On dichotomy above Feder and Vardi's logic

Alexey Barsukov

Université Clermont Auvergne

École Doctorale Sciences pour l'Ingénieur, Clermont-Ferrand

Thèse présentée par
Alexey Barsukov

Pour obtenir le grade de
Docteur d'Université

Spécialité Informatique

On dichotomy above Feder and Vardi's logic

Soutenue publiquement le 13 Décembre 2022 devant le jury constitué de :

Nadia Creignou	*Professeure d'université*	*AMU, Marseille*	Examinatrice
Víctor Dalmau	*Professeur d'université*	*UPF, Barcelone*	Rapporteur
Arnaud Durand	*Professeur d'université*	*Paris 7*	Rapporteur
Florent Foucaud	*Maître de conférences*	*UCA, Aubière*	Invité
Mamadou Kanté	*Maître de conférences (HDR)*	*UCA, Aubière*	Directeur
Florent Madelaine	*Professeur d'université*	*UPEC, Créteil*	Encadrant
Lhouari Nourine	*Professeur d'université*	*UCA, Aubière*	Examinateur
Aline Parreau	*Chargée de recherche*	*CNRS, Lyon 1*	Examinatrice
Yann Strozecki	*Maître de conférences (HDR)*	*UVSQ, Versailles*	Rapporteur

Abstract

A subset of NP is said to have a dichotomy if it contains problem that are either solvable in P-time or NP-complete. The class of finite Constraint Satisfaction Problems (CSP) is a well-known subset of NP that follows such a dichotomy. The complexity class NP does not have a dichotomy unless P = NP. For both of these classes there exist logics that are associated with them.

- NP is captured by Existential Second-Order (ESO) logic by Fagin's theorem, *i.e.*, a problem is in NP if and only if it is expressible by an ESO sentence.

- CSP is a subset of Feder and Vardi's logic, Monotone Monadic Strict NP without inequalities (MMSNP), and for every MMSNP sentence there exists a P-time equivalent CSP problem.

This implies that ESO does not have a dichotomy as well as NP, and that MMSNP has a dichotomy as well as CSP. The main objective of this thesis is to study subsets of NP that strictly contain CSP or MMSNP with respect to the dichotomy existence.

Feder and Vardi proved that if we omit one of the three properties that define MMSNP, namely being monotone, monadic or omitting inequalities, then the resulting logic does not have a dichotomy. As their proofs remain sketchy at times, we revisit these results and provide detailed proofs.

Guarded Monotone Strict NP (GMSNP) is a known extension of MMSNP that is obtained by relaxing the *monadic* restriction of MMSNP. We define similarly a new logic that is called MMSNP with Guarded inequalities ($\mathsf{GMMSNP}_{\neq}$), relaxing the restriction of being *without inequalities*. We prove that $\mathsf{GMMSNP}_{\neq}$ is strictly more expressive than MMSNP and that it also has a dichotomy.

There is a logic MMSNP_2 that extends MMSNP in the same way as MSO_2 extends Monadic Second-Order (MSO) logic. It is known that MMSNP_2 is a fragment of GMSNP and that these two classes either both have a dichotomy or both have not. We revisit this result and strengthen it by proving that, with respect to having a dichotomy, without loss of generality, one can consider only MMSNP_2 problems over one-element signatures, instead of GMSNP problems over arbitrary finite signatures.

We seek to prove the existence of a dichotomy for MMSNP_2 by finding, for every MMSNP_2 problem, a P-time equivalent MMSNP problem. We face some obstacles to build such an equivalence. However, if we allow MMSNP sentences to consist of countably many negated conjuncts, then we prove that such an equivalence exists. Moreover, the corresponding infinite MMSNP sentence has a property of being *regular*. This regular property means that, in some sense, this sentence is still finite. It is known that regular MMSNP problems can be expressed by CSP on ω-categorical templates. Also, there is an algebraic dichotomy characterisation for ω-categorical CSPs that describe MMSNP

problems. If one manages to extend this algebraic characterisation onto regular MMSNP, then our result would provide an algebraic dichotomy for MMSNP_2.

Another potential way to prove the existence of a dichotomy for MMSNP_2 is to mimic the proof of Feder and Vardi for MMSNP. That is, by finding a P-time equivalent CSP problem. The most difficult part there is to reduce a given input structure to a structure of sufficiently large girth. For MMSNP and CSP, it is done using expanders, *i.e.*, structures, where the distribution of tuples is close to a uniform distribution. We study this approach with respect to MMSNP_2 and point out the main obstacles.

We also consider an extension of CSP: the Matrix Partition (MP) problems class. We study it from several perspectives. It is well-known that CSP over an arbitrary finite signature has a dichotomy if and only if CSP on directed graphs has a dichotomy. Motivated by this result, we consider MP problems over arbitrary finite signatures and show that they have a dichotomy if and only if MP problems over one-element signatures have a dichotomy, similarly to our result for MMSNP_2. Another perspective is to characterise MP problems with respect to being definable in First-Order (FO) logic. For CSP, a problem is FO-definable if and only if it has a finitary duality, *i.e.*, a finite family of digraphs such that an input digraph is accepted by the CSP if and only if no digraph from the family is homomorphically mapped to the input one. There have already been some attempts to classify Matrix Partition problems in terms of having finitely many minimal obstructions, *i.e.*, an input graph is accepted by the MP problem if and only if it does not contain an induced subgraph from a given finite family. We manage to show that, for MP problems, these two notions are the same. The third perspective is to find a logic that would be related to MP in a similar way as MMSNP is related to CSP. We introduce, as a potential candidate, a logic obtained from MMSNP by relaxing the *monotone* restriction, and show that it contains MP. However, it is not known how to show the equivalence. At last, we study the notion of *polymorphism* for MP problems. We do it in order to consider the algebraic dichotomy characterisation for finite CSP and see if there is some potential to consider polymorphisms for MP problems. In the case of CSP, a structure has a non-trivial polymorphism if and only if the corresponding CSP is P-time solvable. We manage to provide an MP problem that has only trivial polymorphisms and that is P-time solvable. This means that, for MP problems, the existence of an algebraic characterisation is unlikely.

In an independent chapter, we investigate the Maximum Cut (MaxCut) problem. Although being NP-complete in general, its complexity becomes unknown if we consider only unit interval graphs in the input. Knowing that MaxCut is NP-complete on interval graphs, we approach as close as possible to unit interval graphs by proving that it remains NP-complete even if we are allowed to operate with intervals of only two different lengths.

Dedication

To my parents Igor and Galina for their constant support.

Contents

List of Notations

$[n]$	$\{i \in \mathbb{N} \mid 1 \leq i \leq n\}$
$\alpha(\mathbf{x}), \phi(\mathbf{x}), \ldots$	First-order formulae
$\alpha, \phi, \ldots$	First-order sentences
$\mathcal{A}, \mathcal{F}, \mathcal{G}, \ldots$	Families of objects (graphs, sets, structures, etc.)
$\mathsf{f}, \mathsf{g}, \ldots$	Mappings
$\mathbb{N}$	The set of nonnegative integer numbers
$\mathbb{Z}$	The set of integer numbers
$\mathsf{A}, \mathsf{M}, \ldots$	Matrices
$\Phi(\mathbf{x}), \Psi(\mathbf{x}), \ldots$	Second-order formulae
$\Phi, \Psi, \ldots$	Second-order sentences
$\mathsf{E}, \mathsf{R}, \ldots$	Relation symbols, second-order variables
$\mathsf{E}^{\mathfrak{G}}, \mathsf{R}^{\mathfrak{A}}$	Relations
$\mathfrak{A}, \mathfrak{B}, \ldots$	Relational structures
$\mathfrak{G}, \mathfrak{H}, \ldots$	Graphs
$\tau, \sigma, \rho, \ldots$	Relational signatures
$\mathbf{x}, \mathbf{y}, \mathbf{z}, \ldots$	Tuples of variables or of vertices
$\uplus$	Disjoint union
$\varnothing$	The empty set
$A, G, X, \ldots$	Sets
$i, j, \ldots$	Indices
$k, l, m, n, \ldots$	Natural numbers
$x, y, z, \ldots$	Variables, vertices

Glossary

Complexity

We try to use the same definitions as in Papadimitriou's book [Pap94].

A *Turing machine* is a quadruple $M = (\mathcal{Q}, \Sigma, \delta, q_0)$. Here $\mathcal{Q}$ is a finite set of *states*; $q_0 \in \mathcal{Q}$ is the *initial state*. Σ is a finite set of *symbols* (we say that Σ is the *alphabet* of M). Σ always contains the special symbols $\sqcup$ and $\triangleright$: the *blank* and the *first symbol*. Finally, δ is a *transition function* which maps $\mathcal{Q} \times \Sigma$ to $\big(\mathcal{Q} \cup \{\text{“yes”}, \text{“no”}\}\big) \times \Sigma \times \{\leftarrow, \rightarrow\}$.

For an alphabet Σ, a *string* of symbols of Σ is a finite sequence $s_1 s_2 \dots s_r$, for some r in $\mathbb{N}$. The set of all strings of symbols of Σ is denoted by Σ^*.

The input of a Turing machine is a one-way infinite *tape* with a head that can move freely across the tape. A tape is a sequence of cells, each cell contains one symbol of the alphabet Σ. The first cell of the tape always contains a special symbol $\triangleright$ that represents the left end of the tape. We assume that there exists N in $\mathbb{N}$ such that, for all $n > N$, the nth cell contains $\sqcup$, and, for all i such that $1 < i \leq N$, the ith cell contains a symbol from $\Sigma \setminus \{\triangleright, \sqcup\}$. That is, there is a one-to-one correspondence between one-way infinite tapes and finite strings in $\big(\Sigma \setminus \{\triangleright, \sqcup\}\big)^*$. The head is always in some state of $\mathcal{Q}$, it can move across the tape, and overwrite the symbol of a cell where it is at the moment. At the beginning, the head is at the first cell of the tape, at the cell with $\triangleright$, and it is in the starting state $q_0 \in \mathcal{Q}$. The execution of a Turing machine is a sequence of *iterations*. In this thesis, we always suppose that this sequence is finite. Each iteration is described as follows.

1. The head having a state q in $\mathcal{Q}$ finds itself in some cell of the tape.
2. The head reads the symbol s in Σ of that cell.
3. The transition function $\delta(q, s)$ commands the head what it should do. Suppose $\delta(q, s) = (q', s', D)$, where q' is in $\mathcal{Q} \cup \{\text{“yes”}, \text{“no”}\}$, s' is in Σ, and D is in $\{\leftarrow, \rightarrow\}$; then the head changes its state to q', overwrites s with s', and, depending on D, either moves one cell to the left or one cell to the right.

The head cannot move to the left from the first cell. So we assume that, for any transition rule $\delta(q, \triangleright) = (q', s', D)$, we have $s' = \triangleright$ and $D = \rightarrow$. Also we assume that the first cell is always a unique cell that contains $\triangleright$. So, for any rule $\delta(q, s) = (q', s', D)$, where $s \neq \triangleright$, we have $s' \neq \triangleright$.

Once the head is in a state from $\{\text{“yes”}, \text{“no”}\}$ the machine *halts*. If the state of M is “yes”, then we say that M *accepts* the input, if it is “no”, then M *rejects* the input.

Let M be a Turing machine, and x in Σ^* be a string of symbols. If M reaches one of the states of $\{\text{“yes”}, \text{“no”}\}$ in t iterations, then we say that the *time required by M on input x is t*.

Let $\mathsf{f}\colon \mathbb{N} \to \mathbb{N}$. We say that a Turing machine M *operates within time* $\mathsf{f}(n)$ if, for any input string x, the time required by M on x is at most $\mathsf{f}(|x|)$ (by $|x|$ we denote the *length* of string x).

Let M be a Turing machine, and x in Σ^* be its input. By $M(x)$ we denote the output of the machine. That is a string of $(\Sigma \cup \{\sqcup\})^*$ that is associated with the configuration of the tape at the end of the execution of M.

Let $\mathcal{L} \subset (\Sigma \setminus \{\sqcup\})^*$ be a language. Let M be a Turing machine such that, for any string x in $(\Sigma \setminus \{\sqcup\})^*$, if x is in L, then M accepts x, and, if x is not in L, then M rejects x. Then we say that M *decides* $\mathcal{L}$.

Let $\mathcal{O}(n^k)$ be a set of functions from $\mathbb{N}$ to $\mathbb{N}$ that satisfies the following property. For any function f in $\mathcal{O}(n^k)$ there exist positive integers c and n_0 such that, for any $n \geq n_0$, $\mathsf{f}(n) \leq cn^k$.

We say that M *operates within polynomial time* if it operates within time $\mathsf{f}(n)$, and if there is k in $\mathbb{N}$ such that $\mathsf{f}(n)$ is in $\mathcal{O}(n^k)$.

A *nondeterministic Turing machine* is a quadruple $N = (\mathcal{Q}, \Sigma, \Delta, q_0)$. $\mathcal{Q}, \Sigma$, and q_0 are as before. But now Δ is a relation: $\Delta \subset (\mathcal{Q} \times \Sigma) \times \Big[\big(\mathcal{Q} \cup \{\text{“yes”}, \text{“no”}\}\big) \times \Sigma \times \{\leftarrow, \rightarrow\}\Big]$. For a nondeterministic Turing machine, its execution is not uniquely determined by the input. Suppose that before some iteration the head is in some state q in $\mathcal{Q}$ and is at some cell that contains a symbol s in $(\Sigma \cup \{\sqcup\})$. Then N chooses an element from $\Delta(q, s)$ and the head acts depending on the choice.

For a nondeterministic machine N and some string x in Σ, N *accepts* x if there is an execution of N such that N halts in the “yes” state. Otherwise, N *rejects* x. We say that a nondeterministic Turing machine N *decides* a language $\mathcal{L}$ if, for any x in Σ^*, the following is true: x is in L if and only if N accepts x. We say that N *decides* $\mathcal{L}$ *in time* $\mathsf{f}(n)$, where $\mathsf{f}\colon \mathbb{N} \to \mathbb{N}$, if N decides $\mathcal{L}$, and, for any x in Σ^*, any execution of N with input x has at most $\mathsf{f}(|x|)$ iterations.

Denote by P the set of all the languages $\mathcal{L}$ that can be decided by a Turing machine operating within polynomial time. And denote by NP the set of all the languages $\mathcal{L}$ that can be decided by a nondeterministic Turing machine operating within polynomial time.

Let $\mathsf{f}\colon (\Sigma \setminus \{\sqcup\})^* \to \Sigma^*$. We say that a Turing machine M *computes* f if, for any x in $(\Sigma \setminus \{\sqcup\})^*$, $M(x) = \mathsf{f}(x)$. If M operates within polynomial time, then we write that M *computes* f *in polynomial time*.

Let $\mathcal{L}_1, \mathcal{L}_2$ be languages over an alphabet Σ. We say that $\mathcal{L}_1$ *reduces in polynomial time* to $\mathcal{L}_2$, denoted by $\mathcal{L}_1 \leq_p \mathcal{L}_2$, if there is a function $\mathsf{f}\colon (\Sigma \setminus \{\sqcup\})^* \to \Sigma^*$ and a Turing machine M such that f is computable by M in polynomial time, and that x is in L_1 if and only if $\mathsf{f}(x)$ is in L_2. This is also called *Karp reduction*.

We say that $\mathcal{L}_1$ and $\mathcal{L}_2$ are *polynomial time equivalent*, denoted by $\mathcal{L}_1 \equiv_p \mathcal{L}_2$, if $\mathcal{L}_1 \leq_p \mathcal{L}_2$ and $\mathcal{L}_2 \leq_p \mathcal{L}_1$.

Denote by *NP-complete* the set of all the languages $\mathcal{L}$ such that, for any $\mathcal{L}'$ in NP, we have $\mathcal{L}' \leq_p \mathcal{L}$.

Suppose that P $\neq$ NP. Denote by NP-intermediate the set of all languages $\mathcal{L}$ such that $\mathcal{L}$ is in NP, $\mathcal{L}$ is not in P, and $\mathcal{L}$ is not NP-complete.

For $\mathcal{C} \subseteq$ NP, we say that $\mathcal{C}$ *has a dichotomy* if $\mathcal{C} \cap$ NP-intermediate $= \varnothing$.

Let $\mathcal{C}_1, \mathcal{C}_2$ in NP be two sets of languages. We say that $\mathcal{C}_1$ *is contained in* $\mathcal{C}_2$ *under polynomial time reductions*, denoted by $\mathcal{C}_1 \subseteq_p \mathcal{C}_2$, if for every $\mathcal{L}_1$ in $\mathcal{C}_1$ there exists $\mathcal{L}_2$ in $\mathcal{C}_2$ such that $\mathcal{L}_1 \equiv_p \mathcal{L}_2$. If $\mathcal{C}_1 \subseteq_p \mathcal{C}_2$ and $\mathcal{C}_2 \subseteq_p \mathcal{C}_1$, then we say that $\mathcal{C}_1$ is *equivalent to* $\mathcal{C}_2$ *under polynomial time reductions*, denoted by $\mathcal{C}_1 \equiv_p \mathcal{C}_2$. Clearly, if $\mathcal{C}_1 \subseteq_p \mathcal{C}_2$, then if $\mathcal{C}_2$

has a dichotomy, so does $\mathcal{C}_1$. And, if $\mathcal{C}_1 \equiv_p \mathcal{C}_2$, then $\mathcal{C}_1$ has a dichotomy if and only if $\mathcal{C}_2$ has a dichotomy.

Structures and homomorphisms

We try to use the same definitions as in the book of Hell and Nešetřil [HN04] and as in Bodirsky's book [Bod21].

Let A, B be two sets. The *product of sets* A and B, denoted by $A \times B$, is the set of pairs: $\{(a, b) \mid a \in A, b \in B\}$. For a set A and k in $\mathbb{N}$, the *k-ary product* of A, denoted by A^k, is the following set of k-tuples: $\{(a_1, \ldots, a_k) \mid a_1, \ldots, a_k \in A\}$.

A *relational signature* τ is a set of relation symbols R each of which has an associated finite *arity* k in $\mathbb{N}$. A relational *structure* $\mathfrak{A}$ over a signature τ (also called *τ-structure*) consists of a set A (the *domain*) together with a relation $\mathsf{R}^{\mathfrak{A}} \subseteq A^k$ for each relation symbol R in τ, where k is the associated arity. For any k-tuple $\mathbf{a}$ in A^k, if $\mathbf{a}$ is in $\mathsf{R}^{\mathfrak{A}}$, then we usually write $\mathsf{R}^{\mathfrak{A}}(\mathbf{a})$. We consider only finite relational signatures; and relational structures with at most countable domains.

A *homomorphism* between τ-structures $\mathfrak{A}$ and $\mathfrak{B}$ with domains A and B is a mapping $\mathsf{h}\colon A \to B$ that *preserves* the relations; that is, for any R in τ of some arity k and any $(a_1, \ldots, a_k)$ in A^k, $\mathsf{R}^{\mathfrak{A}}(a_1, \ldots, a_k)$ implies $\mathsf{R}^{\mathfrak{B}}\big(\mathsf{h}(a_1), \ldots, \mathsf{h}(a_k)\big)$. If h is a homomorphism between $\mathfrak{A}$ and $\mathfrak{B}$, then it is denoted by $\mathsf{h}\colon \mathfrak{A} \to \mathfrak{B}$. For a tuple $\mathbf{a} = (a_1, \ldots, a_k)$ in A^k and a homomorphism $\mathsf{h}\colon \mathfrak{A} \to \mathfrak{B}$, we denote $\mathsf{h}(\mathbf{a}) := \big(\mathsf{h}(a_1), \ldots, \mathsf{h}(a_k)\big)$. For τ-structures $\mathfrak{A}, \mathfrak{B}$, if there exist homomorphisms $\mathsf{h}_1\colon \mathfrak{A} \to \mathfrak{B}$ and $\mathsf{h}_2\colon \mathfrak{B} \to \mathfrak{A}$, then $\mathfrak{A}$ and $\mathfrak{B}$ are called *homomorphically equivalent*, it is denoted by $\mathfrak{A} \leftrightarrows \mathfrak{B}$. A homomorphism $\mathsf{h}\colon \mathfrak{A} \to \mathfrak{B}$ is called *surjective* if the mapping $\mathsf{h}\colon A \to B$ is surjective. An *injective* homomorphism is defined similarly.

A homomorphism $\mathsf{f}\colon \mathfrak{A} \to \mathfrak{B}$ is called *full* if, for any R in τ of arity k and any $\mathbf{a}$ in A^k, $\mathsf{R}^{\mathfrak{B}}\big(\mathsf{f}(\mathbf{a})\big)$ implies $\mathsf{R}^{\mathfrak{A}}(\mathbf{a})$. A full homomorphism $\mathsf{i}\colon \mathfrak{A} \to \mathfrak{B}$ is called an *isomorphism* between $\mathfrak{A}$ and $\mathfrak{B}$ if $\mathsf{i}\colon A \to B$ is one-to-one. If there exists an isomorphism $\mathsf{i}\colon \mathfrak{A} \to \mathfrak{B}$, then the two structures are said to be *isomorphic*, we denote it by $\mathfrak{A} \cong \mathfrak{B}$. An isomorphism $\mathsf{a}\colon \mathfrak{A} \to \mathfrak{A}$ from a structure to itself is called an *automorphism*.

Let $\mathfrak{A}$ be a τ-structure for some relational signature τ. Let $B \subseteq A$. A τ-structure $\mathfrak{B}$ with domain B is called a *substructure* of $\mathfrak{A}$ if, for any R in τ, $\mathsf{R}^{\mathfrak{B}} \subseteq \mathsf{R}^{\mathfrak{A}}$. And $\mathfrak{B}$ is called an *induced substructure* of $\mathfrak{A}$ if, for any R in τ of arity k, $\mathsf{R}^{\mathfrak{B}} = B^k \cap \mathsf{R}^{\mathfrak{A}}$. Equivalently, we can say that $\mathfrak{B}$ is the *substructure of* $\mathfrak{A}$ *induced on* B and write $\mathfrak{A}[B]$. When $\mathfrak{B}$ is an induced substructure of $\mathfrak{A}$, we also say that $\mathfrak{B}$ *embeds* into $\mathfrak{A}$ and denote it as follows: $\mathfrak{B} \hookrightarrow \mathfrak{A}$.

Let $\mathsf{h}\colon A \to B$ be a mapping between two sets. The set $\mathsf{h}(A) := \{b \in B \mid \exists a \in A\colon \mathsf{h}(a) = b\}$ is called the *image* of h. A homomorphism $\mathsf{e}\colon \mathfrak{A} \to \mathfrak{B}$ is called an *embedding* if $\mathfrak{A} \cong \mathfrak{B}[\mathsf{e}(A)]$, it is denoted by $\mathsf{e}\colon \mathfrak{A} \hookrightarrow \mathfrak{B}$.

Let $\mathfrak{A}, \mathfrak{B}$ be τ-structures with domains A and B correspondingly. The *union* of two τ-structures $\mathfrak{A}$ and $\mathfrak{B}$, denoted by $\mathfrak{A} \cup \mathfrak{B}$, is a τ-structure with domain $A \cup B$ and, for any R in τ of arity k, $\mathsf{R}^{\mathfrak{A} \cup \mathfrak{B}} = \mathsf{R}^{\mathfrak{A}} \cup \mathsf{R}^{\mathfrak{B}}$. The *intersection* of τ-structures $\mathfrak{A}$ and $\mathfrak{B}$, denoted by $\mathfrak{A} \cap \mathfrak{B}$, is defined similarly. A *disjoint union* of $\mathfrak{A}$ and $\mathfrak{B}$, denoted by $\mathfrak{A} \uplus \mathfrak{B}$, is the union of two isomorphic copies of $\mathfrak{A}$ and $\mathfrak{B}$ with disjoint domains. As disjoint unions are unique up to isomorphism, we usually speak of *the* disjoint union of $\mathfrak{A}$ and $\mathfrak{B}$. If a structure is not isomorphic to the disjoint union of structures, then it is called *connected*.

For two k-tuples $\mathbf{a} = (a_1, \ldots, a_k)$ in A^k, $\mathbf{b} = (b_1, \ldots, b_k)$ in B^k, denote $\mathbf{a} \times \mathbf{b} := \big((a_1, b_1), \ldots, (a_k, b_k)\big)$. For τ-structures $\mathfrak{A}$ and $\mathfrak{B}$, the *product* of $\mathfrak{A}$ and $\mathfrak{B}$, denoted by $\mathfrak{A} \times \mathfrak{B}$, is a τ-structure with domain $A \times B$ and, for any k-ary R in τ and, for any $\mathbf{a}$ in A^k ,$\mathbf{b}$ in B^k, if $\mathsf{R}^{\mathfrak{A}}(\mathbf{a})$ and $\mathsf{R}^{\mathfrak{B}}(\mathbf{b})$, then $\mathsf{R}^{\mathfrak{A}\times\mathfrak{B}}(\mathbf{a} \times \mathbf{b})$.

A binary relation $\mathfrak{e}(\cdot, \cdot)$ is called an *equivalence relation* on a set A if it is reflexive, symmetric and transitive.

Let $\sim$ be an equivalence relation on a set A. For some a in A, a set $[a]_\sim = \{x \in A \mid a \sim x\}$ is called the *$\sim$-equivalence class of a*. For any equivalence relation, its equivalence classes make a partition of A. The set of all the $\sim$-equivalence classes of elements of A is denoted by $A/\sim$.

For a relational structure $\mathfrak{A}$ and an equivalence relation $\sim$ defined on this structure, the *quotient* $\mathfrak{A}/\sim$ is also a relational structure with domain $A/\sim$, and, for some relation R, a k-tuple $\mathbf{a} = ([a_1]_\sim, \ldots, [a_k]_\sim)$ belongs to $\mathsf{R}^{\mathfrak{A}/\sim}$ of $\mathfrak{A}/\sim$ if there exists a tuple $\mathbf{a}' = (a'_1, \ldots, a'_k)$ such that, for all i in $[k]$, a'_i is in the equivalence class $[a_i]_\sim$, and $\mathbf{a}'$ belongs to $\mathsf{R}^{\mathfrak{A}}$. There is always a surjective homomorphism $\pi\colon \mathfrak{A} \to \mathfrak{A}/\sim$ such that $\pi(a) = [a]_\sim$. For a k-tuple $\mathbf{a} = (a_1, \ldots, a_k)$ in A^k, denote $[\mathbf{a}]_\sim := ([a_1]_\sim, \ldots, [a_k]_\sim)$.

Let τ and σ be two relational signatures. For a $\tau \uplus \sigma$-structure $\mathfrak{A}$ and a τ-structure $\mathfrak{A}^\tau$, we say that $\mathfrak{A}^\tau$ is the *τ-reduct* of $\mathfrak{A}$ if the structures have the same domain A and, for any R in τ, $\mathsf{R}^{\mathfrak{A}} = \mathsf{R}^{\mathfrak{A}^\tau}$. In this case, $\mathfrak{A}$ is called a *σ-expansion* of $\mathfrak{A}^\tau$. Observe that there is a unique reduct but there may be many expansions.

A relational structure $\mathfrak{G} = (G, \mathsf{E}^{\mathfrak{G}})$, where E has arity 2, is called a *directed graph*. Elements of G are called *vertices*, tuples of $\mathsf{E}^{\mathfrak{G}}$ are called *arcs*.

Two distinct vertices v, w in G are called *adjacent* if (v, w) is in $\mathsf{E}^{\mathfrak{G}}$ or (w, v) is in $\mathsf{E}^{\mathfrak{G}}$. A sequence of vertices $v_1, \ldots, v_n$ of G is called an *walk*, if, for any i in $[n-1]$, v is adjacent to v_{i+1}. A *path* is a walk $v_1, \ldots, v_n$ such that all its vertices are pairwise distinct. A *cycle* is a path $v_1, \ldots, v_n$ such that v_1 is adjacent to v_n. The *length* of a walk is the number of vertices in the sequence. A digraph is called a *tree* if it contains no cycle.

For a walk $v_1, \ldots, v_n$, an arc $\mathsf{E}^{\mathfrak{G}}(v_i, v_{i+1})$ is called a *forward arc*, and an arc $\mathsf{E}^{\mathfrak{G}}(v_{i+1}, v_i)$ is called a *backward arc*. The *net length* of a walk is the difference between the number of forward arcs and the number of backward arcs. A walk/path/cycle is called *directed* if all its arcs are forward. If a digraph contains no directed cycle, then it is called *acyclic*.

For a directed graph $\mathfrak{G}$, the *girth* of $\mathfrak{G}$ is the shortest length of a cycle of $\mathfrak{G}$.

For a vertex v in G, the *degree* of v is the number of vertices w in G such that v is adjacent to w. The *in-degree* of v is the number of vertices w in G such that (w, v) is in $\mathsf{E}^{\mathfrak{G}}$. The definition of the *out-degree* is similar.

Let $\mathfrak{A}$ be a relational τ-structure. Two elements a_1, a_2 of $\mathfrak{A}$ are called *adjacent* if there is a tuple $\mathbf{a}$ in $\mathsf{R}^{\mathfrak{A}}$, for some R in τ, such that both a_1, a_2 are contained in $\mathbf{a}$. The definitions of a walk, path, cycle, and degree are similar.

We use the same definitions for the notions of the category theory as in the book of Awodey, see [Awo10]. For a relational signature τ, the τ-structures together form a category, where the objects are τ-structures themselves and the arrows are homomorphisms between the corresponding τ-structures. The category of τ-structures is denoted by $\mathbf{Struct}[\tau]$.

Any relational structure can be encoded by a string consisting of 0 and 1. Such string is an input instance of a Turing machine. We explain how the encoding is described in [Lib04]. Consider an n-element relational τ-structure $\mathfrak{A}$, where $\tau = \{\mathsf{R}_1, \ldots, \mathsf{R}_t\}$. Suppose that its elements are linearly ordered: $a_1 < a_2 < \cdots < a_n$. It induces lexico-

graphical orderings of k-tuples consisting of these elements, for any k in $\mathbb{N}$, $(a_1, \ldots, a_1) < (a_1, \ldots, a_1, a_2) < \cdots < (a_n, \ldots, a_n)$. For R of arity k in τ, the relation $\mathsf{R}^{\mathfrak{A}}$ can be encoded by a n^k-bit string $\mathsf{enc}\big(\mathsf{R}^{\mathfrak{A}}\big)$: the ith element of $\mathsf{enc}\big(\mathsf{R}^{\mathfrak{A}}\big)$ equals 1 if and only if the ith tuple in the lexicographical ordering belongs to $\mathsf{R}^{\mathfrak{A}}$. The whole structure $\mathfrak{A}$ can be encoded by the following string:

$$\mathsf{enc}(\mathfrak{A}) := 0^n 1 \cdot \mathsf{enc}\big(\mathsf{R}_1^{\mathfrak{A}}\big) \cdot \cdots \cdot \mathsf{enc}\big(\mathsf{R}_t^{\mathfrak{A}}\big),$$

where 0^n means the string of length n consisting only of 0s, and the sign $\cdot$ means the concatenation of strings.

Remark. Further in this thesis, when we show a reduction from one problem to another, we neither consider the problems as languages of strings nor explicitly describe the Turing machine that computes the function between the languages. Usually, input instances of a problem are relational structures. The reduction is shown by describing an algorithm that takes an input structure $\mathfrak{A}_1$ of the first problem P_1 and returns an input structure $\mathfrak{A}_2$ of the second problem P_2 such that $\mathfrak{A}_1$ is accepted by P_1 if and only if $\mathfrak{A}_2$ is accepted by P_2. And, moreover, the runtime of the algorithm must be polynomial in size of $\mathfrak{A}_1$.

Logic

We try to use the same definitions as in Libkin's book [Lib04].

Let τ be a relational signature. We assume a countably infinite set of (first-order) variables that range over the domains of τ-structures. We inductively define *first-order formulae* over τ as follows.

- If x_1, x_2 are variables, then $x_1 = x_2$ is an (*atomic*) formula.
- If $x_1, \ldots, x_k$ are variables and R in τ is a k-ary relation symbol, then $\mathsf{R}(x_1, \ldots, x_k)$ is an (*atomic*) formula.
- If ϕ_1, ϕ_2 are formulae, then $\phi_1 \wedge \phi_2, \phi_1 \vee \phi_2$, and $\neg\phi_1$ are formulae.
- If ϕ is a formula, then $\forall x\ \phi$ and $\exists x\ \phi$ are formulae.

The set of all first-order formulae is denoted by FO and is called *first-order logic*. We do not consider signatures with constants and function symbols, so we do not need to mention them in the definition of FO. For a formula ϕ in FO, denote by $\mathrm{Var}(\phi)$ its set of variables.

For a relation symbol R in τ, an atomic formula of the form $\mathsf{R}(x_1, \ldots, x_k)$ is called an R-*atom* or a τ-*atom*.

A formula that does not use existential ($\exists$) and universal ($\forall$) quantifiers is called *quantifier-free*. It is called *universal* if all the quantifiers are universal. It is called *existential* if all the quantifiers are existential.

A τ-formula $\phi(x_1, \ldots, x_n)$ is called *primitive positive* if it is of the form:

$$\exists x_{n+1}, \ldots, x_l(\psi_1 \wedge \cdots \wedge \psi_m),$$

where $\psi_1, \ldots, \psi_m$ are τ-atoms.

We use the standard shorthand $\phi \to \psi$ for $\neg\phi \vee \psi$ and $\phi \leftrightarrow \psi$ for $(\phi \to \psi) \wedge (\psi \to \phi)$.

Suppose that a formula ϕ contains a variable x. If x is not quantified neither existentially nor universally, then the variable x is called *free*; otherwise it is called *bound*. If $\mathbf{x}$ is a tuple of all the free variables of ϕ, then we write $\phi(\mathbf{x})$.

If a formula ϕ contains no free variables, then ϕ is called a *sentence*. If we want to precise that it is over a signature τ, then we usually call it a *τ-sentence*.

Given a τ-structure $\mathfrak{A}$, we define inductively, for each first-order formula ϕ with its tuple $\mathbf{a}$ of free variables, the notion $\mathfrak{A} \models \phi(\mathbf{a})$ (*i.e.* $\phi(\mathbf{a})$ is true in $\mathfrak{A}$). That is, we define the semantics of the first-order logic.

- If ϕ is of the form $(x_1 = x_2)$, then, for $\mathbf{a} = (a_1, a_2)$, $\mathfrak{A} \models \phi(a_1, a_2)$ if and only if $a_1 = a_2$.
- If ϕ is of the form $\mathsf{R}(x_1, \ldots, x_k)$, for some k-ary R in τ, then $\mathfrak{A} \models \phi(\mathbf{a})$ if and only if $\mathbf{a}$ is in $\mathsf{R}^{\mathfrak{A}}$.
- $\mathfrak{A} \models \neg\phi(\mathbf{a})$ if and only if $\mathfrak{A} \models \phi(\mathbf{a})$ is false.
- $\mathfrak{A} \models \phi_1(\mathbf{a}_1) \wedge \phi_2(\mathbf{a}_2)$ if and only if $\mathfrak{A} \models \phi_1(\mathbf{a}_1)$ and $\mathfrak{A} \models \phi_2(\mathbf{a}_2)$.
- $\mathfrak{A} \models \phi_1(\mathbf{a}_1) \vee \phi_2(\mathbf{a}_2)$ if and only if $\mathfrak{A} \models \phi_1(\mathbf{a}_1)$ or $\mathfrak{A} \models \phi_2(\mathbf{a}_2)$.
- If $\psi(\mathbf{x})$ is of the form $\exists y \phi(y, \mathbf{x})$, then $\mathfrak{A} \models \psi(\mathbf{a})$ if and only if $\mathfrak{A} \models \phi(a', \mathbf{a})$ for some a' in A.
- If $\psi(\mathbf{x})$ is of the form $\forall y \phi(y, \mathbf{x})$, then $\mathfrak{A} \models \psi(\mathbf{a})$ if and only if $\mathfrak{A} \models \phi(a', \mathbf{a})$ for all a' in A.

A set of first-order τ-sentences is called a *first-order theory*. For some first-order theory T, a τ-structure $\mathfrak{A}$ is called a *model* of T, if any ϕ of T is true in $\mathfrak{A}$. For a τ-structure $\mathfrak{A}$, the *first-order theory* of $\mathfrak{A}$ is the set of all first-order sentences ϕ such that ϕ is true in $\mathfrak{A}$, it is denoted by $\mathrm{Th}(\mathfrak{A})$.

Assume, apart from first-order variables, that for any k in $\mathbb{N}$ there is a countably infinite set of *second-order variables* $\mathsf{X}^k_1, \mathsf{X}^k_2, \ldots$ ranging over k-ary relations. The *second-order logic*, denoted by SO, is the set of all second-order formulae $\phi(\mathbf{x}, \mathbf{X})$; each such formula contains both first and second order variables and is inductively defined as follows.

- Any atomic FO formula is an atomic SO formula; also, for variables $x_1, \ldots, x_k$ and for a k-ary SO variable X, $\mathsf{X}(x_1, \ldots, x_k)$ is an atomic SO formula.
- If ϕ_1, ϕ_2 are in SO, then $\phi_1 \wedge \phi_2, \phi_1 \vee \phi_2, \neg\phi_1, \exists x\ \phi_1, \forall x\ \phi_1$ are in SO.
- If $\phi(\mathbf{x}, \mathsf{Y}, \mathbf{X})$ is in SO, then $\exists \mathsf{Y}\ \phi(\mathbf{x}, \mathbf{X})$ and $\forall \mathsf{Y}\ \phi(\mathbf{x}, \mathbf{X})$ are in SO.

The semantics of the second-order logic extend the semantics of FO as follows.

- Suppose $\phi(\mathbf{x}, \mathsf{X})$ is of the form $\mathsf{X}(x_1, \ldots, x_k)$; then $\mathfrak{A} \models \phi(\mathbf{b}, B)$ if and only if $\mathbf{b}$ is in B, where B is some subset of A^k.
- Suppose $\psi(\mathbf{x}, \mathbf{X})$ is of the form $\exists \mathsf{Y}\ \phi(\mathbf{x}, \mathsf{Y}, \mathbf{X})$, where Y is a second-order variable with arity k in $\mathbb{N}$; then $\mathfrak{A} \models \psi(\mathbf{b}, \mathbf{B})$ if and only if there is $C \subseteq A^k$ such that $\mathfrak{A} \models \phi(\mathbf{b}, C, \mathbf{B})$.

- Suppose $\psi(\mathbf{x}, \mathbf{X})$ is of the form $\forall \mathsf{Y}\ \phi(\mathbf{x}, \mathsf{Y}, \mathbf{X})$, where Y is a second-order variable with arity k in $\mathbb{N}$; then $\mathfrak{A} \models \psi(\mathbf{b}, \mathbf{B})$ if and only if for any subset $C \subseteq A^k$, we have $\mathfrak{A} \models \phi(\mathbf{b}, C, \mathbf{B})$.

Let L be a logic, *i.e.*, a set of formulae. Let $\mathcal{S}$ be a set of τ-structures. We say that $\mathcal{S}$ is *definable* in L (or, L-*definable*), if there is a sentence ϕ in L such that, for any τ-structure $\mathfrak{A}$, $\mathfrak{A} \models \phi$ if and only if $\mathfrak{A}$ is in $\mathcal{S}$.

For a τ-structure $\mathfrak{A}$, we say that a k-ary relation $\mathsf{R} \subseteq A^k$ is *definable* by a formula $\phi(x_1, \ldots, x_k)$ if, for every k-tuple $(a_1, \ldots, a_k)$ in A^k, we have $(a_1, \ldots, a_k) \in \mathsf{R}$ if and only if $\mathfrak{A} \models \phi(a_1, \ldots, a_k)$. In particular, if ϕ is primitive positive, then we say that R has a *primitive positive definition* in $\mathfrak{A}$, or just R is *pp-definable* in $\mathfrak{A}$.

Let ϕ be a conjunction of non-negated atomic formulae over some relational signature τ. Let $\mathfrak{A}$ be a τ-structure. Suppose that there is a one-to-one correspondence f between variables $x_1, \ldots, x_n$ of ϕ and the vertices $a_1, \ldots, a_n$ of $\mathfrak{A}$. Also suppose that, for any R in τ, ϕ contains an atom $\mathsf{R}(\mathbf{x})$ if and only if the tuple $\mathsf{f}(\mathbf{x})$ belongs to the relation $\mathsf{R}^{\mathfrak{A}}$. In this case, ϕ is called the *canonical conjunctive query* of $\mathfrak{A}$, and $\mathfrak{A}$ is called the *canonical database* of ϕ.

Chapter 1

Introduction

1.1 Dichotomy question overview

1.1.1 The beginning

Ladner theorem

P vs NP is one of the most famous open computer science problems. Despite failing to solve it, researchers pose new questions assuming that there is an answer to P vs NP conjecture. In particular, Ladner shows in [Lad75] that the class of NP-intermediate problems is not empty under the assumption that P $\neq$ NP, see Figure 1.1 for an illustration.

Theorem 1.1.1 ([Lad75])**.** *Suppose that $P \neq NP$. Then there is a problem $\mathcal{L}$ in NP that is neither solvable in polynomial time, nor NP-complete.*

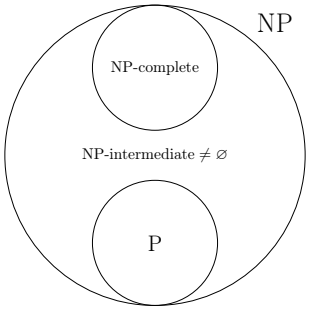


Figure 1.1: An illustration of complexity subclasses of NP assuming that P $\neq$ NP.

1.1.2 Constraint Satisfaction

The *Constraint Satisfaction Problem* enjoys many definitions. A simple one is as follows. Let $\mathfrak{B}$ be a finite relational structure. The problem CSP($\mathfrak{B}$) accepts every finite relational structure $\mathfrak{A}$ that homomorphically maps to $\mathfrak{B}$. The class CSP is the set of all such problems CSP($\mathfrak{B}$), for any $\mathfrak{B}$.

The complexity of this NP problem that generalises both SAT and 3-Colorability is then classified when parametrised by the target of the homomorphism – the so called template of the problem – which is fixed and not part of the input.

Feder and Vardi conjectured a dichotomy for this class in [FV98] in particular because of the following noted results in the boolean case and in the case of undirected graphs.

Theorem 1.1.2 (Schaefer, [Sch78]). *Let $\mathfrak{B}$ be a relational structure with a two-element domain. Then either $(\{0,1\}; \mathsf{NAE})$ has a primitive positive definition in $\mathfrak{B}$, and CSP($\mathfrak{B}$) is NP-complete, or*

1. *$\mathfrak{B}$ is preserved by a constant operation.*
2. *$\mathfrak{B}$ is preserved by $\mathfrak{min}$. In this case, every relation of $\mathfrak{B}$ has a definition by a propositional Horn formula.*
3. *$\mathfrak{B}$ is preserved by $\mathfrak{max}$. In this case, every relation of $\mathfrak{B}$ has a definition by a dual-Horn formula, that is, by a propositional formula in CNF where every clause contains at most one negative literal.*
4. *$\mathfrak{B}$ is preserved by the majority operation. In this case, every relation of $\mathfrak{B}$ can be defined by 2-CNF.*
5. *$\mathfrak{B}$ is preserved by the minority operation. In this case, every relation of $\mathfrak{B}$ can be defined by a conjunction of linear equations modulo 2.*

In each of these five cases, CSP($\mathfrak{B}$) can be solved in P-time.

Theorem 1.1.3 (Hell, Nešetřil, [HN90]). *If an undirected graph $\mathfrak{H}$ is bipartite, then CSP($\mathfrak{H}$) is P-time solvable. Otherwise CSP($\mathfrak{H}$) is NP-complete.*

This conjecture was attacked on several fronts. One direction of research involved the study of special case of digraphs as it was shown that proving the conjecture for digraphs was sufficient.

Theorem 1.1.4 ([FV98]). *For a finite relational structure $\mathfrak{B}$ there exists a balanced directed graph $\mathfrak{H}$ such that CSP($\mathfrak{B}$) is P-time equivalent to CSP($\mathfrak{H}$).*

The approach that proved the most fruitful involved algebra, pushing further the methods initially used by Shaefer : progress was made in the nineties by Cohen, Jeavons and their co-authors. An important milestone was the two dichotomy theorems of Bulatov at the turn of the century, in the three element case and in the conservative case [Bul03, BJK05, Bul06], pushing the methodology further into algebra, using Congruence Tame Theory.

Finally, the dichotomy conjecture on finite relational structures was proved independently by Bulatov in [Bul17] and Zhuk in [Zhu20]. The tractability condition is described by an existence of symmetries within the structure. The symmetries are described by polymorphisms. Below we provide an equivalent formulation of the theorem that uses the notion of Siggers polymorphism, as we also use it in this thesis. The absence of a Siggers polymorphism was already known to imply NP-completeness, and it remained to provide a classification in the presence of a Siggers polymorphism. This was done by providing tractable algorithms in all such remaining cases.

Theorem 1.1.5 (Bulatov, Zhuk). *For every finite relational structure $\mathfrak{A}$, if $\mathfrak{A}$ admits a Siggers polymorphism, then CSP($\mathfrak{A}$) is P-time solvable. Otherwise it is NP-complete.*

Meanwhile researchers investigated many variations of CSP : quantified and in P-space with QCSP [Mar17], with an aspect of optimization with VCSP [KZ17] or more recently approximation with the promise CSP (PCSP) [BBKO21]. An extension of the CSP that shall concern us in this thesis is championed by Bodirsky and involves CSP on infinite domains, most often on languages that can be described by an omega-categorical template. The problem definition is the same as in the finite case, but with an infinite domain structure. Such well known structures in model theory are sufficiently similar to finite structures that some of the algebraic approach can be adapted and extended and allows to prove many dichotomies on natural classes of problems that could not be captured in the finite setting and were already studied in the literature.

The field of infinite CSP is now quite mature, see [Bod21]. There is an analogue of the Feder and Vardi's conjecture. Here, $\mathcal{P}$ denotes the clone of projections:

$$\mathcal{P} := \{\pi_i^n : X^n \to X \mid i, n \in \mathbb{N},\ i < n, \forall x_1, \ldots, x_n\ \pi_i^n(x_1, \ldots, x_n) = x_i\}.$$

Conjecture 1.1.6. *[BKO$^+$17, BMM18] Let $\mathfrak{B}$ be a first-order reduct of a finitely bounded homogeneous structure with finite relational signature. If there is no uniformly continuous minor-preserving map from Pol($\mathfrak{B}$) to $\mathcal{P}$, then CSP($\mathfrak{B}$) is in P. If there is such a mapping, then it is NP-complete.*

1.1.3 Descriptive complexity

Descriptive complexity relates (finite) model theory with complexity theory by showing a correspondence between the amount of logic one needs to describe a problem with the amount of computation problem necessary to solve it. This area goes back to the following theorem.

Theorem 1.1.7 ([Fag74])**.** *For every ESO sentence Φ, the problem SAT(Φ) is in NP. For every property $\mathcal{P}$ defined on the set of finite structures and decidable in P-time, there is a sentence $\Phi_{\mathcal{P}}$ in ESO such that, for any finite structure $\mathfrak{A}$, we have $\mathfrak{A} \models \Phi_{\mathcal{P}}$ if and only if $\mathfrak{A}$ has the property $\mathcal{P}$.*

Feder and Vardi in [FV98] introduced the logic MMSNP as a candidate to express CSP. This logic is a fragment of SNP $\subset$ ESO, which is defined by three conditions: monotone, monadic and without inequalities.

They showed that if MMSNP loses one of them, then it becomes P-time equivalent to NP, and, thus, it has no dichotomy, by Ladner's theorem. We revisit these three classes in Chapter 2.

On the other hand, they showed that MMSNP $\equiv_p$ CSP, which, as they conjectured, should have a dichotomy. Their approach used randomised reductions, that is, the reduction is polynomial but probabilistic. Nonetheless, Kun in [Kun13] provided an explicit P-time equivalence between MMSNP and CSP by determinising this randomised reduction.

THe problems described by sentences of MMSNP may not be in CSP, Madelaine et al. showed in [MS07] that the question of being in CSP is decidable for MMSNP problems. However, they are always P-time equivalent to infinite CSP with an ω-categorical template, see [BD13]. After the algebraic characterisation of CSP given by Bulatov and Zhuk, a similar question was answered for a fragment of CSP on ω-categorical structures

which is definable by MMSNP sentences. Bodirsky, Madelaine and Mottet in [BMM18] proved the following result.

Theorem 1.1.8 ([BMM18])**.** *Let $\mathfrak{B}$ be an ω-categorical structure such that CSP($\mathfrak{B}$) is described by an MMSNP sentence. Then exactly one of the following holds:*

- *There is a uniformly continuous minor-preserving map from Pol($\mathfrak{B}$) to the clone of projections on a 2-element set, and CSP($\mathfrak{B}$) is NP-complete, or*
- *there is no such map and CSP($\mathfrak{B}$) is in P.*

1.1.4 The whole picture

The purpose of this thesis is to investigate classes of problems and fragments of logics that are between the three known non dichotomic logics and above the known logic or classes of problems known to have a dichotomy (CSP, MMSNP) (see Figure 1.2 for an illustration).

The various classes of interest investigated in this thesis are introduced in more details in the next section. MP is a family of combinatorial problems. MMSNP_2 and GMSNP are logics.

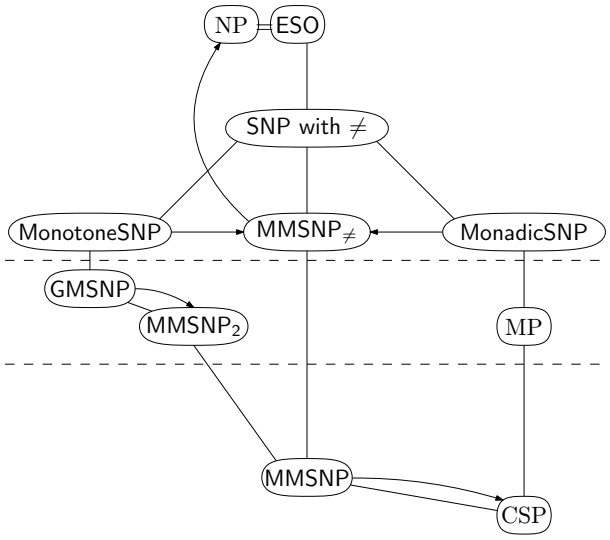


Figure 1.2: An illustration of complexity classes with respect to having a dichotomy. An arc going from a class $\mathcal{C}_1$ to a class $\mathcal{C}_2$ means that $\mathcal{C}_2 \subseteq_p \mathcal{C}_1$ – for every problem of $\mathcal{C}_2$, $\mathcal{C}_1$ contains a P-time equivalent problem. This implies that, if $\mathcal{C}_1$ has a dichotomy, then so does $\mathcal{C}_2$. If a class $\mathcal{C}_1$ is below a class $\mathcal{C}_2$ and if there is an edge between $\mathcal{C}_1$ and $\mathcal{C}_2$, then $\mathcal{C}_1$ is a subclass of $\mathcal{C}_2$.

The main objective is to expand known classes that admit a dichotomy. For some classes like MMSNP_2 that is an example of an infinite CSP à la Bodirsky, we hope to obtain an algebraic dichotomy in the spirit of Theorem 1.1.8.

1.2 Motivation

1.2.1 Investigate MMSNP extensions

Guarded fragments

There are natural ways when can think of to create a logic that lies above MMSNP and below the three fragments that are provably non dichotomic, and drop a restriction on MMSNP, namely to relax this restriction: one canonical example is MMSNP_2 which relaxes the requirement that the logic is not necessarily monadic as in MMSNP, yet does not allow proper binary quantification as in MonotoneSNP, and only allows for colouring tuples of input relations. This is similar in spirit to MSO_2 and MSO_1 considered by Courcelle in [CE12].

MMSNP_2 is subsumed by GMSNP, the guarded fragment of MonotoneSNP, yet they are in fact equally expressive as any sentence of the latter is logically equivalent to a sentence of the former.

Looking at Figure 1.2 on page 22 may give reader a feeling that, similarly to GMSNP being a guarded fragment of MonotoneSNP, there might be fragments of MonadicSNP and of MMSNP with inequalities that are also "guarded" in some sense.

We investigate in Section 3.1 a logic $\mathsf{GMMSNP}_{\neq}$ that extends MMSNP by allowing inequalities within negated conjuncts. But, for each inequality $x \neq y$ there must be a τ-atom $\mathsf{R}(\mathbf{z})$ within the same conjunct such that both x and y are contained in $\mathbf{z}$. We say, in this case, that $\mathsf{R}(\mathbf{z})$ guards $x \neq y$. We manage to show that $\mathsf{GMMSNP}_{\neq} \equiv_p \mathsf{MMSNP}$ and thus it has a dichotomy.

It is not quite clear what it could mean to weaken MonadicSNP (the extension of MMSNP that allows negation) by guarding negation. We study this case in Section 6.5, where we attempt to find a logic that is equivalent to the class of Matrix Partition problems, an extension of CSP that is not monotone to be introduced shortly. This logic is still monadic, it does not have inequalities, but violates the property of being monotone, that is, input atoms may be both negated and non-negated. But we restrict it to be such that, within each negated conjunct, all the input atoms have the same polarity: either all negated or all non-negated. The question of having a dichotomy for this logic is still open.

Edge colouring

It is known, see [BtCLW14], that the guarded fragment GMSNP of MonotoneSNP is strictly more expressible than MMSNP, so showing a dichotomy (or its absence) would be an important result. We study this logic in Section 3.2.

Feder and Vardi showed in [FV98] that the signature of a CSP problem can be simplified up to digraph homomorphisms. We ask a similar question several times in this thesis: for the logic GMSNP in Section 3.2.1 and for generalisations of Matrix partitions in Section 6.3.2. Both times we manage to show that problems over an arbitrary finite signature are P-time equivalent to problems over a signature consisting of a unique relation symbol. However, it is not known in any of these two cases, if we can reduce the arity of this symbol up to 2 as Feder and Vardi did for CSP.

Bienvenu et al. showed in [BtCLW14] that their logic GMSNP is P-time equivalent to a logic MMSNP_2 which had been introduced earlier by Madelaine in [Mad09]. This

class is as expressive as GMSNP but its definition is simpler. It is similar to MMSNP but now we can colour edges (tuples) as well, when in MMSNP we can colour vertices only. In Section 3.2.2 we show that without loss of generality we can study MMSNP_2 problems over one-relation signatures.

Infinite MMSNP

It is not known if MMSNP_2 has a dichotomy. But it is quite natural to try showing that it is equivalent to MMSNP, as these two logics are similar. In Chapter 4 we study connections between them.

As MMSNP_2 colours edges, we try to mimic this by adding to every edge a new vertex whose mission is to represent the edge and its colour. As we all know, an edge is incident to just one pair of vertices. But these new vertices are not aware of this rule, and MMSNP is not expressive enough so that we could inform them about it. Because of this, we obtain an undesirable concept of duplicated edges (tuples). This becomes an obstacle for us to find an equivalent MMSNP problem for a given one from MMSNP_2.

However, in Section 4.3 we construct a countably infinite class of structures that can be thought as an infinite extension of an MMSNP sentence. And we manage to show that for any MMSNP_2 sentence there is an equivalent MMSNP sentence of countable length.

It is known that any MMSNP problem is described by a CSP problem on ω-categorical structure. In Section 4.4 we prove that the countably infinite class of structures that is associated with a MMSNP_2 sentence is regular. The notion of regularity for relational structure families is studied by Hubička and Nešetřil in [HN15], where they show that for any regular class of structures $\mathcal{F}$ there exists an ω-categorical structure $\mathfrak{B}$ such that, for any finite structure $\mathfrak{A}$, $\mathfrak{A}$ homomorphically maps to $\mathfrak{B}$ if and only if no structure from $\mathcal{F}$ maps to $\mathfrak{A}$. This implies that any infinite MMSNP problem that we have obtained in Section 4.3 is described by some ω-categorical CSP. Although it is known (see [Mad09]) that MMSNP_2 problems are already described by ω-categorical CSPs, representing these problems as regular families of forbidden structures seems to be an interesting result.

Bodirsky et al. in [BMM18] characterized a dichotomy condition for CSP on ω-categorical structures that are described by MMSNP sentences. We think that a dichotomy for MMSNP_2 can potentially be investigated by a similar approach.

Expanders for MMSNP_2

Apart from reducing to MMSNP, one could try to use an approach similar to Feder and Vardi's proof of $\mathsf{MMSNP} \equiv_p \mathrm{CSP}$. The most difficult part of their proof was to find, for a given structure $\mathfrak{S}$, a structure $\mathfrak{S}'$ that is equivalent to $\mathfrak{S}$ with respect to the CSP and that contains no short cycles. The authors construct $\mathfrak{S}'$ by replacing each vertex of $\mathfrak{S}$ with a large bag of vertices and then, for each tuple of $\mathfrak{S}$, the tuples are uniformly randomly distributed in its preimage. By manipulating the number of tuples in $\mathfrak{S}'$, this structure becomes both sparse and dense at the same time. It is sparse because it contains few short cycles. And it is dense because of the uniform distribution. These two properties are usually discussed when talking about expander graphs. In particular, when Kun derandomised the construction of $\mathfrak{S}'$ in [Kun13], he also used such graphs.

We face a similar task: to find such $\mathfrak{S}'$. In Section 5.3 we study different approaches to construct a desired expander structure. For every approach, we highlight the obstacle that does not let us complete the task.

1.2.2 Investigate CSP extensions

Overview

Motivated by the CSP conjecture, many homomorphism type problems have been introduced and studied under the realm of a dichotomy, e.g., *full homomorphism* [BNP10], *locally injective/surjective homomorphism* [MS10, BKM12], *list homomorphism* [HR11], *quantified CSP* [ZM20], *infinite CSP* [BMM18], *VCSP* [KZ13], etc. In this thesis, we are interested in the *Matrix Partition* Problem introduced in [FHKM03] which finds its origin in combinatorics as other variants of the CSP conjecture, e.g., list or surjective homomorphism.

Trigraphs

A *trigraph* is a pair $\mathfrak{G} = (G, \mathsf{E}^{\mathfrak{G}})$ where $\mathsf{E}^{\mathfrak{G}} \colon G^2 \to \{0, 1, \star\}$. A *homomorphism* between two trigraphs $\mathfrak{G}$ and $\mathfrak{H}$ is a mapping $\mathfrak{h} \colon G \to H$ such that for all $(x, y) \in G^2$, $\mathsf{E}^{\mathfrak{H}}(\mathfrak{h}(x), \mathfrak{h}(y)) \in \{\mathsf{E}^{\mathfrak{G}}(x, y), \star\}$. As any graph is a trigraph, Hell et al. ([FHKM03,FHX07,Hel14]) proposed a way to consider combinatorial problems on graphs as trigraph homomorphism problems, and called them *Matrix Partition Problems* The term *Matrix Partition Problem* is a natural one because any trigraph can be represented by a matrix where each entry is in $\{0, 1, \star\}$, and a trigraph homomorphism is a partition problem where the edges between two parts V_i and V_j are controlled by the entry of the matrix on (i, j). Particularly, any CSP problem on (directed) graphs can be represented as a Matrix Partition Problem, thus the latter is a generalisation of the class CSP. Motivated by the CSP conjecture, and the similarity of Matrix Partition Problem with CSP, Hell et al. [FHX07,Hel14] asks whether Matrix Partition Problems may follow a dichotomy as CSP does.

Motivation

Motivated by the P-time equivalence between general CSP and CSP on directed graphs [FV98], we investigate a similar question for Matrix Partition Problems. But, contrary to CSP, there are several ways to generalise Matrix Partition Problems on relational structures. We first propose to generalise the definition of trigraphs to relational structures, where a tuple can be now labeled $\star$, and as in the trigraph homomorphism, a tuple labeled 0 can be only mapped to tuples labeled 0 or $\star$, similarly for 1-labeled tuples that can be mapped to 1 or $\star$-labeled tuples, and a tuple labeled $\star$ can be only mapped to tuples labeled $\star$. Another generalisation of Matrix Partition Problem concerns the inputs. While in Hell et al.'s definition of the problem the inputs are graphs, we propose to consider instead trigraphs as inputs, for their definition see [HN07]. We denote such new problems by $\mathrm{MP}_\star(\mathfrak{H})$, and the original ones by $\mathrm{MP}(\mathfrak{H})$ where $\mathfrak{H}$ is the target structure of the problem. As in the CSP case, we wonder whether this generalisation is P-time equivalent to trigraph homomorphism. We prove that $\mathrm{MP}_\star$ and MP are P-time equivalent: $\mathrm{MP}_\star \equiv_p \mathrm{MP}$. Hell and Nešetřil in [HN07] provided a probabilistic proof of this equivalence. In this thesis, we make it deterministic. In doing so, we replace any $\star$-labeled tuple by a large enough *Hadamard matrix* [FRW88]. Hadamard matrices are matrices over $\{1, -1\}$ with the property that any large submatrix is not monochromatic. This property of Hadamard matrices and the pigeonhole principle allow us to show the P-time equivalence (see Section 6.2).

Feder and Vardi in [FV98] showed that a CSP over a finite signature is P-time equivalent to a CSP on directed graphs. Bulin et al. in [BDJN15] gave a more detailed proof of this fact and showed that all the reductions are log-space. In Section 6.3.2, we raise similar questions about Matrix Partition Problems. Using the result achieved in Section 6.3.1, we show that any problem in MP over any finite signature is P-time equivalent to a problem in MP on relational structures with one single relation.

We then turn our attention to the P-time equivalence between MP on relational structures with a single relation to MP on directed graphs. While we think that, contrary to the CSP case, MP on relational structures is richer than MP on directed graphs, we fail to prove it. Instead, we analyse the type of reductions used in the CSP case and show that it is unlikely that such reductions work for MP, unless MP $\equiv_p$ CSP. In order to show this, we introduce another generalisation of Matrix Partition Problems, denoted by $\mathrm{MP}_\varnothing$. We first encode any problem in MP by a CSP problem by identifying for each tuple whether it is labeled 1 or 0 (we introduce for each relation R two relations R_0, for 0-labeled tuples, and R_1 for 1-labeled tuples). Therefore, any MP problem is a CSP problem, but restricted to "complete structures", *i.e.*, any tuple should be in either R_0 or in R_1. When we relax this completeness property, we obtain the class of problems $\mathrm{MP}_\varnothing$, where we introduce a new value for tuples, namely $\varnothing$, which can be mapped to any value among $\{0, 1, \star\}$. Firstly, we show in Section 6.2 that $\mathrm{MP}_\varnothing \equiv_p$ CSP, and that the correspondence between the classes of problems is one-to-one. We later use this correspondence to show in Section 6.3.2 that any reduction similar to the one of Bulin et al. cannot prove the P-time equivalence between MP over any finite signature and MP on directed graphs, unless MP $\equiv_p$ CSP. The four values $\varnothing, 0, 1\star$ are ordered in a form of a Boolean lattice. We also show in Section 6.2 that, if relational tuples have values in an arbitrary finite lattice, then the corresponding class of MP problems is embedded into CSP under P-time reductions, and thus has a dichotomy.

A natural way to prove that a problem is in P is to show that it is described by a finite set of obstructions. In the case of CSP, $\mathcal{F}$ is called a *duality set* for an instance $\mathrm{CSP}(\mathfrak{H})$ if for any structure $\mathfrak{G}$, $\mathfrak{G}$ does not homomorphically map to $\mathfrak{H}$ if and only if there is $\mathfrak{F} \in \mathcal{F}$ such that $\mathfrak{F}$ homomorphically maps to $\mathfrak{G}$. It is known that $\mathrm{CSP}(\mathfrak{H})$ has a finite set of obstructions if and only if it is definable by a *first-order formula* [Ats08]. Feder, Hell and Xie proposed in [FHX07] to study Matrix Partition Problems with finite sets of (inclusion-wise minimal) obstructions, that is a graph admits a partition if and only if it does not have an induced subgraph that belongs to a finite family $\mathcal{F}$ of forbidden graphs. They proposed a necessary (but not sufficient) condition for a matrix M to have finitely many obstructions, and Feder, Hell and Shklarsky later showed in [FHS14] that any Matrix Partition Problem has finitely many obstructions if the input consists only of split graphs. In Section 6.4, we show that a Matrix Partition Problem has finitely many inclusion-wise minimal obstructions if and only if there are finitely many of them for the $\mathrm{MP}_\star$ case. We also consider duality sets for Matrix Partition Problems. We show that the following are equivalent for a trigraph $\mathfrak{H}$ (it holds also for relational structures):

1. $\mathrm{MP}(\mathfrak{H})$ has a finite duality set.
2. $\mathrm{MP}(\mathfrak{H})$ has a finite set of inclusion-wise minimal obstructions.
3. $\mathrm{MP}_\star(\mathfrak{H})$ has a finite duality set.
4. $\mathrm{MP}_\star(\mathfrak{H})$ has a finite set of inclusion-wise minimal obstructions.

Apart from it, we study how the finiteness of obstruction sets for the CSPs is related to the finiteness for trigraphs. We demonstrate that if $\mathrm{MP}_{\varnothing}(\mathfrak{H})$ (that is associated with a CSP, see Section 6.2) has a finite set of obstructions, then $\mathrm{MP}(\mathfrak{H})$ has also a finite set of obstructions. We show that the other direction is false by giving an example of a $\star$-graph $\mathfrak{H}$ such that $\mathrm{MP}_{\star}(\mathfrak{H})$ has finitely many obstructions and $\mathrm{MP}_{\varnothing}(\mathfrak{H})$ has an infinite set of obstructions.

Any MP problem can be described by a sentence from MonadicSNP. This means that potentially there exists a logic that is P-time equivalent to MP. Finding such a logic will provide more tools to study MP with respect to a possible dichotomy. In Section 6.5 we provide a fragment of MonadicSNP and show that it strictly contains MP.

Being inspired by the result of Bulatov and Zhuk for CSP, we study polymorphisms on MP instances. In Section 6.6 we define what is a polymorphism in the MP case and provide an MP instance that is solvable in P-time and that does not have a Siggers polymorphism. This means that MP does not have a characterization similar to CSP.

This is not completely surprising as MP can be seen as a CSP with restricted input. There are numerous examples of CSP that are NP-complete, yet become tractable when their input is restricted. For example, when the input has bounded tree-width, see [Gro07]

Chapter 2

Strict NP and its three syntactic fragments

We revisit the three logics from Feder and Vardi's paper [FV98] that extend Monotone Monadic SNP without inequality (MMSNP) by omitting one of the three conditions that define MMSNP. The authors state that none of these classes has a dichotomy, however some of their proofs are given rather as sketches of proofs. We show the absence of a dichotomy for these three fragments of SNP by providing detailed proofs.

Denote by τ a signature that contains the input relation symbols. An *SNP sentence* is an existential second-order τ-sentence with the universal first-order part, *i.e.*, a sentence of the form

$$\exists \mathsf{X}_1, \ldots, \mathsf{X}_s \; \forall \mathbf{x} \; \phi(x_1, \ldots, x_n),$$

where ϕ is a quantifier-free first-order formula over the signature $\tau \uplus \{\mathsf{X}_1, \ldots, \mathsf{X}_s\}$, and $\mathbf{x}$ is the tuple of (first-order) variables $x_1, \ldots, x_n$. Denote the set of the existentially quantified second-order variables $\{\mathsf{X}_1, \ldots, \mathsf{X}_s\}$ by σ. We call τ the *input signature* and σ the *existential signature*. Relation symbols of τ are called *input relations* and relation symbols of σ are called *existential relations*.

Let a τ-sentence $\Phi \in \mathsf{SNP}$ be written in the following form:

$$\exists \mathsf{X}_1, \ldots, \mathsf{X}_s \; \forall \mathbf{x} \; \bigwedge_i \neg(\alpha_i \wedge \beta_i).$$

Here α_i is a conjunction of atomic or negated atomic formulae involving relation symbols of τ and variables from $\mathbf{x}$; and β_i is a conjunction of atomic or negated atomic formulae involving relation symbols of σ and variables from $\mathbf{x}$. Every subformula $\neg(\alpha_i \wedge \beta_i)$ is called a *negated conjunct*. A sentence Φ is called *monotone* if each conjunction α_i does not contain negated atomic formulae.

Remark. We frequently replace a negated conjunct $\neg(\alpha_i \wedge \beta_i)$ with an equivalent implication $(\alpha_i \to \neg\beta_i)$.

An SNP-sentence Φ is called *monadic* if all existential relation symbols M in σ have arity 1.

Denote by MMSNP the class of SNP-sentences that are monotone, monadic, and *without inequality*, that is, they do not contain literals of the form $x \neq y$. Similarly, denote by

- $\mathsf{MonotoneSNP}$ those SNP-sentences that are monotone and without inequality;
- $\mathsf{MonadicSNP}$ those ones that are monadic and without inequality;
- $\mathsf{MMSNP}_{\neq}$ those ones that are monotone and monadic, and possibly with inequality.

If Φ is a τ-sentence of SNP, then *the problem of satisfiability of* Φ, denoted by $\mathrm{SAT}(\Phi)$, is a set of relational τ-structures that satisfy the sentence Φ:

$$\mathrm{SAT}(\Phi) := \{\mathfrak{A} \mid \mathfrak{A} \models \Phi\}.$$

Example 2.0.1. We provide three sentences from each of these classes. Let Φ_1 be the following sentence:

$$\exists \mathsf{A}, \mathsf{B}\ \forall x, y \begin{pmatrix} \neg\big(\mathsf{A}(x) \wedge \mathsf{B}(x)\big) \wedge \neg\big(\neg\mathsf{A}(x) \wedge \neg\mathsf{B}(x)\big) \wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{A}(x) \wedge \mathsf{A}(y)\big) \wedge \neg\big(\mathsf{E}(x,y) \wedge \mathsf{B}(x) \wedge \mathsf{B}(y)\big) \wedge \\ \neg\big(\neg\mathsf{E}(x,y) \wedge \mathsf{A}(x) \wedge \mathsf{B}(y)\big) \wedge \neg\big(\neg\mathsf{E}(x,y) \wedge \mathsf{B}(x) \wedge \mathsf{A}(y)\big) \end{pmatrix}$$

This sentence is not monotone because the last two negated conjuncts contains a negated τ-atom $\neg\mathsf{E}(x,y)$. So it belongs to $\mathsf{MonadicSNP} \setminus \mathsf{MMSNP}$. For a directed graph $\mathfrak{G}$, we have $\mathfrak{G} \models \Phi_1$ if and only if $\mathfrak{G}$ is a complete bipartite graph.

Let Φ_2 be the following sentence:

$$\exists \mathsf{T}\ \forall x, y, z \begin{pmatrix} \neg\mathsf{T}(x,x) \wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \neg\mathsf{T}(x,y)\big) \wedge \\ \neg\big(\mathsf{T}(x,y) \wedge \mathsf{T}(y,z) \wedge \neg\mathsf{T}(x,z)\big) \end{pmatrix}$$

This sentence is not monadic because the existentially quantified relation T is not unary. So it belongs to $\mathsf{MonotoneSNP} \setminus \mathsf{MMSNP}$. For a digraph $\mathfrak{G}$, we have $\mathfrak{G} \models \Phi_2$ if and only if $\mathfrak{G}$ is acyclic.

Let Φ_3 be the following sentence:

$$\forall x, y, z \begin{pmatrix} \neg\big(\mathsf{E}(x,y) \wedge \mathsf{E}(x,z) \wedge y \neq z\big) \wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{E}(z,y) \wedge x \neq z\big) \end{pmatrix}$$

This sentence contains inequalities, so it belongs to $\mathsf{MMSNP}_{\neq} \setminus \mathsf{MMSNP}$. For a digraph $\mathfrak{G}$, $\mathfrak{G} \models \Phi_3$ if and only if the in- and out-degrees of any vertex are not greater than 1. This means that, in this case, $\mathfrak{G}$ is a disjoint union of directed paths and cycles. $\triangle$

Our aim is to reprove the three following theorems. All of them are stated in [FV98]. However, their proofs are not very detailed there: the authors only provide the main ideas how the reductions must look like. Our goal is to make them more precise. The proof of Theorem 2.0.1 is almost the same as the original one, here, we also explicitly show the two reduction directions. The proof of Theorem 2.0.2 uses the same constructions as in [FV98], but we also explicitly write both reductions, and they are non-trivial. The proof of Theorem 2.0.3 is different in the sense that the original proof was only sketched and referred to another paper of Vardi on fragments of Datalog with no explicit result that allowed us to be fully convinced. Our proof is complete and uses the key idea of oblivious Turing machines as proposed in the sketched proof of Feder and Vardi.

Theorem 2.0.1 (Theorem 2 in [FV98]). *Any non trivial MMSNP$_{\neq}$ problem is P-time equivalent to a problem in MonadicSNP.*

Theorem 2.0.2 (Theorem 3 in [FV98]). *Any non trivial MMSNP$_{\neq}$ problem is P-time equivalent to a problem in MonotoneSNP.*

Theorem 2.0.3 (Theorem 1 in [FV98]). *Any non trivial NP problem is P-time equivalent to a problem in MMSNP$_{\neq}$.*

Remark. When we prove each of these three theorems, we do not consider trivial problems: a problem that either accepts any input or rejects any input. Each of the three classes: MonadicSNP, MonotoneSNP and MMSNP$_{\neq}$, contains both of the trivial problems. The problem that accepts all the input is described by the following MMSNP sentence, for example:

$$\exists \mathtt{Red}(\cdot), \mathtt{Blue}(\cdot)\ \forall x\ \neg \mathtt{Red}(x),$$

as we can colour all the input structure elements in $\mathtt{Blue}$. The problem that rejects any input can be described by the following MMSNP sentence:

$$\exists \mathtt{Red}(\cdot)\ \forall x\ \neg\big(\mathtt{Red}(x) \wedge \neg \mathtt{Red}(x)\big).$$

2.1 MMSNP with $\neq$ embeds into Monadic SNP without $\neq$

First, we explain how to construct, for a sentence Φ of MMSNP$_{\neq}$, the corresponding sentence Φ' in MonadicSNP in order to prove Theorem 2.0.1. Let Φ in MMSNP$_{\neq}$ be a τ-sentence as follows:

$$\exists \mathsf{M}_1, \ldots, \mathsf{M}_s\ \forall \mathbf{x}\ \phi(\mathbf{x}),$$

where $\tau = \{\mathsf{R}_1, \ldots, \mathsf{R}_t\}$ is the input relational signature and $\sigma = \{\mathsf{M}_1, \ldots, \mathsf{M}_s\}$ is the existential relational signature consisting of unary relation symbols. We are going to construct a sentence Φ' in MonadicSNP with an input signature τ' and the same existential signature σ.

Construction 1. The new input signature τ' is obtained from τ by adding a new relation symbol $\mathtt{eq}(\cdot,\cdot)$. Firstly, we require that $\mathtt{eq}$ is an equivalence relation. We denote this formula by ϵ_1.

$$\epsilon_1 := \forall x, y, z\Big(\mathtt{eq}(x,x) \wedge \big(\mathtt{eq}(x,y) \rightarrow \mathtt{eq}(y,x)\big) \wedge \Big(\big(\mathtt{eq}(x,y) \wedge \mathtt{eq}(y,z)\big) \rightarrow \mathtt{eq}(x,z)\Big)\Big).$$

Remark. By construction, $\mathtt{eq}$ violates the monotonicity property in the sentence Φ' in ϵ_1.

Any equivalence relation $\sim$ defined on a set A can be extended to an equivalence relation on the set A^k of k-tuples of A as follows: $(x_1, \ldots, x_k)$ is equivalent to $(y_1, \ldots, y_k)$ if, for any i in $[k]$, we have $x_i \sim y_i$. The next condition, denoted by $\epsilon_{2,\tau}$, states that if a τ-relation R contains a tuple $\mathbf{x}$, then it contains every other tuple $\mathbf{y}$ that is equivalent to $\mathbf{x}$ in a sense described above.

$$\epsilon_{2,\tau} := \bigwedge_{\mathsf{R}\in\tau} \forall x_1, \ldots, x_{k_\mathsf{R}}, y_1, \ldots, y_{k_\mathsf{R}}$$
$$\Big(\big(\mathsf{R}(x_1, \ldots, x_{k_\mathsf{R}}) \wedge \mathtt{eq}(x_1, y_1) \wedge \cdots \wedge \mathtt{eq}(x_{k_\mathsf{R}}, y_{k_\mathsf{R}})\big) \rightarrow \mathsf{R}(y_1, \ldots, y_{k_\mathsf{R}})\Big).$$

Similarly, for any σ-relation M, we require that any two $\mathbf{eq}$-equivalent elements agree on M.

$$\epsilon_{2,\sigma} := \bigwedge_{\mathsf{M}\in\sigma} \forall x, y\Big(\big(\mathsf{M}(x) \wedge \mathbf{eq}(x,y)\big) \rightarrow \mathsf{M}(y)\Big).$$

Let ϕ' be the quantifier-free first-order $(\tau \uplus \sigma)$-formula obtained from the quantifier-free part ϕ of Φ, where each literal $x \neq y$ is replaced by $\neg\mathbf{eq}(x,y)$. Let Φ' be the following sentence:

$$\Phi' := \exists \mathsf{M}_1, \ldots, \mathsf{M}_s\big(\epsilon_1 \wedge \epsilon_{2,\tau} \wedge \epsilon_{2,\sigma} \wedge \forall \mathbf{x}\ \phi'(\mathbf{x})\big),$$

By construction, Φ' is monadic and without inequality. This is the end of Construction 1.

Remark. We assume without loss of generality that Φ of $\mathsf{MMSNP}_{\neq}$ is non trivial in the sense that there is a structure that does not satisfy Φ. Such structure is called a *NO instance*. We can assume that Φ always has a NO instance.

In order to prove Theorem 2.0.1, we show that $\mathrm{SAT}(\Phi) \leq_p \mathrm{SAT}(\Phi')$, and then we prove the other direction. Thus, together, the following Lemmas 2.1.1 and 2.1.2 prove Theorem 2.0.1.

Lemma 2.1.1. *Suppose that Φ in $\mathsf{MMSNP}_{\neq}$ and Φ' in $\mathsf{MonadicSNP}$ are as in Construction 1. Then SAT(Φ) reduces in P-time to SAT(Φ').*

Proof. For any τ-structure $\mathfrak{A}$, we construct a τ'-structure $\mathfrak{A}'$, by keeping all relations from τ the same and interpreting $\mathbf{eq}$ as the identity, that is, we set

$$\mathbf{eq}^{\mathfrak{A}'} := \{(x,x) \mid x \in A'\}.$$

This interpretation satisfies ϵ_1, $\epsilon_{2,\tau}$, and $\epsilon_{2,\sigma}$. And also, for all x, y in A, $x \neq y$ if and only if $\neg\mathbf{eq}(x,y)$, so $\mathfrak{A} \models \Phi$ if and only if $\mathfrak{A}' \models \Phi'$. As we can construct $\mathfrak{A}'$ in time polynomial in the size of $\mathfrak{A}$, we have $\mathrm{SAT}(\Phi) \leq_p \mathrm{SAT}(\Phi')$. □

Lemma 2.1.2. *Suppose Φ in $\mathsf{MMSNP}_{\neq}$ and Φ' in $\mathsf{MonadicSNP}$ are as in Construction 1. Then SAT(Φ') reduces in P-time to SAT(Φ).*

Proof. Consider a τ'-structure $\mathfrak{A}'$. We can check in P-time in $|A'|$ whether $\mathfrak{A}' \models \epsilon_1 \wedge \epsilon_{2,\tau}$. If this is false, then $\mathfrak{A}' \not\models \Phi'$, and we reduce $\mathfrak{A}'$ to a fixed NO instance of Φ. Otherwise we reduce $\mathfrak{A}'$ to its $\mathbf{eq}$-quotient. That is, we set $\mathfrak{A} := \mathfrak{A}/\mathbf{eq}$.

We now prove that $\mathfrak{A} \models \Phi$ if and only if $\mathfrak{A}' \models \Phi'$. Suppose that $\mathfrak{A}' \not\models \Phi'$ and, on the contrary, $\mathfrak{A} \models \Phi$. Let $\mathsf{M}_1^{\mathfrak{A}}, \ldots, \mathsf{M}_s^{\mathfrak{A}}$ be the choice of σ-relations for $\mathfrak{A}$ that satisfies the sentence Φ. It is associated with a unique choice $\mathsf{M}_1^{\mathfrak{A}'}, \ldots, \mathsf{M}_s^{\mathfrak{A}'}$ for $\mathfrak{A}'$: such that for any M in σ and $[a]_{\mathbf{eq}}$ in $\mathfrak{A}$, $\mathsf{M}^{\mathfrak{A}}([a]_{\mathbf{eq}}) \leftrightarrow \mathsf{M}^{\mathfrak{A}'}(a)$, its uniqueness is provided by $\epsilon_{2,\sigma}$. By assumption, there is a tuple $\mathbf{a}' = (a'_1, \ldots, a'_n)$ from $\mathfrak{A}'$ such that some negated conjunct $\neg\phi'_i(\mathbf{a}'_i)$ of $\phi'(\mathbf{a}')$ is false. Take the tuple $[\mathbf{a}']_{\mathbf{eq}} := ([a'_1]_{\mathbf{eq}}, \ldots, [a_n]'_{\mathbf{eq}})$. By $\epsilon_{2,\tau}$ and $\epsilon_{2,\sigma}$, the corresponding negated conjunct $\neg\phi_i([\mathbf{a}']_{\mathbf{eq}})$ is false, this is a contradiction.

Suppose now that $\mathfrak{A} \not\models \Phi$ and, on the contrary, $\mathfrak{A}' \models \Phi'$. Let $\mathsf{M}_1^{\mathfrak{A}'}, \ldots, \mathsf{M}_s^{\mathfrak{A}'}$ be the choice of the σ-relations for $\mathfrak{A}'$ that satisfies the sentence Φ'. Let us choose $\mathsf{M}_1^{\mathfrak{A}}, \ldots, \mathsf{M}_s^{\mathfrak{A}}$ such that for any a' in A' and M in σ, we have $\mathsf{M}^{\mathfrak{A}}([a']_{\mathbf{eq}}) \leftrightarrow \mathsf{M}^{\mathfrak{A}'}(a')$. For this choice of $\mathsf{M}_1^{\mathfrak{A}}, \ldots, \mathsf{M}_s^{\mathfrak{A}}$ there is a tuple $\mathbf{a}$ such that some negated conjunct $\neg\phi_i(\mathbf{a})$ is false. Pick any tuple $\mathbf{a}'$ such that $\mathbf{a} = [\mathbf{a}']_{\mathbf{eq}}$. Then, by $\epsilon_{2,\tau}$ and $\epsilon_{2,\sigma}$, the corresponding negated conjunct $\phi'_i(\mathbf{a}')$ is false in $\mathfrak{A}'$, this is a contradiction. □

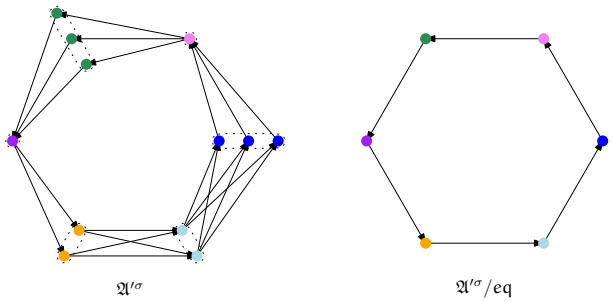


Figure 2.1: On the left, we depict a σ-extension of a structure $\mathfrak{A}'$. On the right, we depict its eq-quotient. Arcs denote binary input relations. Coloured dots denote monadic existential relations. Dotted closed curves denote eq-equivalence classes.

Example 2.1.1. We provide an example of the reduction from Lemma 2.1.2, see Figure 2.1 on page 33.

We can observe that if a map $\mathsf{e}\colon A'/\mathsf{eq}^{\mathfrak{A}'} \to A'$ sends any eq-equivalence class $[a]_{\mathsf{eq}}$ to some element of the same class, then this map is an embedding. In particular, this means that if $\mathfrak{A}'$ satisfies Φ', then $\mathfrak{A}'/\mathsf{eq}$ satisfies Φ' as well. △

2.2 MMSNP with $\neq$ embeds into Monotone SNP without $\neq$

Let Φ in $\mathsf{MMSNP}_{\neq}$ be a τ-sentence with the existential signature $\sigma = \{\mathsf{M}_1, \ldots, \mathsf{M}_s\}$ consisting of unary relation symbols. We construct a τ'-sentence Φ' in $\mathsf{MonotoneSNP}$ with the existential signature σ' such that $\mathrm{SAT}(\Phi) \equiv_p \mathrm{SAT}(\Phi')$, which will prove Theorem 2.0.2.

Construction 2. The new input signature τ' is obtained from τ by adding two relation symbols: $\mathsf{special}(\cdot)$ and $\mathsf{succ}(\cdot,\cdot)$. The new existential signature σ' is obtained from σ by adding three relation symbols: $\mathsf{Marked}(\cdot)$, $\mathsf{Eq}(\cdot,\cdot)$ and $\mathsf{Pred}(\cdot,\cdot)$. We have certain requirements for every new relation.

Every $\mathsf{special}$ element must be Marked:

$$\theta_1 := \forall x\big(\mathsf{special}(x) \to \mathsf{Marked}(x)\big).$$

Any element that is connected (in any direction) to a Marked element by a succ-arc must also be Marked:

$$\theta_2 := \forall x, y\Big(\Big(\big(\mathsf{Marked}(x) \wedge \mathsf{succ}(x,y)\big) \to \mathsf{Marked}(y)\Big)\wedge$$

$$\wedge\Big(\big(\mathsf{Marked}(x) \wedge \mathsf{succ}(y,x)\big) \to \mathsf{Marked}(y)\Big)\Big).$$

The relation Pred must be irreflexive and transitive:

$$\theta_3 := \forall x, y, z\Big(\neg\mathsf{Pred}(x,x)\wedge\Big(\big(\mathsf{Pred}(x,y) \wedge \mathsf{Pred}(y,z)\big) \to \mathsf{Pred}(x,z)\Big)\Big).$$

Any succ-arc is a Pred-arc, so Pred must contain the transitive closure of succ:

$$\theta_4 := \forall x, y\big(\mathsf{succ}(x,y) \to \mathsf{Pred}(x,y)\big).$$

Similarly as in the proof of Theorem 2.0.1, Eq must be an equivalence relation, that is, ϵ_1 must be satisfied:

$$\epsilon_1 := \forall x, y, z \Big(\mathsf{Eq}(x,x) \wedge \big(\mathsf{Eq}(x,y) \rightarrow \mathsf{Eq}(y,x)\big) \wedge \Big(\big(\mathsf{Eq}(x,y) \wedge \mathsf{Eq}(y,z)\big) \rightarrow \mathsf{Eq}(x,z)\Big)\Big).$$

Every two $\mathtt{special}$ elements must be Eq-equivalent:

$$\theta_5 := \forall x, y \Big(\big(\mathtt{special}(x) \wedge \mathtt{special}(y)\big) \rightarrow \mathsf{Eq}(x,y)\Big).$$

The relation Pred must be preserved under replacing each of the elements by an equivalent one:

$$\theta_6 := \forall x, y, z \Big(\Big(\big(\mathsf{Pred}(x,y) \wedge \mathsf{Eq}(y,z)\big) \rightarrow \mathsf{Pred}(x,z)\Big) \wedge$$

$$\wedge \Big(\big(\mathsf{Pred}(x,y) \wedge \mathsf{Eq}(x,z)\big) \rightarrow \mathsf{Pred}(z,y)\Big)\Big).$$

The successors and predecessors of two equivalent elements must also be equivalent with one another:

$$\theta_7 := \forall x, y, x', y' \Big(\Big(\big(\mathsf{Eq}(x,y) \wedge \mathtt{succ}(x,x') \wedge \mathtt{succ}(y,y')\big) \rightarrow \mathsf{Eq}(x',y')\Big) \wedge$$

$$\wedge \Big(\big(\mathsf{Eq}(x,y) \wedge \mathtt{succ}(x',x) \wedge \mathtt{succ}(y',y)\big) \rightarrow \mathsf{Eq}(x',y')\Big)\Big).$$

Also, all relations M from σ keep holding after replacing one of their elements x by another element that is equivalent to x, similarly to $\epsilon_{2,\sigma}$ from Theorem 2.0.1:

$$\theta_8 := \bigwedge_{\mathsf{M} \in \sigma} \forall x, y \Big(\big(\mathsf{M}(x) \wedge \mathsf{Eq}(x,y)\big) \rightarrow \mathsf{M}(y)\Big).$$

Note that we can not require the something similar for the input relations as in $\epsilon_{2,\tau}$ in the proof of Theorem 2.0.1, otherwise, we would violate the monotonicity.

We can suppose without loss of generality that, for an $\mathsf{MMSNP}_{\neq}$ sentence Φ, its quantifier-free part $\phi(\mathbf{x})$ is as follows:

$$\neg\phi_1(\mathbf{x}_1) \wedge \cdots \wedge \neg\phi_n(\mathbf{x}_n),$$

where each $\phi_i(\mathbf{x}_i)$ is of the following form:

$$\Big(\alpha_i(\mathbf{x}_i^{\tau}) \wedge \beta_i(\mathbf{x}_i^{\sigma}) \wedge \gamma_i\Big(\mathbf{x}_i^{\neq}\Big)\Big).$$

Here, α_i is a conjunction of τ-atoms, β_i is a conjunction of σ-atoms or negated σ-atoms, and γ_i is a conjunction of inequalities. There are no literals of the form $x = y$, as every literal of such form can be removed if we replace y with x everywhere in the conjunct. In order to construct the sentence Φ', for which we will later show the P-time equivalence with Φ, we do the following transformations within every negated conjunct $\neg\phi_i$.

- Suppose that the part α_i of ϕ_i contains two τ-atoms $\mathsf{R}_1(\mathbf{x}_1)$ and $\mathsf{R}_2(\mathbf{x}_2)$ such that both tuples $\mathbf{x}_1$ and $\mathbf{x}_2$ share a variable x. In this case, we introduce a new variable x', add an atom $\mathsf{Eq}(x,x')$ to ϕ_i, and replace x with x' in the tuple $\mathbf{x}_2$. Repeat this procedure until no two τ-atoms of α_i share variables.

- After that, within γ_i, we replace all inequalities $x \neq y$ by negated σ-atoms $\neg \mathsf{Eq}(x, y)$.
- For every variable x of ϕ_i, we add an atom $\mathsf{Marked}(x)$ to this conjunct. This means that now we can restrict any input structure on Marked elements. In model theory, the substructure consisting of Marked elements is usually called the *relativised reduct* with respect to the relation Marked, see [Hod97].

For ϵ_1 and $\theta_3, \dots, \theta_8$, that is, for all the introduced formulae that do not describe the Marked relation, we also require that all their variables are Marked, as well as for any negated conjunct of $\phi(\mathbf{x})$. To do this, we add the atom $\mathsf{Marked}(x)$ into the implication body, for every variable x of each rule.

Example 2.2.1. The rule θ_5 that forces two $\mathsf{special}$ elements be Eq-equivalent is transformed to the following one:

$$\theta_5 := \forall x, y \Big(\big(\mathsf{special}(x) \wedge \mathsf{special}(y) \wedge \mathsf{Marked}(x) \wedge \mathsf{Marked}(y) \big) \rightarrow \mathsf{Eq}(x, y) \Big).$$

△

Let ϕ' be obtained from the quantifier-free part ϕ of Φ after doing the procedures described above. Then, the desired sentence Φ' is obtained by requiring all the conditions from above:

$$\exists \mathsf{M}_1, \dots, \mathsf{M}_s, \mathsf{Marked}, \mathsf{Eq}, \mathsf{Pred}\, \forall \mathbf{x} \big(\epsilon_1 \wedge \theta_1 \wedge \dots \wedge \theta_8 \wedge \phi'(\mathbf{x}) \big).$$

This is the end of Construction 2.

We prove a well-known fact which states that any SNP sentence is preserved under taking induced substructures. So that we are sure that without loss of generality the Marked relation can be chosen to be minimal possible by inclusion. That is, the one that contains all $\mathsf{special}$ elements, by θ_1, and all the elements that are connected to a $\mathsf{special}$ element by a sequence of succ-arcs, by θ_2, and contains nothing else.

Lemma 2.2.1. *Let Φ in SNP with some input and existential signatures τ and σ, $\mathfrak{A}$ is a relational τ-structure, and $\mathfrak{B}$ is an induced substructure of $\mathfrak{A}$. Then $\mathfrak{A} \models \Phi$ implies that $\mathfrak{B} \models \Phi$.*

Proof. Suppose that $\mathfrak{A} \models \Phi$. Let $\mathfrak{A}' := (\mathfrak{A}, \mathsf{X}_1^{\mathfrak{A}'}, \dots, \mathsf{X}_s^{\mathfrak{A}'})$ be a σ-expansion of $\mathfrak{A}$ that satisfies the first-order part $\forall \mathbf{x}\, \phi(\mathbf{x})$ of the sentence Φ. Observe that $\forall \mathbf{x}\, \phi(\mathbf{x})$ is a universal first-order formula. Let $\mathfrak{B}' := (\mathfrak{B}, \mathsf{X}_1^{\mathfrak{B}'}, \dots, \mathsf{X}_s^{\mathfrak{A}'})$ be the substructure of $\mathfrak{A}'$ induced on the set B. It is well-known that universal first-order formulae are preserved under taking induced substructures, see [Hod97]. Then $\mathfrak{B}' \models \forall \mathbf{x}\, \phi(\mathbf{x})$. Thus, $\mathfrak{B} \models \Phi$. □

We have split the proof of Theorem 2.0.2 in two parts: Lemma 2.2.2 and Lemma 2.2.4. Together, they imply Theorem 2.0.2.

Lemma 2.2.2. *For any sentence Φ in $\mathsf{MMSNP}_{\neq}$, the problem SAT(Φ) reduces to SAT(Φ'), where Φ' is a sentence in $\mathsf{MonotoneSNP}$ which is obtained from Φ by Construction 2.*

Proof. For a τ-structure $\mathfrak{A}$ with domain $A = \{a_1, \dots, a_n\}$, we construct a τ'-structure $\mathfrak{A}'$ with the same domain $A' := A$ such that, for any R in τ, we have $\mathsf{R}^{\mathfrak{A}} = \mathsf{R}^{\mathfrak{A}'}$, and the other two input relation symbols $\mathsf{special}$ and succ are interpreted as follows:

$$\mathsf{special}^{\mathfrak{A}'} := \{a_1\},\ \mathsf{succ}^{\mathfrak{A}'} := \{(a_i, a_{i+1}) \mid i \in [n-1]\}.$$

For such interpretations there is a unique choice of the existential relations $\mathtt{Marked}$, $\mathtt{Eq}$, and $\mathtt{Pred}$ satisfying $\epsilon_1, \theta_1, \ldots, \theta_7$, namely,

$$\mathtt{Marked}^{\mathfrak{A}'} := A,\ \mathtt{Eq}^{\mathfrak{A}'} := \{(a,a) \mid a \in A\},\ \mathtt{Pred}^{\mathfrak{A}'} := \{(a_i, a_j) \mid i < j\}.$$

Similarly as in the proof of Theorem 2.0.1, for all x, y in A, $x \neq y$ if and only if $\neg\mathtt{Eq}(x,y)$.

There is a one-to-one correspondence between interpretations of the existential relations $\mathsf{M}_1, \ldots, \mathsf{M}_s$ of σ, for $\mathfrak{A}$ and $\mathfrak{A}'$, as these two structures have the same domain. Let $\mathfrak{A}^\sigma$ and $\mathfrak{A}'^{\sigma'}$ be two expansions such that all $\mathsf{M}_1, \ldots, \mathsf{M}_s$ are interpreted identically. If there is a negated conjunct $\neg\phi_i(\mathbf{x}_i)$ of Φ such that, for some tuple $\mathbf{a}_i$, we have $\mathfrak{A}^\sigma \models \phi_i(\mathbf{a}_i)$, then we also have $\mathfrak{A}'^{\sigma'} \models \phi'_i(\mathbf{a}'_i)$, where $\mathbf{a}'_i$ is associated with $\mathbf{a}_i$ similarly as the variables of ϕ'_i are associated with the variables of ϕ_i. Recall that the variables may not be the same because we may have replaced a variable x, that belonged to two distinct τ-atoms, with two variables x and x' now related by $\mathtt{Eq}$. If two variables of ϕ'_i are related by the $\mathtt{Eq}$-atom, then we have to assign the same value to them. Backwards, if $\mathfrak{A}'^{\sigma'} \models \phi'_i(\mathbf{a}_i)$, then, for any two variables of ϕ'_i that are related by an $\mathtt{Eq}$-atom, there is a common element of A that is assigned to them. Then we uniquely define which element of A should be assigned to any variable of ϕ_i, then, for some tuple $\mathbf{a}_i$, we have $\mathfrak{A}^\sigma \models \phi_i(\mathbf{a}_i)$. This means that $\mathfrak{A} \models \Phi$ if and only if $\mathfrak{A}' \models \Phi'$. □

Before proving the other reduction direction, we want to assume without loss of generality that any element a of $\mathfrak{A}$ is forced to be $\mathtt{Marked}$ by θ_1 and θ_2. That is, either a is $\mathtt{special}$ or it is connected to a $\mathtt{special}$ element by a sequence of $\mathtt{succ}$-arcs.

Lemma 2.2.3. *Let $\mathfrak{A}'$ be a τ'-structure. Denote by $\mathfrak{A}' \mid_{\mathsf{Marked}}$ its substructure induced on the elements that are either* special *or connected to a* special *element by a sequence of* succ*-arcs. Then we have $\mathfrak{A}' \models \Phi'$ if and only if $\mathfrak{A}' \mid_{\mathsf{Marked}} \models \Phi'$, where Φ' is as in Construction 2.*

Proof. Lemma 2.2.1 states that every SNP-sentence is closed under embeddings. This means that $\mathfrak{A}' \mid_{\mathsf{Marked}} \models \Phi'$ if $\mathfrak{A}'$ does. Suppose now that $\mathfrak{A}' \mid_{\mathsf{Marked}} \models \Phi'$. Choose the interpretation for the relation symbol $\mathtt{Marked}$ in $\mathfrak{A}$ to consist exactly of the elements from $\mathfrak{A}' \mid_{\mathsf{Marked}}$. By definition, it satisfies both conditions θ_1 and θ_2. All other conditions of Φ' are applied only to $\mathtt{Marked}$ elements. All of them are satisfied in $\mathfrak{A}' \mid_{\mathsf{Marked}}$, so they are also satisfied in $\mathfrak{A}'$. □

Lemma 2.2.4. *Let Φ in $\mathsf{MMSNP}_{\neq}$ and Φ' in* MonotoneSNP *be as in Construction 2. Then SAT(Φ') reduces to SAT(Φ).*

Proof. Let $\mathfrak{A}'$ be a τ'-structure. By Lemma 2.2.3, we can assume without loss of generality that any element a in A' is either $\mathtt{special}$ or connected by a sequence of $\mathtt{succ}$-arcs to a $\mathtt{special}$ element.

It turns out that the structure $\mathfrak{A}$ has a σ'-extension that satisfies $\epsilon_1, \theta_1, \ldots, \theta_7$ if and only if $\mathfrak{A}_\sim := \mathfrak{A}'/\sim$ is a balanced graph with respect to $\mathtt{succ}$-arcs, where $\sim$ is the minimal by inclusion equivalence relation that identifies precisely all $\mathtt{special}$ elements, that is,

$$\forall x, y \in A' \Big(\big(\mathtt{special}(x) \wedge \mathtt{special}(y)\big) \leftrightarrow (x \sim y)\Big).$$

Recall that a directed graph $\mathfrak{G}$ is called *balanced* if it admits a *height function*, *i.e.*, a function $\mathsf{f}\colon G \to \mathbb{Z}$ such that, for any two vertices x, y in G, if there is an arc $\mathsf{E}(x,y)$, then we have $\mathsf{f}(y) = \mathsf{f}(x) + 1$. One can check in P-time if a digraph is balanced, see [BDJN15].

Claim 2.2.5. *$\mathfrak{A}'$ has a σ'-extension that satisfies the conditions $\epsilon_1 \wedge \theta_1 \wedge \cdots \wedge \theta_7$ if and only if $(A_\sim, \mathsf{succ}^{\mathfrak{A}_\sim})$ is a balanced digraph.*

Proof of the Claim. We are going to prove the claim by arguing, first, that a height function yields an interpretation of $\mathtt{Eq}$ and $\mathtt{Pred}$ that satisfy the formulae; and then we argue that if it not balanced, then no choice for $\mathtt{Eq}, \mathtt{Pred}$ can satisfy the formulae.

Suppose $(A_\sim, \mathsf{succ}^{\mathfrak{A}_\sim})$ is balanced; then let $\mathtt{f}_\sim \colon A_\sim \to \mathbb{Z}$ be its height function. Let $\mathtt{f} \colon A' \to \mathbb{Z}$ satisfy

$$\forall x \in A \left(\mathtt{f}(x) = \mathtt{f}_\sim([x]_\sim)\right),$$

where $[x]_\sim$ is the $\sim$-equivalence class that contains x. Construct the σ'-extension $\mathfrak{A}'^{\sigma'}$ as follows. Choose $\mathtt{Eq}^{\mathfrak{A}'^{\sigma'}} := \{(x,y) \in (A')^2 \mid \mathtt{f}(x) = \mathtt{f}(y)\}$, and $\mathtt{Pred}^{\mathfrak{A}'^{\sigma'}} = \{(x,y) \in (A')^2 \mid \mathtt{f}(x) < \mathtt{f}(y)\}$ (recall that we assume without loss of generality that $\mathtt{Marked}^{\mathfrak{A}'^{\sigma'}} = A'$). All the other relations from σ' can be chosen arbitrarily.

ϵ_1 is true as "=" is an equivalence relation on $\mathbb{Z}$, θ_1, θ_2 are satisfied by the assumption, θ_3 is satisfied because "<" is irreflexive and transitive on $\mathbb{Z}$, θ_4 is true because $\mathtt{f}_\sim$ is a height function, θ_5 is true by the construction of $\mathtt{f}$ from $\mathtt{f}_\sim$, θ_6 is true because, for a, b, c in $(\mathbb{Z}, <)$, $a < b$ and $b = c$ imply $a < c$, θ_7 is true because $\mathtt{f}_\sim$ is a height function. We have proved one direction.

For the other direction, suppose that there is a σ'-extension $\mathfrak{A}'^{\sigma'}$ that satisfies $\theta_1, \ldots, \theta_7$. Let $\mathbf{s} = (a_1, \ldots, a_s)$ be a sequence of elements of $\mathfrak{A}'^{\sigma'}$ such that, for any i in $[s-1]$, either $\mathsf{succ}^{\mathfrak{A}'^{\sigma'}}(a_i, a_{i+1})$ or $\mathsf{succ}^{\mathfrak{A}'^{\sigma'}}(a_{i+1}, a_i)$, or $\mathtt{Eq}^{\mathfrak{A}'^{\sigma'}}(a_i, a_{i+1})$. Denote the family of all such sequences by $\mathcal{S}$. Any $\mathbf{s}$ in $\mathcal{S}$ of length s is associated with a tuple $\mathbf{t_s} = (t_1, \ldots, t_{s-1})$ such that, for any i in $[s-1]$,

- if $\mathsf{succ}^{\mathfrak{A}'^{\sigma'}}(a_i, a_{i+1})$, then $t_i = 1$;
- if $\mathsf{succ}^{\mathfrak{A}'^{\sigma'}}(a_{i+1}, a_i)$, then $t_i = -1$;
- if $\mathtt{Eq}^{\mathfrak{A}'^{\sigma'}}(a_i, a_{i+1})$, then $t_i = 0$.

For any $\mathbf{s}$ in $\mathcal{S}$, denote $\mathtt{h}(\mathbf{s}) := \sum_{i=1}^{s-1} t_i$. This function $\mathtt{h}$ computes the difference between the number of forth-coming succ-arcs of the sequence and the number of back-coming ones. In particular, if $\mathbf{s}$ is a succ-path, then $\mathtt{h}$ returns its *net length*.

We want to show that, for any sequence $\mathbf{s} = (a_1, \ldots, a_s)$ in $\mathcal{S}$, if $\mathtt{h}(\mathbf{s}) = 0$, then (a_1, a_s) belongs to $\mathtt{Eq}^{\mathfrak{A}'^{\sigma'}}$. By construction, this property holds for any $\mathbf{s}$ in $\mathcal{S}$ of length 2. Let $\mathbf{s}$ be a sequence of the shortest length such that this property does not hold. Observe that $\mathbf{t_s}$ cannot contain two consecutive 0s, as $\mathtt{Eq}$ is transitive, thus $\mathbf{t_s}$ must contain at least one element from $\{-1, 1\}$. Consider the four following cases for the tuple $\mathbf{s}$ that are displayed on Figure 2.2 on page 37.

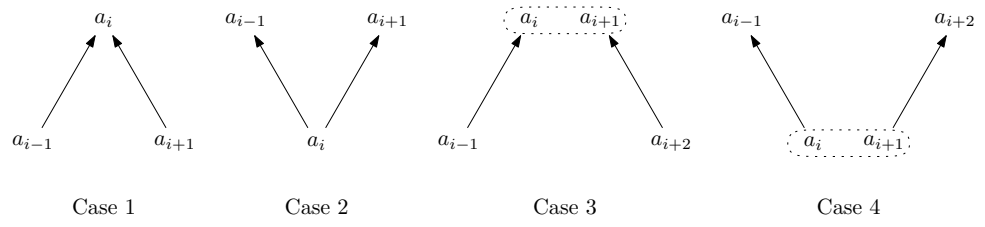


Figure 2.2: The four cases that may occur in $\mathbf{s}$. Dotted closed curves denote $\mathtt{Eq}$-tuples.

If none of these cases happens, then all the $\mathtt{succ}$-arcs of the sequence s have the same direction, so $\mathtt{h}(\mathbf{s}) \neq 0$. Cases 1 and 2 imply that $\mathtt{Eq}^{\mathfrak{A}'^{\sigma'}}(a_{i-1}, a_{i+1})$, by θ_7. Thus, we can delete a_i from $\mathbf{s}$, and obtain a shorter sequence that violates the property. Cases 3 and 4, similarly, imply that $\mathtt{Eq}^{\mathfrak{A}'^{\sigma'}}(a_{i-1}, a_{i+2})$. Thus, if we delete a_i, a_{i+1} from $\mathbf{s}$, then the resulting sequence is shorter and it violates the property.

We want to show that, for any sequence $\mathbf{s} = (a_1, \ldots, a_s)$ in $\mathcal{S}$, if $\mathtt{h}(\mathbf{s}) > 0$, then (a_1, a_s) belongs to $\mathtt{Pred}^{\mathfrak{A}'^{\sigma'}}$. Suppose the opposite, then choose $\mathbf{s}$ to be a sequence of the shortest length that violates the property. If one of the four cases from Figure 2.2 happens, then we can reduce the length of $\mathbf{s}$ without changing the value of $\mathtt{h}$. If none of them happens, then either $\mathbf{t_s}$ consists only of 0s and 1s, or it consists only of 0s and -1s. Without loss of generality, suppose the first case. Then we have $\mathtt{h}(\mathbf{s}) > 0$, and $\theta_3, \theta_4, \theta_6$ provide that (a_1, a_s) belongs to $\mathtt{Pred}^{\mathfrak{A}'^{\sigma'}}$, this is a contradiction.

Suppose that $(A_\sim, \mathtt{succ}^{\mathfrak{A}_\sim})$ is not balanced; then there exist two elements x, y in $A_\sim$ and two sequences $\mathbf{s}_1 = (a_1^1, \ldots, a_{s_1}^1), \mathbf{s}_2 = (a_1^2, \ldots, a_{s_2}^2)$ in $\mathcal{S}$ such that $x = a_1^1 = a_1^2$, $y = a_{s_1}^1 = a_{s_2}^2$, and $\mathtt{h}(\mathbf{s}_1) \neq \mathtt{h}(\mathbf{s}_2)$. Otherwise we could construct a height function $\mathtt{f}$ by setting $\mathtt{f}([x]_\sim) = 0$, for x in $\mathtt{special}^{\mathfrak{A}'}$, and by setting $\mathtt{f}(y) = \mathtt{h}([x]_\sim, \ldots, y)$, for some sequence of $\mathcal{S}$ connecting $[x]_\sim$ and y. Let $\mathbf{s}_2^{-1}$ be the result of reversing $\mathbf{s}_2$, that is, $\mathbf{s}_2^{-1} = (a_{s_2}^2, \ldots, a_1^2)$. For any sequence $\mathbf{s}$ in $\mathcal{S}$, we have $\mathtt{h}(\mathbf{s}) = -\mathtt{h}(\mathbf{s}^{-1})$. Let $\mathbf{s}_1\mathbf{s}_2^{-1}$ be the concatenation of $\mathbf{s}_1$ and $\mathbf{s}_2^{-1}$. Then $\mathtt{h}(\mathbf{s}_1\mathbf{s}_2^{-1}) = \mathtt{h}(\mathbf{s}_1) + \mathtt{h}\left(\mathbf{s}_2^{-1}\right) = \mathtt{h}(\mathbf{s}_1) - \mathtt{h}(\mathbf{s}_2) \neq 0$. This implies that (x, x) belongs to $\mathtt{Pred}^{\mathfrak{A}'^{\sigma'}}$, that is forbidden by θ_3. □

Suppose that $(\mathfrak{A}_\sim, \mathtt{succ}^{\mathfrak{A}_\sim})$ is a balanced graph. Then there is only one way to choose the relations $\mathtt{Eq}$ and $\mathtt{Pred}$ as they depend on the values of the function $\mathtt{h}$. We know that, if $\mathtt{h}(x, \ldots, y) = 0$, then $(x, y) \in \mathtt{Eq}^{\mathfrak{A}'}$; and if $\mathtt{h}(x, \ldots, y) > 0$, then $(x, y) \in \mathtt{Pred}^{\mathfrak{A}'}$. And also we know that, for any pair (x, y) in $(A')^2$, it cannot belong to $\mathtt{Pred}^{\mathfrak{A}'}$ and to $\mathtt{Eq}^{\mathfrak{A}'}$ at the same time, this is forbidden by θ_3 and θ_6. We can conclude now that, for any x, y in A', $(x, y) \in \mathtt{Eq}^{\mathfrak{A}'}$ if and only if $\mathtt{h}(x, \ldots, y) = 0$ and $(x, y) \in \mathtt{Pred}^{\mathfrak{A}'}$ if and only if $\mathtt{h}(x, \ldots, y) > 0$. This implies a uniqueness of the choice of these two relations. So we can assume that either Φ' rejects $\mathfrak{A}'$ in P-time or there is a unique interpretation $\mathtt{Eq}^{\mathfrak{A}'^{\sigma'}}, \mathtt{Pred}^{\mathfrak{A}'^{\sigma'}}$ that satisfies $\epsilon_1 \wedge \theta_1 \wedge \cdots \wedge \theta_7$.

We are ready to construct the corresponding τ-structure $\mathfrak{A}$. It is the τ-reduct of the quotient of $\mathfrak{A}'$ with respect to $\mathtt{Eq}^{\mathfrak{A}'^{\sigma'}}$: $\mathfrak{A} := \left(\mathfrak{A}'/\mathtt{Eq}^{\mathfrak{A}'^{\sigma'}}\right)^\tau$. The condition θ_8 requires that any two elements of the same $\mathtt{Eq}^{\mathfrak{A}'}$-equivalence class have to agree on any σ-relation, thus there is a one-to-one correspondence between choices of σ-relations $\mathsf{M}_1^{\mathfrak{A}}, \ldots, \mathsf{M}_s^{\mathfrak{A}}$ for $\mathfrak{A}$ and choices for $\mathfrak{A}'$.

We need to prove that $\mathfrak{A} \models \Phi$ if and only if $\mathfrak{A}' \models \Phi'$. Let $\mathfrak{A}'^{\sigma'}$ be a σ'-expansion of $\mathfrak{A}'$ that satisfies all the conditions $\epsilon_1, \theta_1, \ldots, \theta_8$. Let $\mathfrak{A}^\sigma$ be the σ-expansion of $\mathfrak{A}$ that is associated with $\mathfrak{A}'^{\sigma'}$. It suffices to show that each negated conjunct $\neg\phi_i(\mathbf{a})$ of the quantifier-free part ϕ of Φ is true in $\mathfrak{A}^\sigma$, for any choice of variables $\mathbf{a}$, if and only if the corresponding negated conjunct $\neg\phi_i'(\mathbf{a}')$ is true in $\mathfrak{A}'^{\sigma'}$, for any choice of variables $\mathbf{a}'$.

Suppose that, for some $\mathbf{a}$ and some negated conjunct $\phi_i(\mathbf{a})$, every atom of $\phi_i(\mathbf{a})$ is true in $\mathfrak{A}^\sigma$. We need to find a tuple $\mathbf{a}'$ such that every atom in $\phi_i'(\mathbf{a}')$ is true in $\mathfrak{A}'^{\sigma'}$. For any atom $\mathsf{R}(\mathbf{b})$ of a τ-relation R, we know that it holds in $\mathfrak{A}$. By the construction of $\mathfrak{A}$, there exists a tuple $\mathbf{b}'$ in $\mathfrak{A}'$ such that $\mathbf{b} = [\mathbf{b}']_{\mathtt{Eq}}$ and that $\mathsf{R}(\mathbf{b}')$ holds in $\mathfrak{A}'$. By the construction of ϕ_i', we know that any variable of ϕ_i' may appear only in one atom with a τ-relation; this means that the choice of the tuple $\mathbf{b}'$ that is associated with $\mathbf{b}$ does not influence

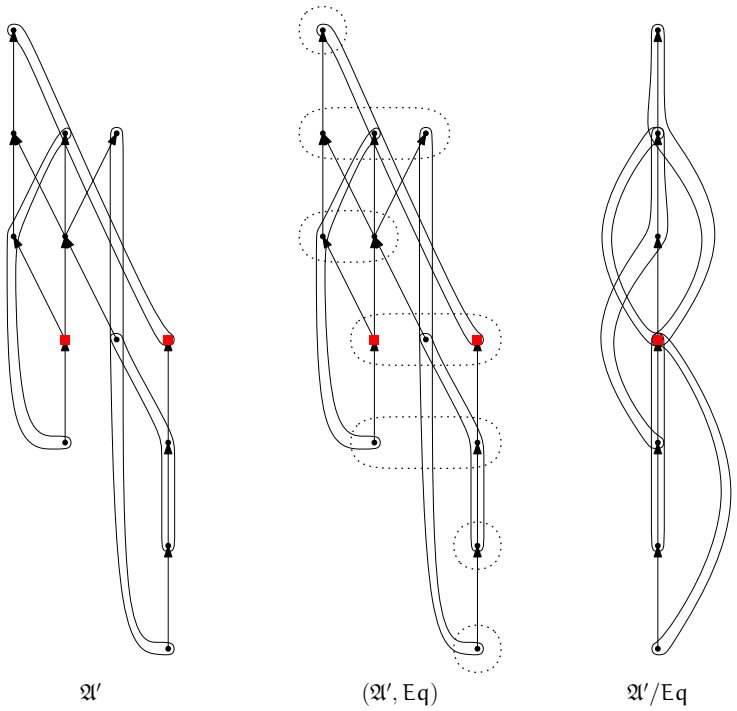


Figure 2.3: From left to right, a τ'-structure $\mathfrak{A}'$, the same structure expanded with the existential equivalence relation $\mathtt{Eq}$, and the $\mathtt{Eq}$-quotient of this structure. Red squares are $\mathtt{special}$ elements, closed curves are τ-relational tuples, black arcs are $\mathtt{succ}$-arcs, dotted closed curves are $\mathtt{Eq}$-equivalence classes.

the choices for other τ-relational atoms of ϕ_i. We can suppose now that we can choose elements of $\mathfrak{A}'$ such that any τ-relational atom of ϕ'_i is true. Consider a σ-relational atom $\mathsf{M}(d)$ of $\phi_i(\mathbf{a})$. Suppose that the element d appears in a tuple $\mathbf{b}$ of some τ-relational atom $\mathsf{R}(\mathbf{b})$: $d = b_i$; then we have already chosen an element of $\mathfrak{A}'$ that is associated with d: it is b'_i of the tuple $\mathbf{b}'$. By θ_8 and, by the construction of $\mathfrak{A}$, b_i belongs to M if and only if $[b_i]_{\mathtt{Eq}}$ belongs to M. Suppose that the element d does not appear in any τ-relational atom of ϕ_i; then we choose any element d' in A' such that $[d']_{\mathtt{Eq}} = d$. We have shown that any τ-relational and any σ-relational atom is satisfied in $\mathfrak{A}'^{\sigma'}$. Recall that, for any a'_1, a'_2 in A', $\neg\mathtt{Eq}(a'_1, a'_2) \Leftrightarrow [a'_1]_{\mathtt{Eq}} \neq [a'_2]_{\mathtt{Eq}}$; thus any atom of ϕ'_i that has the form $\neg\mathtt{Eq}(x, y)$ is satisfied. The rest consists of atoms of the form $\mathtt{Eq}(x, x')$ that are added during the construction of ϕ_i. Both of the elements of any such atom appear in τ-relational atoms, and we know that $[x]_{\mathtt{Eq}} = [x']_{\mathtt{Eq}}$, because they are associated with the same element of $\mathfrak{A}$. We have shown that, for our choice, any atom of ϕ'_i is true in $\mathfrak{A}'^{\sigma'}$.

Let us show the other direction. Suppose that, for some tuple $\mathbf{a}'$, any atom of some ϕ'_i is true in $\mathfrak{A}'^{\sigma'}$. For any element b' of $\mathbf{a}'$, we associate $[b']_{\mathtt{Eq}}$ to it. By the construction of $\mathfrak{A}$, every atom of ϕ_i is true in $\mathfrak{A}^{\sigma}$. □

Example 2.2.2. Have a look at a structure $\mathfrak{A}'$ that is displayed on Figure 2.3 on page 39. Here, the signature τ' is obtained from τ by adding $\mathtt{special}$ and $\mathtt{succ}$ relations, and τ consists of just one ternary relation. By default, there may be more than one $\mathtt{special}$ element, but they belong to the same $\mathtt{Eq}$-equivalence class. Then, their $\mathtt{succ}$-neighbours must also belong to the same $\mathtt{Eq}$-equivalence class, and so on. If $\mathtt{succ}^{\mathfrak{A}'}$ was not a balanced graph, then we would have two elements connected by a $\mathtt{succ}$-arc within the same $\mathtt{Eq}$-equivalence class.

△

2.3 NP embeds into MMSNP with $\neq$

We show that for any nondeterministic Turing machine there exists a P-time equivalent $\mathsf{MMSNP}_{\neq}$ sentence. Feder and Vardi in [FV98] consider only *oblivious* Turing machines. An oblivious Turing machine is defined by the following property: for any two input instances of the same size, the head movements during the machine executions are the same. That is, the head movement does not depend on the symbols written on the tape, it depends only on the size of the input. More formally, a (possibly non-deterministic) Turing machine M_o is called *oblivious* if there exists a function $\mathsf{f}\colon \mathbb{N} \to \mathbb{N}$ such that, for any input $\mathbf{x}$ of size $|\mathbf{x}| = n$, M_o decides $\mathbf{x}$ in precisely $\mathsf{f}(n)$ steps. And that, for every two inputs $\mathbf{x}_1, \mathbf{x}_2$ such that $|\mathbf{x}_1| = |\mathbf{x}_2|$, and, for any moment t in $[\mathsf{f}(n)]$ of each execution, the positions of the head are the same. The concept of oblivious Turing machines is introduced by Pippenger and Fischer in [PF79].

The proof of Theorem 2.0.3 is only vaguely sketched in [FV98]. We provide a complete proof in this section. This proof relies on a specific choice of the way in which we expect an oblivious Turing machine to work. That is, we do not know how to provide a proof of Theorem 2.0.3 that would use [PF79] as a black box reduction from non-deterministic Turing machines to oblivious ones.

Before explaining how NP can be embedded into $\mathsf{MMSNP}_{\neq}$, we argue that without loss of generality one can consider only oblivious Turing machines with a one-way infinite tape. Then, for any oblivious Turing machine, we construct an $\mathsf{MMSNP}_{\neq}$ sentence Φ and then show that the corresponding problems are P-time equivalent.

In the reduction, an input instance of a Turing machine is associated with a structure that looks like a two-dimensional grid, which represents the set of tape configurations at any moment of the execution. That is, the structure shall be used to represent a canvas for the space-time diagram of the oblivious Turing machine. A row consists of cells that are either occupied by the input or visited by the head by this time. When the head reaches the rightmost cell of the tape it does one more step to the right and, thus, increases the row size by 1, as a new cell is visited. The first row is associated with the initial configuration of the tape ($t = 0$): the head is at the leftmost cell, the string of symbols is the initial input string. The next row is associated with the next moment of time ($t = 1$), the head has done one step to the right. The movements of the head are independent of the input contents. So it shall be possible to check that the grid is in the correct shape in $\mathsf{MMSNP}_{\neq}$.

We assume that a Turing machine that we consider is non-trivial, *i.e.*, it neither accepts all input instances nor rejects all of them.

Oblivious Turing machines

For a Turing machine M, we say that the *running time of M depends only on the size of the input* if there is a function $\mathsf{f}\colon \mathbb{N} \to \mathbb{N}$ such that, for any input $\mathbf{x}$ of size $|\mathbf{x}| = n$, M decides $\mathbf{x}$ in precisely $\mathsf{f}(n)$ steps. The following lemma allows us to consider without loss of generality only Turing machines whose running time depends only on the size of the input.

Lemma 2.3.1. *Let M be a nondeterministic Turing machine. Let $\mathsf{f}\colon \mathbb{N} \to \mathbb{N}$ be a function such that M operates within time $\mathsf{f}(n)$. Then there is a nondeterministic Turing machine*

M' that is P-time equivalent to M and whose running time depends only on the size of the input.

Proof. We know that, for some c, k in $\mathbb{N}$, $\mathsf{f}(n) \leq cn^k$. We construct a Turing machine M' that decides an input of M of size n precisely in time $\mathsf{f}'(n)$, for some other function $\mathsf{f}' \colon \mathbb{N} \to \mathbb{N}$. The machine M' has two tapes. The alphabet of the first tape is the same as the alphabet of M. The alphabet of the second tape has m symbols, where $m = 2^k + c'$ such that $m^{\log_2 n} > cn^k$. Such m exists and it is fixed and does not depend on n. Suppose without loss of generality that it consists of the symbols $1, \ldots, m$. Then every execution of M' is as follows:

- The head of the first tape of M' scans it and writes on the second tape the sequence of 1s of length $\lceil \log_2 n \rceil$. This procedure takes the same amount of time for any input on the tape 1 of size n.
- Then, M' does the same things as M for this input. And, for the tape 2, M' acts as a deterministic Turing machine. It consequently writes all the $\lceil \log_2 n \rceil$-sequences consisting of numbers from 1 to m. For m large enough, the execution for the tape 2 always stops after the execution for the tape 1. At the end, it is always in the accepting state.

The machine M' accepts the input if it both tapes answer YES. Otherwise it rejects the input. By construction, it is equivalent to M. And it is well-known, see [Pap94], that a machine with two tapes can be simulated by a machine with one tape. Thus, it is possible to construct an equivalent machine such that its running time depends only on the size of the input. □

The next lemma states that without loss of generality we can assume that the movement of the head of the machine is the same for any two input instances of the same size. For a Turing machine M, we construct an oblivious Turing machine M_o. The movement of head of M_o is always the same: for every iteration of M, the head of M_o walks to the right end of the tape, increases the tape size by 1, and then walks to the left end of the tape, see Figure 2.4 on page 41. If M halts in precisely $\mathsf{f}(n)$ steps, for an input of size n,

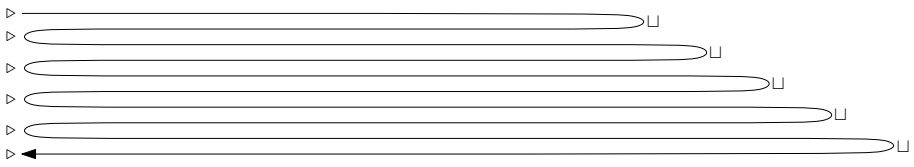


Figure 2.4: The trajectory of the head movement of M_o.

then M_o halts, for any input of size n, in $\mathsf{f}_o(n)$ steps, where:

$$\mathsf{f}_o(n) := 2n + 2(n+1) + \ldots + 2(n + \mathsf{f}(n)) = (\mathsf{f}(n) + 1)(2n + \mathsf{f}(n)).$$

Lemma 2.3.2. *For any nondeterministic Turing machine M with a one-way infinite tape whose running time depends only on the size of the input, there exists a P-time equivalent oblivious nondeterministic Turing machine M_o with a one-way infinite tape.*

Proof. Let M have the alphabet $\Sigma = \{s_1, \ldots, s_k, \triangleright, \sqcup\}$, where $\triangleright$ is the first symbol and $\sqcup$ is the blank symbol; and the states $\mathcal{Q} = \{q_0, \ldots, q_m\}$, where q_0 is the starting state. Firstly, we describe the alphabet Σ_o and the set of states $\mathcal{Q}_o$ of M_o.

- For s in $\Sigma \setminus \{\triangleright, \sqcup\}$, add to Σ_o two symbols: s, s^h. The h-superscript helps to remember the position of the head of M. Instead of one blank symbol in Σ, Σ_o has three symbols: $\sqcup, \sqcup', \sqcup'^h$. $\sqcup$ is written in cells that are not visited by the head of M_o. $\sqcup'$ is written in those cells that are blank but the head of the machine M potentially could have already visited them. $\sqcup'^h$, similarly as s^h, informs the head of M_o that the head of M is at the same cell. At last, add the first symbol $\triangleright$ to Σ_o.

- Any state q_i in $\mathcal{Q}$ is replaced by two states: q_i^L, q_i^R in $\mathcal{Q}_o$ – they highlight the head movement direction (Left or Right). Then, for every pair of states q_i, q_j of $\mathcal{Q}$, we add four states to $\mathcal{Q}_o$: $q^L_{i\to j}$ and $q^R_{i\to j}$ and $q^R_{i\to j(write)}, q_{i\to j(write)}$ – they are used when M_o simulates the transition of M. And also we add to $\mathcal{Q}_o$ states $q_{start}, q_{search}, q_{return}$ that are used at the start of the execution in order to check if the input of M_o is associated with the input of M; q_{start} is the starting state. And, finally, for every newly added state, we add a copy of it that is labeled with the word "error", *e.g.* for q_i^R we add $q_i^{R,error}$. For every state of the type q, the corresponding state of the type *error* does the same function but, in this case, at the end of the execution, the machine always answers YES.

We explain how M_o works. Our informal explanation is followed by an explicit description of M_o.

We assume that for every input tape there is N in $\mathbb{N}$ such that the first cell contains $\triangleright$, next N cells contain symbols of $\Sigma_o \setminus \{\triangleright, \sqcup\}$, and every other cell contains the blank symbol $\sqcup$. We also assume that the head of M_o is in the first cell of the tape and in the state q_{start} when the execution starts.

Informal description of M_o. We want to check if the input of the machine M_o is associated with an input of the machine M. It must have the following form: the first cell must contain the first symbol $\triangleright$, the next n cells must contain an element of $\Sigma \setminus \{\triangleright, \sqcup\}$, and all other cells must contain $\sqcup$.

Check if the input of M_o is exactly like an input of M. At the start of any execution, the head turns into the state q_{search} and moves to the right end of the input in this state. When it reads the blank symbol $\sqcup$, it turns into the state q_{return} and moves back to the initial position in this state. During the left-to-right walk, the head scans every cell and changes the state to q^{error}_{search} if the cell contains an element not from $\Sigma \setminus \{\triangleright, \sqcup\}$. The head edits the bad cell by writing in it a symbol from $\Sigma \setminus \{\triangleright, \sqcup\}$. Thus, when the head returns back to the initial position, the input of the machine is of the right form. But the machine already knows that there is an error, and it shall accept this input at the end of the execution anyway.

Simulate one move of M by two walks: left to right, then right to left. After this scanning and editing procedure, the machine M_o starts the simulation of the execution of M. Every move of the head of M is associated with the walk of the head of M_o towards the right end of the tape and its return to the left end. The head of M_o moves to the right, until it scans the first cell with the blank symbol $\sqcup$. It rewrites $\sqcup$ to $\sqcup'$ in order to mark that during the next walk it does not have to stop at the same cell and can make one more step to the right. Then it returns to the left end of the input string.

If the head is at a cell with a symbol s without the h-superscript, then the head of M_o does not rewrite the symbol and keeps walking in the same direction.

Any cell, except for the first and the last ones, is visited by the head of M_o exactly two times: when it goes left-to-right and right-to-left. Suppose that the head of M_o scans a cell that contains s^h and that the original Turing machine M is in the state q_i now. As it is non-deterministic, there is more than one possible transition from this configuration. For every transition when M decides to move to the right, M_o simulates this transition during the left-to-right walk, and for every transition when M decides to move to the left, M_o does this when the head is returning to the left end. Suppose that the head of M_o is moving towards the right end now, that it is in the state q_i^R, and that it reads s^h from the current cell; then the machine M_o can do one of the following things.

right : M_o decides to simulate one of the possible transitions of M for the configuration (q_i, s) when, during the transition, the head of M moves to the right, *e.g.* $(q_i, s) \to (q_j, s', \to)$. In this case, M_o changes its state to $q^R_{i\to j(write)}$, writes s', and moves one cell to the right. Then it changes to the state $q^R_{i\to j}$ and, for the symbol s_1 of this new cell, writes s_1^h and keeps walking to the right end. Then it changes to the state $q^L_{i\to j}$ and walks towards the left end. And, at the left end, it changes to the state q_j^R and repeats the walk. See Figure 2.5 on page 44.

left : The machine M_o decides to simulate one of the transitions when the head of M walks to the left. In this case, M_o does not change its state q_i^R until it is at the right end of the tape. Then it turns to the state q_i^L, arrives to the cell with s^h during the right-to-left walk and does the steps similar to those from the previous case.

If the head of M is at the first cell, then it always moves to the right. This means that M_o must simulate this movement when it is going to the right from the first cell.

Any state of the type *error* is similar to the corresponding one of the type q. The only difference is that at the end of the execution the machine in an *error*-type state always answers YES.

Explicit description of M_o We give below the transitions of M_o. For the sake of brevity, we do not write the transition rules for the states of the type *error* as they just repeat the transition rules of the usual states, and, once M_o is in an *error* state, it will always be in an *error* state.

Transitions checking the input. If the cell, where the head of M_o is before the run, contains something different from $\triangleright$, then the machine turns to the q^{error}_{search} state, writes $\triangleright$ in the first cell, and starts walking to the right. If the first cell contains $\triangleright$, then the machine writes nothing and just walks to the right while being in the state q_{search}.

$$\forall s \neq \triangleright \big((q_{start}, s) \to (q^{error}_{search}, \triangleright, \to)\big),$$

$$(q_{start}, \triangleright) \to (q_{search}, \triangleright, \to).$$

When the head of M_o walks to the right: either it meets a symbol not from Σ, edits it to a symbol from Σ, changes the state to q^{error}_{search}, and keeps moving to the right; or it edits nothing and keeps moving to the right.

$$\forall s \notin \Sigma \big((q_{search}, s) \to (q^{error}_{search}, s_1, \to)\big), \text{ for some arbitrary } s_1 \in \Sigma,$$

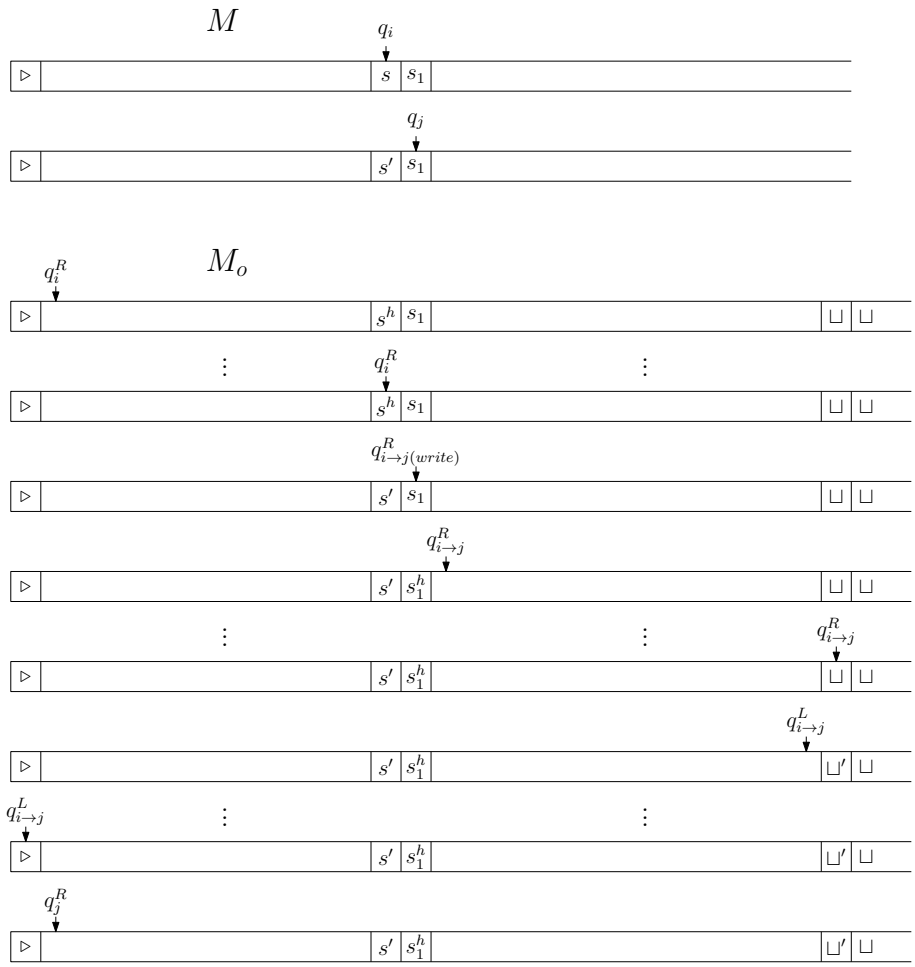


Figure 2.5: A move of the head of a machine M is simulated by a walk of the head of the corresponding oblivious machine M_o.

$$\forall s \in \Sigma \setminus \{\triangleright, \sqcup\} \left((q_{search}, s) \to (q_{search}, s, \to)\right).$$

Now we describe what the head does when it reaches the right end. Observe that, when the machine is at the right end and in the state q_{search}, the tape already has a good form, so we do not need to edit it anymore.

$$(q_{search}, \sqcup) \to (q_{return}, \sqcup', \leftarrow),\ \forall s \in \Sigma_o \setminus \{\triangleright\} \left((q_{return}, s) \to (q_{return}, s, \leftarrow)\right).$$

When the head of M_o returns to the first cell, it simulates the first move of M. Suppose that M has the following transition rule:

$$(q_0, \triangleright) \to \left\{(q_{i_1}, \triangleright, \to), \ldots, (q_{i_k}, \triangleright, \to)\right\};$$

then M_o has the following transition rule:

$$(q_{search}, \triangleright) \to \left\{(q^R_{0\to i_1(write)}, \triangleright, \to), \ldots, (q^R_{0\to i_k(write)}, \triangleright, \to)\right\}.$$

For q_i, q_j in $\mathcal{Q}$, the states $q^L_{i\to j(write)}, q^R_{i\to j(write)}$ are used when the head of M_o simulates the change of the head position of the initial machine M. The head of M_o adds the h-superscript to the symbol of the next cell. Thus, M_o must have the following rules for any symbol of $\Sigma_o \setminus \{\sqcup, \triangleright\}$ without the h-superscript:

$$\forall q_i, q_j \in \mathcal{Q}, s \in \left(\Sigma \cup \{\sqcup'\}\right) \setminus \{\sqcup, \triangleright\} \left((q^R_{i\to j(write)}, s) \to (q^R_{i\to j}, s^h, \to)\right),$$

$$\forall q_i, q_j \in Q, s \in \big(\Sigma \cup \{\sqcup'\}\big) \setminus \{\sqcup, \triangleright\} \ \big((q^L_{i\to j(write)}, s) \to (q^L_{i\to j}, s^h, \leftarrow)\big).$$

Note that by construction, the head of M_o cannot be simultaneously in the state of the type $q_{i\to j(write)}$ and scan the symbol $\sqcup$. The head cannot be in the first cell (that contains $\triangleright$) in a state of the types $q^R_{i\to j}, q^R_{i\to j(write)}$, as there is no cell to the left from the first cell. However, it can be at the first cell in a state of the type $q^L_{i\to j(write)}$. We will describe this case later.

If the head of M_o, being in a state not of the type $q_{i\to j(write)}$, reads a symbol without the h-superscript, then it does not change the symbol and keeps moving further:

$$\forall q_i, q_j \in Q, s \in \big(\Sigma \cup \{\sqcup'\}\big) \setminus \{\sqcup\} \ \big((q^R_i, s) \to (q^R_i, s, \to)\big),$$

$$\forall q_i, q_j \in Q, s \in \big(\Sigma \cup \{\sqcup'\}\big) \setminus \{\sqcup\} \ \big((q^R_{i\to j}, s) \to (q^R_{i\to j}, s, \to)\big),$$

and similarly for the right-to-left direction:

$$\forall q_i, q_j \in Q, s \in \big(\Sigma \cup \{\sqcup'\}\big) \setminus \{\sqcup\} \ \big((q^L_i, s) \to (q^L_i, s, \leftarrow)\big),$$

$$\forall q_i, q_j \in Q, s \in \big(\Sigma \cup \{\sqcup'\}\big) \setminus \{\sqcup\} \ \big((q^L_{i\to j}, s) \to (q^L_{i\to j}, s, \leftarrow)\big),$$

If the head of M is not at the first cell, then, for a rule

$$(q_i, s) \to \big\{(q_{j_1}, s_{j_1}, D_1), \ldots, (q_{j_m}, s_{j_m}, D_m)\big\}$$

of M, we write a rule

$$(q^R_i, s^h) \to (q^R_{i\to j_k(write)}, s_{j_k}, \to)$$

for any k in $[m]$ such that $D_k = \to$; and if $D_k = \leftarrow$, then we write a rule

$$(q^L_i, s^h) \to (q^L_{i\to j_k(write)}, s_{j_k}, \leftarrow),$$

and we also want to let the head of M_o pass s^h when it moves left-to-right:

$$(q^R_i, s^h) \to (q^R_i, s^h, \to).$$

If the head of M_o is in a state of the type $q_{i\to j}$, then it has simulated the current transition of M and is returning now to the first cell in order to change its state to q^R_j.

$$\forall s \in \Sigma_o \setminus \{\triangleright, \sqcup\} \ \big((q^R_{i\to j}, s) \to (q^R_{i\to j}, s, \to)\big),$$

$$\forall s \in \Sigma_o \setminus \{\triangleright, \sqcup\} \ \big((q^L_{i\to j}, s) \to (q_{i\to j}, s, \leftarrow)\big).$$

By construction, M_o reaches the right end in one of the states of the type: $q^R_i, q^R_{i\to j}$, for some i, j. In this case, it pushes the right end one cell to the right by changing $\sqcup$ to $\sqcup'$ and starts moving to the left:

$$(q^R_i, \sqcup) \to (q^L_i, \sqcup', \leftarrow), \ (q^R_{i\to j}, \sqcup) \to (q^L_{i\to j}, \sqcup', \leftarrow).$$

When it returns to the first cell, it turns into the state q^R_j and moves to the right:

$$(q^L_{i\to j}, \triangleright) \to (q^R_j, \triangleright, \to).$$

It remains to explain what M_o does when the head of M moves to the first cell. Consider the following situation, it is displayed on Figure 2.6 on page 46. The head of M

is in the second cell with a symbol s, being in the state q_i. Then it rewrites s to s', moves one cell to the left and arrives to the first cell being in the state q_j. And then it moves one cell to the right and turns into the state q_k. In total, there are two moves of the head of M. As each move of M is associated with a walk of M_o left-to-right and back, the head of M_o must simulate these two moves with two walks. It starts the first walk at the first cell and turns into the state q_i^R. Then it passes the second cell from left to right and does nothing. Then it reaches the end of the tape and turns into the state q_i^L. When it visits the second cell again, it turns into the state $q^L_{i\to j(write)}$, rewrites s^h to s', and moves to the first cell. Then it acts according to the following rule:

$$(q^L_{i\to j(write)}, \triangleright) \to (q^R_{j\to k(write)}, \triangleright, \to).$$

As it arrives to the second cell again, being in the state $q^R_{j\to k(write)}$, it has to rewrite s' to s'^h, turn into the state $q^R_{j\to k}$ and continue walking to the right. When it arrives to the end of the tape, it turns into the state $q^L_{j\to k}$ and walks left. Eventually, it arrives to the first cell and turns into the state q_k^R. With this setting, one move of M is associated with one left-right-left walk, thus, M_o does not violate the obliviousness property.

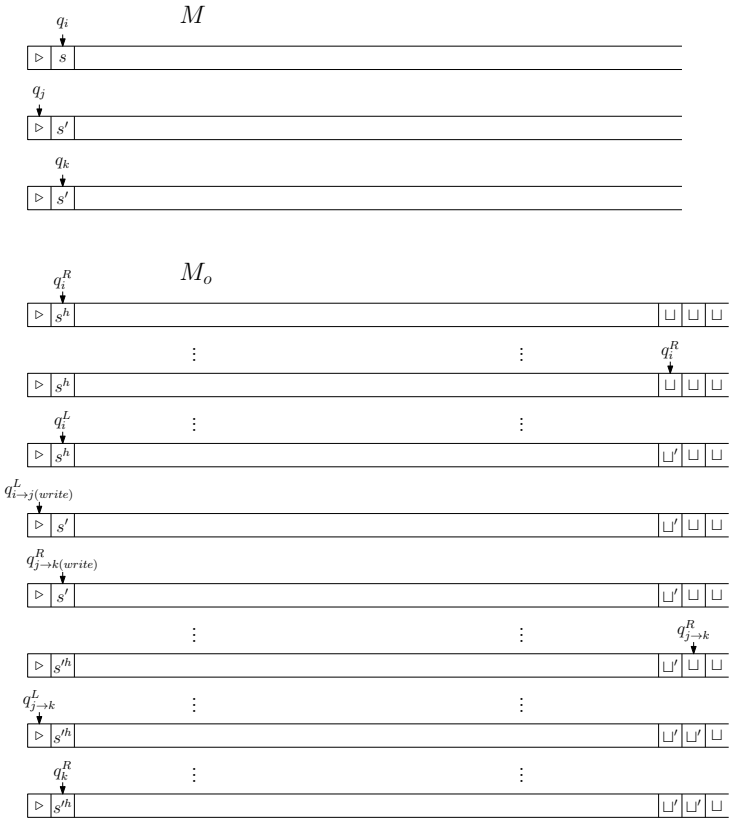


Figure 2.6: Description of the M_o simulation when the head of M is at the first cell.

We have explained how M_o executes. Every execution for the input of size n will take the same amount of time $\mathtt{f}_o(n)$ as the running time of the machine M depends only on the size of the input. The movement of the head of M_o is also the same for all input instances of the same size. We conclude that M_o is oblivious. Moreover, M_o is P-time equivalent to M, because if the input of M_o is not of the form that fits M, then we reduce to a fixed YES instance of M, and if it is of the good form, then we reduce to the same input tape of M. The reduction from M to M_o is the identity. □

Lemmas 2.3.1 and 2.3.2 together imply the following.

Corollary 2.3.3. *For any Turing machine there exists a P-time equivalent oblivious Turing machine with a one-way-infinite tape.*

By Corollary 2.3.3, it is sufficient to show the existence of an $\mathsf{MMSNP}_{\neq}$ sentence that is P-time equivalent to a given oblivious Turing machine M_o with a one-way-infinite tape, for any such M_o.

Construction of the $\mathsf{MMSNP}_{\neq}$ sentence

Let an oblivious Turing machine M_o be the same as in the proof of Lemma 2.3.2.

- M_o has the alphabet $\Sigma_o := \Sigma \uplus \{s^h \mid s \in \Sigma \setminus \{\triangleright, \sqcup\}\} \uplus \{\sqcup', \sqcup'^h\}$.
- It has the states $\mathcal{Q}_o$: $q_{start}, q_{search}, q_{return}$ in $\mathcal{Q}_o$; then, for every q_i in $\mathcal{Q}$, q_i^L, q_i^R in $\mathcal{Q}_o$; and, for every two distinct q_i, q_j in $\mathcal{Q}$: $q^L_{i\to j}, q^R_{i\to j}, q^L_{i\to j(write)}, q^R_{i\to j(write)} \in \mathcal{Q}_o$; and $\mathcal{Q}_o$ contains all the *error*-copies of each of these states.

The input $\mathfrak{A}$ of the $\mathsf{MMSNP}_{\neq}$ sentence Φ to be constructed shall be a canvas for a space-time diagram of a run of M_o. The expressing power of the inequality allows us to forbid the structure $\mathfrak{A}$ to be over-complete, that is, to have more than one tuple of the same relation that are incident to the same element. For example, by using of the inequality, we can require that all the out- and in-degrees of digraph elements are at most one.

However, we cannot forbid the structure $\mathfrak{A}$ to be incomplete, that is, not to have tuples where they ought to be. This is because any $\mathsf{MMSNP}_{\neq}$ formula is monotone with respect to τ-relations. For example, we cannot require that all the out- and in-degrees are precisely one.

The main difficulty of the proof is to come up with an encoding that can be dealt with using the expressivity of $\mathsf{MMSNP}_{\neq}$ when $\mathfrak{A}$ is not of the correct form.

We are planning to construct Φ in $\mathsf{MMSNP}_{\neq}$ such that, for any input structure $\mathfrak{A}$:

- if $\mathfrak{A}$ is over-complete, then $\mathfrak{A} \not\models \Phi$,
- if $\mathfrak{A}$ is not complete enough to simulate a run of M_o, then $\mathfrak{A} \models \Phi$,
- if $\mathfrak{A}$ is in an appropriate form and can simulate a run of M_o that returns YES, then $\mathfrak{A} \models \Phi$,
- if $\mathfrak{A}$ is in an appropriate form and can simulate a run of M_o that returns NO, then $\mathfrak{A} \not\models \Phi$.

By an appropriate form we mean the form of a two-dimensional (space-time) grid. It is formed by horizontal succ-arcs directed from left to right and by vertical next-arcs directed from top to bottom. Every row of this grid represents the tape at some moment of time t. The row below is the tape at the next moment $t+1$, and the row above stands for the moment $t-1$. Every element of a row is connected to the first and the last member of the row with a ternary relation between, this will allow us to require some properties precisely from all the elements of the tape at some fixed moment of time. The contents

of the tape – the symbols, the states, the positions of the head – are encoded with both input and existential relations.

The input relations are represented by a signature τ:

$$\tau = \{\mathsf{s}(\cdot) \mid s \in \Sigma\} \uplus \{\mathsf{start}(\cdot), \mathsf{succ}(\cdot,\cdot), \mathsf{next}(\cdot,\cdot), \mathsf{between}(\cdot,\cdot,\cdot)\},$$

where

- $\mathsf{start}(x)$ represents a constant that highlights the grid element that is associated with the leftmost element of the tape, where the head must be at the start of the execution;
- $\mathsf{succ}(x, y)$ – a horizontal arc, it means that the cell corresponding to y is the right neighbour of the cell corresponding to x on the tape;
- $\mathsf{next}(x, x')$ – a vertical arc, it means that x' represents some cell of the tape at the time t, and x represents the same cell at the time $t + 1$;
- $\mathsf{s}(x)$, for each s in Σ, means that the symbol s is written in the cell corresponding to an element x; these relations matter only in the first row, where they set the initial values for the corresponding existential relations;
- $\mathsf{between}(x, y, z)$ means that y represents an element of a tape that starts at x and ends at z.

The existential relations are represented by a signature σ. All of them must be unary. In order to distinguish τ-relations from σ-relations, we write the name of each σ-relation with an initial upper case letter while every τ-relation starts with a lower case letter. This is convenient because there are both input and existential relations that describe symbols on the tape and that are named identically up to the letter case.

$$\sigma = \{\mathsf{S}(\cdot) \mid s \in \Sigma_o\} \uplus \{\mathsf{Marked}(\cdot), \mathsf{Head}(\cdot)\} \uplus \{\mathsf{Q}(\cdot) \mid q \in \mathcal{Q}\},$$

where

- $\mathsf{S}(x)$, for s in Σ_o, means that the cell corresponding to x contains a symbol s;
- $\mathsf{Marked}(x)$ has the same role as in the proof of Theorem 2.0.2: it marks the elements that need to be considered;
- $\mathsf{Initial}(x)$ highlights the elements corresponding to the initial configuration of the tape, it is used in order to generate all Marked elements;
- $\mathsf{Head}(x)$ means that the head of M_o is at the cell corresponding to x;
- $\mathsf{Q}(x)$ means that the cell corresponding to x belongs to the tape at the moment of time when the machine is in the state q.

Below, we provide the conditions that force an element to be Marked. We require that the elements of the first row are $\mathsf{Initial}$, where the first row consists of all x such that $\mathsf{between}(a, x, a') \wedge \mathsf{start}(a)$. We demand every $\mathsf{Initial}$ element to be Marked. And then we force Marked relation to spread from an element to its bottom next-neighbour if the structure contains all the necessary tuples around this point.

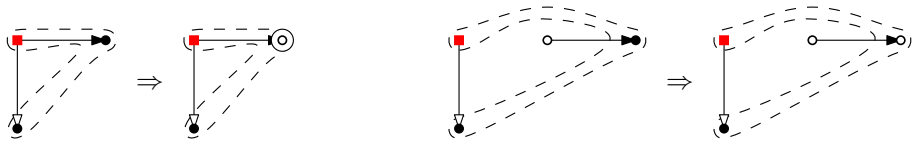


Figure 2.7: A graphical representation of formulae that set the Initial relation on the first row of the grid. Arcs with black heads are succ, arcs with white heads are next, dashed closed curves are between-triples, red squares are start elements, small white circles are Initial elements, large white circles are Head elements. Relations s and S are implicit.

Simultaneously, we set the values for existential relations that are associated with alphabetical symbols as we do not want to have an Initial element that is associated with a cell with no symbol. For any s in Σ, we add the following conjuncts to Φ:

$$\mathsf{start}(a) \wedge \mathsf{succ}(a,x) \wedge \mathsf{next}(a,a') \wedge \mathsf{between}(a,x,a') \wedge \mathsf{s}(x) \rightarrow \\ \mathsf{Initial}(x) \wedge \mathsf{S}(x) \wedge \mathsf{Head}(x) \wedge \mathsf{Q}_{start}(x),$$

$$\mathsf{start}(a) \wedge \mathsf{Initial}(x) \wedge \mathsf{succ}(x,y) \wedge \mathsf{between}(a,y,a') \wedge \mathsf{next}(a,a') \wedge \mathsf{s}(y) \rightarrow \\ \mathsf{Initial}(y) \wedge \mathsf{S}(y),$$

we require that every element of the row highlighted with start must be Initial, except for the start itself. We also want the structure to be complete with respect to between and next, *i.e.*, an element does not need to be Initial if any such tuple is missing. In the same rule we assign the existential symbol relations $\mathsf{S}(\cdot)$ depending on the input symbol relations $\mathsf{s}(\cdot)$, and also demand that the Head is in the leftmost cell (the right succ-neighbour of the start element) in the state Q_{start}. See Figure 2.7 on page 49 for a graphical representation.

Initial elements generate all the Marked elements. Firstly, we require that all the elements of the first row are Marked: $\mathsf{Initial}(x) \rightarrow \mathsf{Marked}(x)$. Secondly, we force the relation Marked to spread from top to bottom via next-arcs. If the neighbourhood of an element y' has one of the four appropriate patterns from Figure 2.8 on page 50 and if its next-predecessor y is Marked, then y' must also be Marked. It will be more convenient to display them as graphs, see Figure 2.8. As the corresponding formulae are rather difficult to read. We explicitly write only the rule corresponding to Figure 2.8a. It goes as follows.

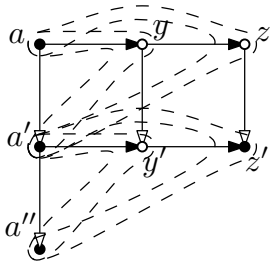
$$\begin{pmatrix} \mathsf{succ}(a,y) \wedge \mathsf{succ}(y,z) \wedge \mathsf{succ}(a',y') \wedge \mathsf{succ}(y',z') \wedge \\ \mathsf{next}(a,a') \wedge \mathsf{next}(y,y') \wedge \mathsf{next}(z,z') \wedge \mathsf{next}(a',a'') \wedge \\ \mathsf{between}(a,y,a') \wedge \mathsf{between}(a,z,a') \wedge \mathsf{between}(a',y',a'') \wedge \\ \mathsf{between}(a',z',a'') \wedge \mathsf{Marked}(y) \wedge \mathsf{Marked}(z) \end{pmatrix} \rightarrow \mathsf{Marked}(y')$$

We want from the relations that represent the states to be a partition:

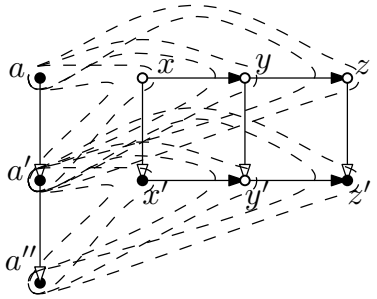
$$\neg \left(\bigwedge_{q \in \mathfrak{Q}} \neg \mathsf{Q}(x) \right) \wedge \bigwedge_{q,q' \in \mathfrak{Q}} \neg \big(\mathsf{Q}(x) \wedge \mathsf{Q}'(x) \big),$$

and that two points at the same row of the grid (*i.e.*, from the tape at the same moment of time) can not have different states, thus, they are in precisely one state:

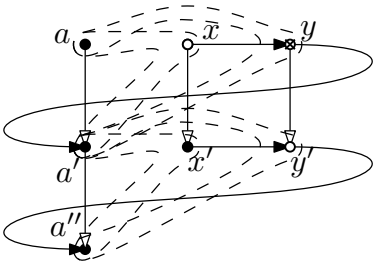
$$\bigwedge_{q,q' \in \mathfrak{Q}} \neg \left(\mathsf{between}(a,x,a') \wedge \mathsf{between}(a,y,a') \wedge \mathsf{next}(a,a') \wedge \mathsf{Q}(x) \wedge \mathsf{Q}'(y) \right).$$



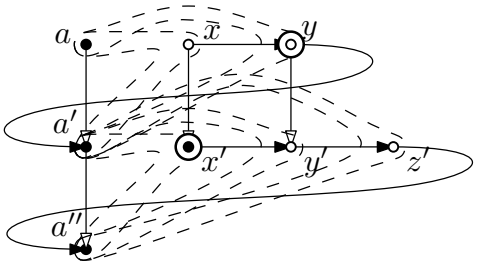
(a) y is the first element of the tape. If y, z are Marked, then y' has to be Marked.



(b) The general case, when y is in the middle of the tape. If x, y, z are Marked, then y' has to be Marked.



(c) y is at the end of the tape. The head is not at y, this is highlighted with a cross. Then, at the next moment of time, the tape will not be augmented. If x, y are Marked, then y' has to be Marked.



(d) y is at the end of the tape. The head is at y, this is highlighted with a large circle. Then the tape will be augmented with z'. Both y', z' have to be Marked and the head has to move upon x'.

Figure 2.8: All the four cases when an element y' is forced to become Marked. All the elements are required to be pairwise distinct, using $\neq$. Black-headed arcs are succ. White-headed arcs are next. Dashed closed curves are between. Small white discs are Marked elements.

The relations S, for s in Σ_o form a partition, as a cell cannot contain two different symbols simultaneously:

$$\neg\left(\bigwedge_{s\in\Sigma_o}\neg\mathsf{S}(x)\right)\wedge\bigwedge_{s,s'\in\Sigma_o}\neg\big(\mathsf{S}(x)\wedge\mathsf{S}'(x)\big).$$

The symbol predicates $\mathsf{S}(\cdot)$, for s in $\Sigma_o \setminus \{\sqcup\}$, are propagated from top to bottom when the predicate Head is not upon the considered element. These symbol predicates can change only when Head is present. This is because a machine can change the contents of the tape only with its head. For every s in $\Sigma_o \setminus \{\sqcup\}$ and the corresponding relation S written in upper-case, we write:

$$\mathsf{Marked}(x)\wedge\neg\mathsf{Head}(x)\wedge\mathsf{S}(x)\wedge\mathsf{next}(x,x')\rightarrow\mathsf{S}(x'). \tag{2.1}$$

We shall add constraints for the Head movement when we provide the formulae that describe the transitions of M_o. There is only one other constraint for the Head relation. It states that at any moment of time there cannot be two different cells of the tape where the Head is at:

$$\mathsf{between}(a,x,a')\wedge\mathsf{between}(a,y,a')\wedge\mathsf{next}(a,a')\wedge\mathsf{Head}(x)\wedge\mathsf{Head}(y)\rightarrow x=y. \tag{2.2}$$

The head of M_o is on the initial cell containing $\triangleright$ at the start. We have one element to the left of $\triangleright$ in our structure so the conjunct will not be $\mathsf{start}(x)\rightarrow\mathsf{Head}(x)$ but something morally similar.

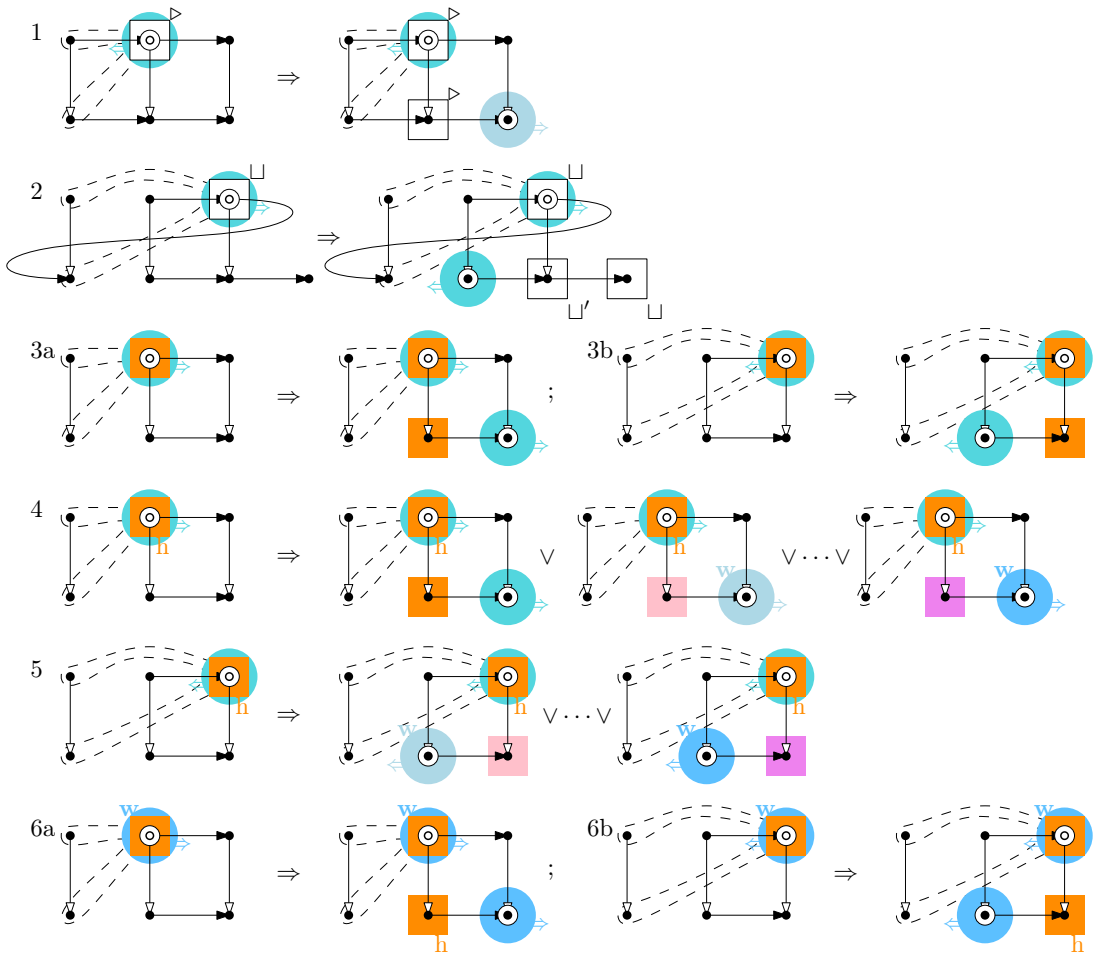


Figure 2.9: A graphical representation of the formulae that describe the transition of M_o. Horizontal black-headed arcs are succ. Vertical white-headed arcs are next. Dashed closed curves are between-tuples. Small white circles are Marked elements. Medium white circles are Head elements. Large squares are existential symbol relations S. If two squares on the same figure have the same colour, then they are associated with the same relation. There is a letter h near the square if and only if the corresponding symbol has the h-superscript. A square with $\triangleright$ is associated with the first symbol relation $\triangleright(\cdot)$. Squares with $\sqcup, \sqcup'$ are relations corresponding to the blank symbol that highlights the end of the tape and to the blank symbol that can be written in the middle of the tape. Large circles stand for the state relations. If two circles have the same colour, then they are associated with the same relation. A letter w near the circle means that the head is in a state of the type $q_{i \to j(write)}$. Symbols $\Leftarrow$ and $\Rightarrow$ near circles denote the direction of the walk. A symbol $\Rightarrow$ between structures means that the left structure represented as the conjunction of all the corresponding atomic formulae implies any atomic formula that is present on the right structure and absent on the left. Symbols $\vee \cdots \vee$ mean that the structure on the left implies one of the structures on the right.

The formulae are displayed on Figure 2.9 on page 51. The figure has 6 lines. Every line is associated with a class of head transitions. Every such class is described by an implication between $(\tau \uplus \sigma)$structures. The caption of Figure 2.9 explains the meaning of every detail of the figure. We describe the transitions one by one. Observe that in the base of every implication the element with the Head is Marked. This means that if the Head comes into a non-Marked element, then no transition rule has to be used.

1. The first line describes the case when the Head is in the first cell. The transition of M_o is always deterministic when the first symbol $\triangleright$ is read. During the transition the Head writes the same symbol and then moves right. The new state is uniquely determined by the state in which the machine was before the transition. The corresponding rule of M_o is of the following form: $(q^L_{i\to j}, \triangleright) \to (q^R_j, \triangleright, \to)$. Below, we write the formula that is associated with this rule:

$$\begin{pmatrix} \mathsf{between}(a,y,a') \wedge \\ \mathsf{succ}(a,y) \wedge \mathsf{succ}(y,z) \wedge \mathsf{succ}(a',y') \wedge \mathsf{succ}(y',z') \wedge \\ \mathsf{next}(a,a') \wedge \mathsf{next}(y,y') \wedge \mathsf{next}(z,z') \wedge \\ \mathsf{Marked}(y) \wedge \mathsf{Head}(y) \wedge \mathsf{S}_\triangleright(y) \wedge \mathsf{Q}^L_{i\to j}(y) \end{pmatrix} \to \begin{pmatrix} \mathsf{S}_\triangleright(y') \wedge \\ \mathsf{Head}(z') \wedge \mathsf{Q}^R_j(z') \end{pmatrix}.$$

   For brevity, we write the formula only for one case, as, for other cases, the formulae can be written similarly.

2. The second line describes the situation when the Head is in the last cell. By the form of the input tape, the last cell is the cell that contains the blank symbol $\sqcup$. When the Head of M_o is reads $\sqcup$, then the transition is deterministic. For any possible state of M_o, the Head rewrites $\sqcup$ to $\sqcup'$ and moves to the left. The new state is uniquely determined by the previous state. As the length of the grid increases every time when the head reaches its end, we also assign the relation $\sqcup(\cdot)$ to the newly added element.

3. The third line contains two rules that are similar to each other. They describe the case when the Head reads a symbol in the middle of the tape that does not contain the h-superscript. This means that the transition is deterministic: the Head does not rewrite the symbol and keeps moving in the same direction. This line also describes the situation when the Head has done the transition and now is returning to the first cell in order to start a new tour.

4. The fourth line is associated with the case when the Head is moving towards the right end and reads a symbol with the h-superscript. This means that it can imitate some of the transitions of the original Turing machine M, where the Head moves to the right. In this case, the Head of M_o has a choice: either to continue moving further and imitate the transition when it visits this cell for the second time, or to imitate one of the transitions. If it decides to imitate them now, then it rewrites the symbol according to the transition, and turns into a state of the type $q_{i\to j(write)}$ in order to add the h-superscript to the next symbol.

5. The fifth line is associated with the case when the Head is moving towards the left end and reads a symbol with the h-superscript. This case is similar to the fourth line, but now the Head cannot do nothing and keep moving. It is the last time when it visits this cell, so it has to do the transition now.

6. The sixth line contains two similar rules. Each of them is associated with the case when the Head is in a state of the type $q_{i\to j(write)}$ and it reads a symbol without the h-superscript. In this case it adds the h-superscript to the symbol highlighting that the head of the original machine M changes its position. After doing this it keeps moving to the same direction.

We have described how to represent the transition rules of M_o in the language of $\mathsf{MMSNP}_{\neq}$. We know that M_o rejects the input if at the end of the execution it is in a rejecting state. We are going to reject the structure if, for some point x, it has both Head and Q_{reject} at this point (for any Q_{reject} that is associated with a rejecting state of $\mathfrak{Q}$):

$$\neg\big(\mathsf{Marked}(x) \wedge \mathsf{Head}(x) \wedge \mathsf{Q}_{reject}(x)\big) \tag{2.3}$$

We have explained all the rules that describe the behaviour of the existential predicates. This is sufficient to show that the problem corresponding to M_o can be reduced to SAT(Φ), where Φ is a sentence from $\mathsf{MMSNP}_{\neq}$ with the existential predicates from σ, the input relations from τ, and the quantifier-free part is the conjunction of all the formulae mentioned earlier in this proof.

Lemma 2.3.4. *Let $\mathcal{L}(M_o)$ be the set of tapes accepted by M_o. Then the problem $\mathcal{L}(M_o)$ reduces in P-time to SAT(Φ).*

Proof. For a tape of length n, we reduce to a grid that consists of $\mathsf{f}(n)$ rows, where f is the running time of M_o. The horizontal arcs are succ, the vertical are next, each element of the row is connected by $\mathsf{between}$ with the start of this row and with the start of the next row. The start of a row is the succ-successor of the end of the precedent row. As every time the the head reaches the right end of the tape, it extends the tape by one cell, we need to augment the length of each row by 1 every time the head reaches its end. So, the first n rows of the grid have length n, the next $2(n+1)$ rows have length $n+1$, and so on. The top-left element of the grid has the predicate start. And the elements of the first row have predicates from $\{s(\cdot) \mid s \in \Sigma_o\}$ that are associated with symbols in the cells of the tape. Such a grid is displayed on Figure 2.10 on page 53.

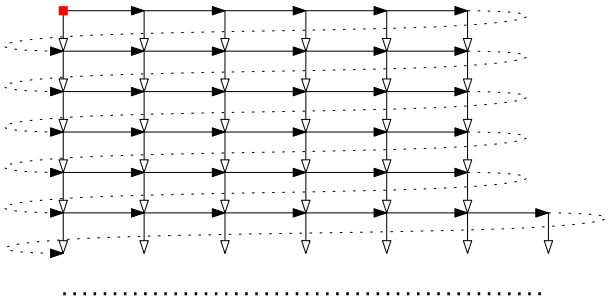


Figure 2.10: An example of the grid that is associated with the tape of M_o. The horizontal arcs are succ, the red square element is the start. All tuples of the relation $\mathsf{between}$ and the symbol relations $\mathsf{s}(\cdot)$ are not drawn in order to make the figure look less complex.

By the construction of Φ and the construction of the grid, every grid element has to be Marked unless it is the leftmost element of a row. For the predicate Head there is only one possible interpretation on the set of Marked elements. See Figure 2.11 on page 54.

As any element that is associated with a cell of the tape is Marked and as the Head movement coincides with the one of M_o, we conclude that this grid is accepted by Φ if

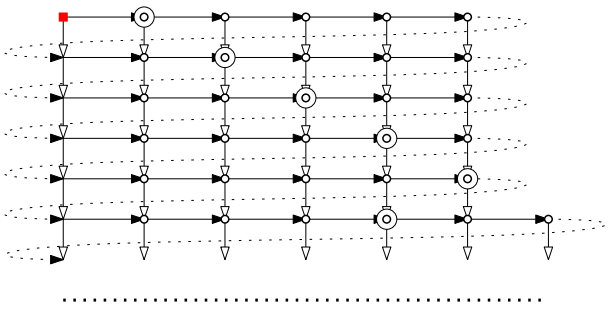


Figure 2.11: The colouring of the grid from Figure 2.10 with the predicates Marked (small white circles) and Head (large white circles).

and only if all the conjuncts from eq. (2.3) are satisfied. That is, if and only if M_o does not reject its input tape. This proves the reduction. □

We need to show the other direction of P-time equivalence. That is, for any relational τ-structure, either we need to solve SAT(Φ) in time polynomial in the structure size or we need to reduce this τ-structure to an equivalent tape of M_o.

In order to do this we have to treat all the input that does not look like a grid from Figure 2.11. We are going to use $\neq$ to forbid all the inputs that are over-complete. Informally speaking, $\neq$ helps us to forbid degrees two or more and to forbid two or more appearances of a predicate.

Let us list these conditions, which are displayed on Figure 2.12 on page 55:

- there can be at most one point highlighted with start:

$$\neg(\mathsf{start}(x) \wedge \mathsf{start}(y) \wedge x \neq y),$$

- every element may have at most one in-neighbour and at most one out-neighbour with respect to succ and next:

$$\neg(\mathsf{succ}(x,y) \wedge \mathsf{succ}(x,y') \wedge y \neq y'),\ \neg(\mathsf{succ}(x,y) \wedge \mathsf{succ}(x',y) \wedge x \neq x'),$$

$$\neg(\mathsf{next}(x,y) \wedge \mathsf{next}(x,y') \wedge y \neq y'),\ \neg(\mathsf{next}(x,y) \wedge \mathsf{next}(x',y) \wedge x \neq x'),$$

- every element can not belong to several tapes at the same time, that is, it can participate in at most one between-tuple:

$$\neg(\mathsf{between}(x,y,z) \wedge \mathsf{between}(x',y,z') \wedge x \neq x'),$$

$$\neg(\mathsf{between}(x,y,z) \wedge \mathsf{between}(x',y,z') \wedge z \neq z'),$$

- if an element is on the 1st or on the 3rd coordinate in a between-tuple, then it can not be on the second coordinate of another between-tuple:

$$\neg\big(\mathsf{between}(a,x,b) \wedge \mathsf{between}(a',a,b')\big),\ \neg\big(\mathsf{between}(a,x,b) \wedge \mathsf{between}(a',b,b')\big),$$

- loops of any kind are forbidden:

$$\neg\mathsf{succ}(x,x),\ \neg\mathsf{next}(x,x'),\ \neg\mathsf{between}(a,x,a),$$

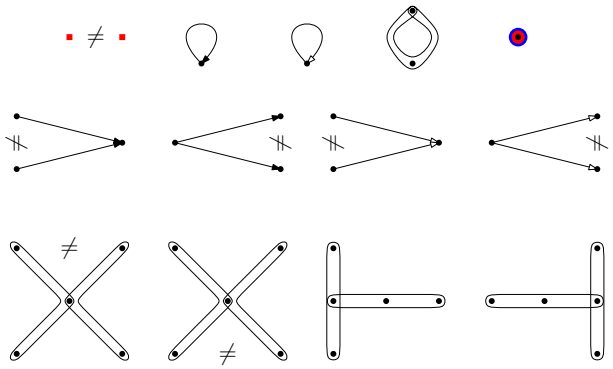


Figure 2.12: A graphical representation of forbidden conjunctions. Red squares are start elements. Arcs with black heads are succ-arcs. Arcs with white heads are next-arcs. Closed curves are between-triples. Red and blue disks are different input symbol relations s, s'.

- a point can belong to at most one symbol relation, that is, for each s, s' in Σ_o, we add:
$$\neg\big(\mathsf{s}(x) \wedge \mathsf{s}'(x)\big).$$

Observe that grids corresponding to tapes satisfy all these sentences. Thus, if we add them to Φ, the reduction from NP to $\mathsf{MMSNP}_{\neq}$ still works correctly.

Informal outline of our reduction from SAT(Φ) to M_o. For some input τ-structure $\mathfrak{A}$, we first check if it omits all the forbidden patterns from Figure 2.12. If it does, then we can understand which elements must be Marked, and at which elements the Head has to be. We shall argue that without loss of generality one needs to consider the minimal by inclusion interpretations of the Marked and Head relations. For an oblivious Turing machine, we know exactly how much time we do need to finish the execution. Thus, we can just see if the size of the substructure induced on Marked elements is large enough to achieve it. If the size is not large enough, then we are certain that the head never reaches a rejecting state, thus, all the conjuncts of Φ are satisfied. In this case we reduce the structure to some input instance of M_o that is accepted by the machine. If the substructure induced on Marked elements is sufficiently large to finish the execution, then the input instance of M_o is provided by the first row of the grid.

Lemma 2.3.5. *SAT(Φ) reduces in P-time to $\mathcal{L}(M_o)$.*

Proof. It takes P-time to check if the input structure $\mathfrak{A}$ of SAT(Φ) satisfies conditions given on Figure 2.12 as they all are first-order formulae. If it is rejected, then we reduce it to some input instance rejected by M_o. If not, then we use the following claim in order to understand which elements of $\mathfrak{A}$ must be Marked and reconstruct the movement of the Head.

Claim 2.3.6. *Among all valid interpretations of the relations* Marked *and* Head *there exist unique minimal by inclusion interpretations, and, if Φ holds on $\mathfrak{A}$, then there exists a valid assignment of σ-relations such that* $\mathsf{Marked}^{\mathfrak{A}}$ *and* $\mathsf{Head}^{\mathfrak{A}}$ *are minimal by inclusion.*

Proof of the Claim. Denote by a_0 an element of $\mathfrak{A}$ such that $\mathsf{start}(a_0)$; if such element does not exist, then no element is forced to be Marked. Denote by $C_m \subseteq A$ the set of elements that are forced to be Marked. It is defined inductively. Firstly, we add to C_m

the right succ-neighbour of a_0; then we put the neighbour of the neighbour and so on, by using the formulae from Figure 2.7 on page 49. Eventually, we add all the elements that are forced to be Initial. Then, we use the formulae from Figure 2.8 on page 50 and add every element that is forced to be Marked by one of the rules. Repeat this until no more element can be added.

We now know that in- and out-degrees of succ- and next-arcs of $\mathfrak{A}$ are at most one. Thus, we can assign to every element of C_m a pair of natural numbers. Assign $(1,0)$ to the right succ-neighbour of the start, assign $(2,0)$ to the neighbour of the neighbour, and so on. For the bottom next-neighbour of (i,j), we assign $(i,j+1)$. One can prove by induction that, for any x, y in C_m with coordinates $(i_x, j_x), (i_y, j_y)$, we have $\mathsf{succ}^{\mathfrak{A}}(x,y)$ if and only if $i_x + 1 = i_y$ and $j_x = j_y$; also, a similar statement is true for next-arcs, and for between-triples as well: x and y are on the same row if and only if $j_x = j_y$.

We show that any interpretation $\mathsf{Marked}^{\mathfrak{A}}$ either contains C_m as a subset or does not satisfy Φ. If there is x in C_m such that $x \notin \mathsf{Marked}^{\mathfrak{A}}$, then choose such x that is added to C_m at the earliest possible moment. Then, as x is not Marked, it does not satisfy one of the formulae from Figures 2.7 and 2.8, as all its predecessors are Marked.

Suppose that $\mathfrak{A} \models \Phi$ and $\mathsf{Marked}^{\mathfrak{A}} = B$ is the interpretation of a solution. We show now that if we set $\mathsf{Marked}^{\mathfrak{A}} = C_m$, then this assignment is still satisfiable. Suppose that it is not satisfiable, thus, some conjunct of Φ is false. It cannot be a rule from Figures 2.7 and 2.8, as C_m satisfies them. And there are no conjuncts with an atomic Marked formula that has an even number of negations before it. This means that all other conjuncts cannot be violated because of the Marked relation. All other relations being the same, we conclude that Φ is also satisfied when $\mathsf{Marked}^{\mathfrak{A}} = C_m$. We have proved that without loss of generality we can assume that the relation Marked is interpreted as the minimal possible one.

Now we proceed similarly for the relation Head. We define a set C_h as follows. Firstly, add the right succ-neighbour of a_0 to C_h (only if it belongs to C_m). Then, follow the rules that are responsible for the head movement, *i.e.*, those that simulate the transitions of the Turing machine displayed on Figure 2.9 on page 51. Recall that any transition rule is fired only if the current Head element is Marked; also recall that the movement of the Head is unambiguous. This means that there is at most one Head element that is not in C_m: it is the last element that is added to C_h.

Similarly as for C_m, for any interpretation $\mathsf{Head}^{\mathfrak{A}}$, either $C_h \subseteq \mathsf{Head}^{\mathfrak{A}}$ or a rule of Φ that describes the transition is not satisfied in $\mathfrak{A}$.

Suppose that there is a satisfiable assignment for the existential relations of Φ, where $\mathsf{Head}^{\mathfrak{A}} = B \supsetneq C_h$. We show now that if we set $\mathsf{Head}^{\mathfrak{A}} = C_h$, then the assignment is still valid. Suppose that it is not valid when $\mathsf{Head}^{\mathfrak{A}} = C_h$; then there is a conjunct of Φ that is not satisfied. This conjunct has to contain a Head-atom that is true when $\mathsf{Head}^{\mathfrak{A}} = B$ and that is false when $\mathsf{Head}^{\mathfrak{A}} = C_h$. This means that this atom has an even number of negations before it. The only conjunct that can be false and that has a Head-atom is eq. (2.1), it means that, if there is no Head at a cell, then this cell contains the same symbol as before at the next moment of time. This means that, when $\mathsf{Head}^{\mathfrak{A}} = B$, it can put the Head where it is not supposed to be and change the symbol written on the tape such that the Head is never in a rejecting state.

When $\mathsf{Head}^{\mathfrak{A}} = C_h$, we need to satisfy this conjunct, so we need to change the interpretation of the relation S such that the new interpretation satisfies the conjunct. That is, if the head is not at the cell, the symbol written in the cell stays the same at the

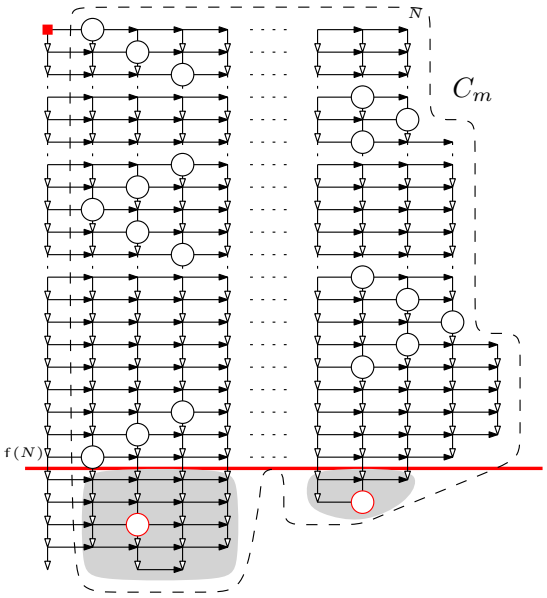


Figure 2.13: The substructure of $\mathfrak{A}$ induced on C_m. The red square is $\mathtt{start}$, horizontal arcs are $\mathtt{succ}$, vertical are $\mathtt{next}$. White circles with black boundaries are the elements of C_h. White circles with red boundaries are the elements of $B \setminus C_h$. Thick red line highlights the moment of time when the execution of M_o stops. There cannot be an element of $B \setminus C_h$ above this red line. The region where they could be is highlighted with grey.

next moment of time. If with this new interpretation $\mathsf{S}^{\mathfrak{A}}$ every conjunct is satisfied, then we are done. Suppose that it is not satisfied; then the only type of negated conjuncts that are false is the one that is associated with M_o being in a rejecting state:

$$\neg\big(\mathsf{Marked}(x) \wedge \mathsf{Head}(x) \wedge \mathsf{Q}_{reject}(x)\big).$$

We use the fact that one can assign a pair of coordinates to any element of C_m. By eq. (2.2), no two points x, y with coordinates $(i_x, j_x), (i_y, j_y)$ can belong to $\mathsf{Head}^{\mathfrak{A}}$ if $j_x = j_y$. By the construction of C_h, we know that, if an element x with coordinates (i_x, j_x) belongs to C_h, then, for any $j' < j_x$ there exists x' in C_h with its second coordinate equal to j'. So, for any b in $B \setminus C_h$ and for any x in C_h, the second coordinates cannot be the same: $j_b \neq j_x$. Thus, in the case when $\mathsf{Head}^{\mathfrak{A}} = B$, changing a symbol on the tape does not change the rejecting state of the machine at the last point of C_h. This is a contradiction as we suppose that Φ is satisfied when $\mathsf{Head}^{\mathfrak{A}} = B$. See Figure 2.13 on page 57 for a graphical representation of this proof.

□

Now we can prove Lemma 2.3.5. We know now that without loss of generality we can interpret Marked and Head as C_m and C_h. We can construct these two sets in time linear of the size of $\mathfrak{A}$. We know that M_o decides an input of size n in exactly $\mathtt{f}(n)$ steps. The size of the input is equal to $N = \max\{i \mid (i, 0) \in C_m\}$. Then we see if $C_h \cap C_m$ contains an element with its second coordinate equal to $\mathtt{f}(N)$. If it does, then we know that the structure is large enough to simulate the execution of M_o, and we associate it with a string of symbols written in cells $(1, 0), \ldots, (N, 0)$. Otherwise any element of $C_h \cap C_m$ has a too small second coordinate, so no negated conjunct of the following type can be false:

$$\neg\big(\mathsf{Marked}(x) \wedge \mathsf{Head}(x) \wedge \mathsf{Q}_{reject}(x)\big).$$

This means that $\mathfrak{A} \models \Phi$, and we associate with $\mathfrak{A}$ some YES instance of M_o.

□

Together, Lemmas 2.3.4 and 2.3.5 prove Theorem 2.0.3.

Why oblivious Turing machines are necessary? In this paragraph we explain why the proof of Theorem 2.0.3 works only for oblivious Turing machines. We use this property in the proof of Lemma 2.3.5, where we reduce $\mathsf{MMSNP}_{\neq}$ back to NP. Suppose that the Turing machine is not oblivious; then the head movements and the execution time can be arbitrary. We can make the input to look like a two-dimensional grid. But we cannot manipulate its height (the number of rows) and width (the number of columns). If the Turing machine is oblivious, then we know the exact head movement and the time of the execution. Consider two grids that are displayed on Figure 2.14. If the head movement is oblivious, then it will perform as we expect, for every grid that is in a good form. If the grid is too short in terms of height and width, then the execution will not be finished, then the rejecting state will never be reached, this means that such short input will never be rejected. However, if the movement is not oblivious, then, for some input string, it can be rejected arbitrarily quickly. This means that we cannot predict if a short grid is sufficient to imitate the whole execution leading to a rejecting state. The right two-dimensional grid is an example of a short grid: the red dashed line marks the boundary, everything beyond it is deleted. The white circles mark the head movement. The execution time is small enough to reject the input encoded in the first row.

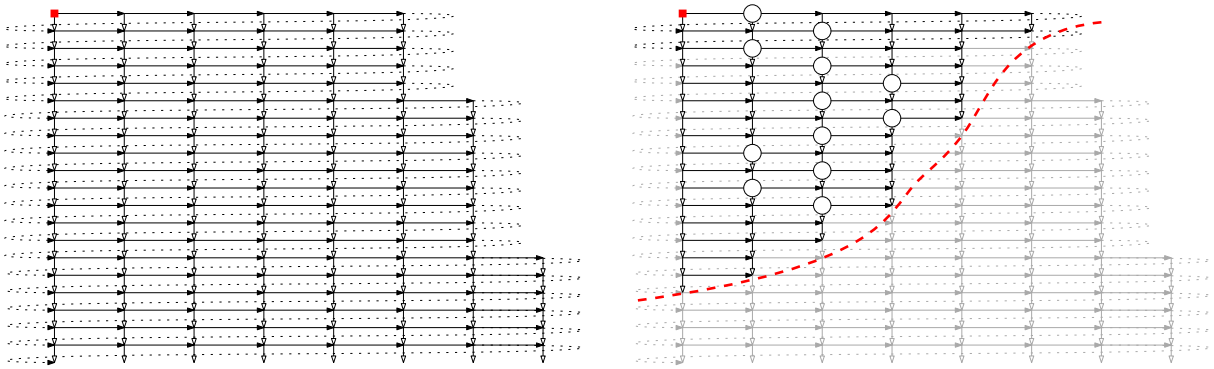


Figure 2.14: On the left, there is a space-time grid in a good form. On the right, there is a space-time grid in a bad form.

Chapter 3

Guarded extensions of MMSNP

We consider extensions of MMSNP that allow for a limited relaxation of some of the constraints imposed on MMSNP, namely absence of inequalities and monotonicity in the input signature. This limitation on the relaxation of a constraint involves guardedness. We describe two such guarded fragments of $\mathsf{MMSNP}_{\neq}$ and MonotoneSNP that both extend MMSNP. The objective is to study these two classes with respect to having a dichotomy. We first show that MMSNP with guarded inequalities has a dichotomy. The guarded fragment of MonotoneSNP, denoted by GMSNP, is a well-studied class, see [BtCLW14]. We then show that, for any GMSNP problem over any signature there is a P-time equivalent MMSNP_2 problem over a signature consisting of a unique relation symbol. Here, MMSNP_2 is a fragment of GMSNP that extends MMSNP by being able to colour tuples as well as vertices.

3.1 Dichotomy for MMSNP with guarded inequalities

Let τ be an input signature and $\sigma = \{\mathsf{M}_1, \dots, \mathsf{M}_s\}$ be the signature of the existentially quantified unary relations.

Definition 1. The *guarded MMSNP with* $\neq$ ($\mathsf{GMMSNP}_{\neq}$) is an extension of MMSNP such that every sentence Φ in $\mathsf{GMMSNP}_{\neq}$ is of the following form:

$$\exists \mathsf{M}_1, \dots \mathsf{M}_s \, \forall \mathbf{x} \bigwedge_i \neg(\alpha_i \wedge \beta_i \wedge x_1 \neq x_1' \wedge \cdots \wedge x_{k_i} \neq x_{k_i}'),$$

where, for each i, α_i is a conjunction of non-negated τ-atoms, β_i is a conjunction of σ-atoms and negated σ-atoms, and, for each inequality $x_j \neq x_j'$ of the ith negated conjunct, there exists a τ-atom in α_i that contains both x_j and x_j'.

Example 3.1.1. Let $\tau = \{\mathsf{E}(\cdot,\cdot)\}$ be the directed graph signature. The following first-order τ-sentence belongs to $\mathsf{GMMSNP}_{\neq}$:

$$\forall x, y \, \neg\big(\mathsf{E}(x,y) \wedge x \neq y\big).$$

This sentence describes the class of graphs where every arc is a loop. Such class is not closed under inverse homomorphisms as any directed graph can be mapped to a loop. As any MMSNP problem is closed under inverse homomorphisms, see *e.g.* [Bod21], we conclude that $\mathsf{GMMSNP}_{\neq}$ is strictly more expressive than MMSNP. △

The goal of this section is to show that such extension of MMSNP has a dichotomy. For simplicity, we consider the case $\tau = \{\mathsf{R}\}$, where R has arity n. The proof for an arbitrary finite signature τ is similar, we can independently do the same steps for any relation of τ as we do for R in this proof.

Lemma 3.1.1. *Any GMMSNP$_{\neq}$ sentence Φ is logically equivalent to a GMMSNP$_{\neq}$ sentence Ψ such that, for any negated conjunct $\neg\psi_i$ of Ψ and for any two different variables x, y that appear within some R-atom of ψ_i, this conjunct contains the inequality $x \neq y$.*

In the following, we explain how to obtain the sentence Ψ from the statement of Lemma 3.1.1.

Construction 3. For every negated conjunct of Φ and for any two variables that appear in the same τ-atom, replace this negated conjunct with two negated conjuncts: the first one is obtained from the original one by adding the inequality $x \neq y$ to the conjunction, the second one is obtained from the original one by replacing every occurence of the variable y with the variable x. Denote this new τ-sentence by Ψ.

Example 3.1.2. If Φ contains the following negated conjunct:

$$\neg\big(\mathsf{R}(x,y) \wedge \mathsf{M}_1(x) \wedge \mathsf{M}_2(y)\big),$$

then this negated conjunct is replaced by the two following negated conjuncts:

$$\neg\big(\mathsf{R}(x,y) \wedge \mathsf{M}_1(x) \wedge \mathsf{M}_2(y) \wedge x \neq y\big) \wedge \neg\big(\mathsf{R}(x,x) \wedge \mathsf{M}_1(x) \wedge \mathsf{M}_2(x)\big).$$

$\triangle$

Remark. This transformation is similar to one used in the proof of Theorem 3 in [FV03]. The theorem states that, for classes of finite structures closed under inverse homomorphisms, MonadicSNP is as expressive as MMSNP.

Proof of Lemma 3.1.1. Let Φ and Ψ be as in Construction 3, and $\mathfrak{A}$ be a τ-structure. Suppose that $\mathfrak{A} \not\models \Phi$; then for any σ-expansion there exists a negated conjunct $\neg\phi_i$ of Φ that is false for some assignment of elements of A to the conjunct variables. Among the negated conjuncts of Ψ that are obtained from ϕ_i we choose the one that is associated with this assignment: if two variables x, y of ϕ have the same element a in A assigned to them, then x and y must be identified in the corresponding conjunct of Ψ; if two different elements a, a' in A are assigned to x and y, then the corresponding conjunct of Ψ must contain the inequality $x \neq y$. By the construction of Ψ, such a conjunct exists. If $\mathfrak{A} \not\models \Psi$, then for any σ-expansion there is a negated conjunct of Ψ that is false for some assignment. It is associated with some negated conjunct $\neg\phi_i$ of Φ, and, by construction, ϕ_i is also false for the same assignment. We conclude that $\mathfrak{A} \models \Phi \leftrightarrow \mathfrak{A} \models \Psi$, for any τ-structure $\mathfrak{A}$. $\square$

Consider any sentence Φ of GMMSNP$_{\neq}$. By Lemma 3.1.1, we assume that, without loss of generality, for any two variables x, y that appear within the same τ-atom in some negated conjunct, this negated conjunct contains the inequality $x \neq y$.

We need to introduce some necessary notations. For n in $\mathbb{N}$, denote by $\epsilon(n)$ the number of equivalence relations on a set of n elements: $\sim_1, \ldots, \sim_{\epsilon(n)}$. For each $\sim_k$, denote by $n_k := |[n]/\sim_k|$ the number of equivalence classes of this relation. Observe that every

n-tuple $\mathbf{x}$ is associated with exactly one equivalence relation $\sim_k$ on the set $[n]$ such that $x_i = x_j$ if and only if $i \sim_k j$. If $\sim_k$ is associated with $\mathbf{x}$ then we say that this tuple has *equivalence type* k. For every equivalence class $\{x_{c_1}, \ldots, x_{c_l}\}$ of an equivalence relation $\sim_k$, denote it by $[c]_{\sim_k}$, where $c = \min\{c_1, \ldots, c_l\}$ – the smallest number of this set. Then, introduce a linear ordering $\prec_k$ on the set $[n]/\sim_k$ by setting $[x]_{\sim_k} \prec_k [y]_{\sim_k}$ if $x < y$.

Definition 2. For a set X, define a function $\mathfrak{p}\colon \biguplus_{n=1}^{\infty} X^n \to \biguplus_{n=1}^{\infty} X^n$ as follows. Let $\mathbf{x} = (x_1, \ldots, x_n)$ in X^n be a n-tuple of equivalence type k, for some n in $\mathbb{N}$. Let $[s_1]_{\sim_k}, \ldots, [s_{n_k}]_{\sim_k}$ be the $\sim_k$-equivalence classes such that $[s_i]_{\sim_k} \prec_k [s_j]_{\sim_k}$ if and only if $i < j$. Then we say that $\mathfrak{p}(\mathbf{x}) := (x_{s_1}, \ldots, x_{s_{n_k}})$ belongs to X^{n_k}.

Example 3.1.3. Informally, the function $\mathfrak{p}$ removes an element from a tuple if it is not its first occurence. Consider a 3-tuple $\mathbf{t} = (y, x, x)$, it is associated with the equivalence relation $\sim_4$ from Figure 3.1 on page 61, the equivalence classes of $\sim_4$ on the set $[3] = \{1, 2, 3\}$ are $[1]_{\sim_4} = \{1\}$ and $[2]_{\sim_4} = \{2, 3\}$. Then $\mathfrak{p}(\mathbf{t}) = (y, x)$ because y is on the first coordinate of $\mathbf{t}$, x is on the second, and $[1]_{\sim_4} \prec_4 [2]_{\sim_4}$. △

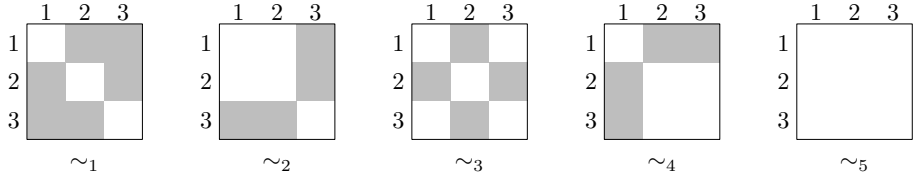


Figure 3.1: All the 5 equivalence relations on the 3-element set depicted as 3×3 logical matrices (white cells stand for ones, grey cells stand for zeros.) That is, for i, j in $[3], k$ in $[\epsilon(3)]$, the cell (i, j) of the kth matrix is white if and only if $i \sim_k j$.

In the following construction we explain how to obtain an equivalent MMSNP sentence Φ', for a given GMMSNP$_{\neq}$ sentence Φ, in order to prove Theorem 3.1.2 on page 62.

Construction 4. We need to find, for any Φ in GMMSNP$_{\neq}$, an MMSNP sentence Φ' over some input and existential signatures τ' and σ' such that $\mathrm{SAT}(\Phi) \equiv_p \mathrm{SAT}(\Phi')$. At first, we set $\sigma' = \sigma$ to be the same signature. In order to construct the new input signature τ', we consider any R in τ that has arity n, and we add to τ' new relation symbols $\mathsf{R}_1, \ldots, \mathsf{R}_{\epsilon(n)}$. Each R_i is associated with one of the $\epsilon(n)$ equivalence relations $\sim_i$ on the set $[n]$, the arity of R_i equals to n_i – the number of equivalence classes of $\sim_i$.

Example 3.1.4. If R is ternary, then we add to τ' five relation symbols:

$$\mathsf{R}_1(\cdot,\cdot,\cdot), \mathsf{R}_2(\cdot,\cdot), \mathsf{R}_3(\cdot,\cdot), \mathsf{R}_4(\cdot,\cdot), \mathsf{R}_5(\cdot).$$

The corresponding equivalence relations on the set $[3] = \{1, 2, 3\}$ are shown on Figure 3.1. △

We are going to describe how to construct the MMSNP sentence Φ' from Φ. At first, we get rid of the inequalities as they are not allowed in MMSNP. And then we show how to represent them with the relations of τ'.

- Firstly, for every negated conjunct $\neg\phi_i(\mathbf{x}_i)$, we delete all the inequalities from it. Recall that, before this procedure, every negated conjunct contained an inequality $x \neq y$ for every pair of distinct variables x, y that appeared together in some relational τ-tuple.

- After doing that, change every τ-atom in any negated conjunct $\neg\phi_i(\mathbf{x}_i)$ to the corresponding τ'-atom. Consider a relational τ-tuple $\mathsf{R}(\mathbf{x})$, where $\mathbf{x}$ is an n-tuple of equivalence type k. Then, replace $\mathsf{R}(\mathbf{x})$ with $\mathsf{R}_k\big(\mathbf{p}(\mathbf{x})\big)$.

- Finally, we require that an element cannot appear in a τ'-relational tuple more than once. In order to do this, for any R_k in τ' and for any $i < j$ in $[n_k]$, we add to Φ' the following negated conjunct consisting of one atom:

$$\neg\mathsf{R}_k(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_{j-1}, x_i, x_{j+1}, \ldots, x_{n_k}), \tag{3.1}$$

  that is the elements at the ith and the jth coordinates are the same.

This is the end of Construction 4.

Example 3.1.5. Let Φ in $\mathsf{GMMSNP}_{\neq}$ describe bipartite directed graphs, but loops are not prohibited. It can written as follows:

$$\exists \mathsf{M}_1, \mathsf{M}_2\ \forall x, y \begin{pmatrix} \neg\big(\neg\mathsf{M}_1(x) \wedge \neg\mathsf{M}_2(x)\big) \wedge \neg\big(\mathsf{M}_1(x) \wedge \mathsf{M}_2(x)\big) \wedge \\ \neg\big(\mathsf{M}_1(x) \wedge \mathsf{M}_1(y) \wedge x \neq y \wedge \mathsf{E}(x,y)\big) \wedge \\ \neg\big(\mathsf{M}_2(x) \wedge \mathsf{M}_2(y) \wedge x \neq y \wedge \mathsf{E}(x,y)\big) \end{pmatrix}$$

Observe that all the negated conjuncts contain inequalities for every pair of distinct variables. So we can follow the stages from Construction 4. At first, we delete all the inequalities from Φ and the sentence is now in the following form:

$$\Phi_1 := \exists \mathsf{M}_1, \mathsf{M}_2\ \forall x, y \begin{pmatrix} \neg\big(\neg\mathsf{M}_1(x) \wedge \neg\mathsf{M}_1(x)\big) \wedge \neg\big(\mathsf{M}_1(x) \wedge \mathsf{M}_2(x)\big) \wedge \\ \neg\big(\mathsf{M}_1(x) \wedge \mathsf{M}_1(y) \wedge \mathsf{E}(x,y)\big) \wedge \neg\big(\mathsf{M}_2(x) \wedge \mathsf{M}_2(y) \wedge \mathsf{E}(x,y)\big) \end{pmatrix}$$

There are two equivalence relations on a 2-element set. This means that the signature τ' consists of a binary relation symbol E_1 and of a unary relation symbol E_2. As there are no atoms of the form $\mathsf{E}(x,x)$, all τ-atoms are replaced by E_1-atoms:

$$\Phi_2 := \exists \mathsf{M}_1, \mathsf{M}_2\ \forall x, y \begin{pmatrix} \neg\big(\neg\mathsf{M}_1(x) \wedge \neg\mathsf{M}_1(x)\big) \wedge \neg\big(\mathsf{M}_1(x) \wedge \mathsf{M}_2(x)\big) \wedge \\ \neg\big(\mathsf{M}_1(x) \wedge \mathsf{M}_1(y) \wedge \mathsf{E}_1(x,y)\big) \wedge \neg\big(\mathsf{M}_2(x) \wedge \mathsf{M}_2(y) \wedge \mathsf{E}_1(x,y)\big) \end{pmatrix}$$

Finally, we add the negated conjunct $\neg\mathsf{E}_1(x,x)$ and we obtain the desired MMSNP sentence Φ':

$$\Phi' := \exists \mathsf{M}_1, \mathsf{M}_2\ \forall x, y \begin{pmatrix} \neg\mathsf{E}_1(x,x) \wedge \neg\big(\neg\mathsf{M}_1(x) \wedge \neg\mathsf{M}_1(x)\big) \wedge \neg\big(\mathsf{M}_1(x) \wedge \mathsf{M}_2(x)\big) \wedge \\ \neg\big(\mathsf{M}_1(x) \wedge \mathsf{M}_1(y) \wedge \mathsf{E}_1(x,y)\big) \wedge \neg\big(\mathsf{M}_2(x) \wedge \mathsf{M}_2(y) \wedge \mathsf{E}_1(x,y)\big) \end{pmatrix}$$

△

Theorem 3.1.2. *For any formula Φ in $\mathsf{GMMSNP}_{\neq}$ there exists a formula Φ' in MMSNP such that the problems SAT(Φ) and SAT(Φ') are P-time equivalent.*

Proof. Let Φ' be obtained from Φ by Construction 4. We need to show that the problems $\mathrm{SAT}(\Phi) \equiv_p \mathrm{SAT}(\Phi')$. We describe how to construct an equivalent τ'-structure $\mathfrak{A}'$ from a given τ-structure $\mathfrak{A}$, and then show the P-time equivalence between the problems.

Let $\mathfrak{A}$ be a τ-structure. The corresponding τ'-structure $\mathfrak{A}'$ has the same domain $A' = A$. Consider any n-ary R in τ. Then, for every relation R_k in τ' and for every tuple $\mathbf{a} = (a_1, \ldots, a_n)$ of equivalence type k:

$$\mathsf{R}_k^{\mathfrak{A}'}\big(\mathfrak{p}(\mathbf{a})\big) \leftrightarrow \mathsf{R}^{\mathfrak{A}}(\mathbf{a}). \tag{3.2}$$

For a τ'-structure $\mathfrak{A}'$ there exists a corresponding τ-structure $\mathfrak{A}$ if and only if any relation R_k of $\mathfrak{A}$ does not contain a tuple with some element appearing in it more than once. Also observe that such $\mathfrak{A}'$ is constructible from $\mathfrak{A}$ in P-time.

It remains to prove the correctness of two reductions: we show that $\mathrm{SAT}(\Phi) \leq_p \mathrm{SAT}(\Phi')$ and then do the reverse.

For any relational τ-structure $\mathfrak{A}$, we construct the τ'-structure $\mathfrak{A}'$. By construction, $\mathfrak{A}'$ satisfies every negated conjunct of the type described in eq. (3.1). It is sufficient to show that, for the same choice of any σ-relation for both structures, the corresponding expansions $\mathfrak{A}^\sigma$ and $\mathfrak{A}'^\sigma$ either both satisfy the first-order parts of corresponding formulae or both do not satisfy them. That is, we assume that, for any M in σ, a in A, we have $\mathsf{M}^{\mathfrak{A}^\sigma}(a)$ if and only if $\mathsf{M}^{\mathfrak{A}'^\sigma}(a)$. Observe that, for any negated conjunct $\neg\phi_i(\mathbf{x}_i)$ of Φ, the corresponding negated conjunct of Φ' uses the same variables: $\phi_i'(\mathbf{x}_i)$. Let $\mathfrak{a}(\mathbf{x}_i)$ be an assignment of elements of A to the variables of $\mathbf{x}_i$. For any τ-atom $\mathsf{R}(\mathbf{y})$ of $\phi_i(\mathbf{x}_i)$ we know that, for any two different variables y_i, y_j within $\mathbf{y}$, the negated conjunct $\neg\phi_i(\mathbf{x}_i)$ also contains the inequality $y_i \neq y_j$. The conjunct ϕ_i' is obtained from ϕ_i by replacing every τ-atom $\mathsf{R}(\mathbf{y})$ and the corresponding inequalities $y_i \neq y_j$ with a τ'-atom $\mathsf{R}_k\big(\mathfrak{p}(\mathbf{y})\big)$. So, if the tuple $\mathfrak{a}(\mathbf{y})$ is in $\mathsf{R}^{\mathfrak{A}}$ and all the inequalities: $\mathfrak{a}(y_i) \neq \mathfrak{a}(y_j)$, then, by the construction of $\mathfrak{A}'$, the tuple $\mathfrak{p}(\mathfrak{a}(y))$ belongs to $\mathsf{R}_k^{\mathfrak{A}'}$. If $\mathfrak{p}(\mathfrak{a}(\mathbf{y}))$ is in $\mathsf{R}_k^{\mathfrak{A}'}$, then, by the construction of $\mathfrak{A}'$, $\mathsf{R}(\mathfrak{a}(\mathbf{y}))$ is satisfied in $\mathfrak{A}$; and also any inequality is satisfied because the coordinates of $\mathfrak{p}(y)$ are pairwise distinct, by the definition of the function $\mathfrak{p}$. Hence, for any assignment $\mathfrak{a}$, $\phi_i(\mathfrak{a}(\mathbf{x}_i))$ holds in $\mathfrak{A}^\sigma$ if and only if $\phi_i'(\mathfrak{a}(\mathbf{x}_i))$ holds in $\mathfrak{A}'^\sigma$. This proves that $\mathrm{SAT}(\Phi) \leq_p \mathrm{SAT}(\Phi')$.

We prove now the reduction from $\mathrm{SAT}(\Phi')$ to $\mathrm{SAT}(\Phi)$. Consider any relational τ'-structure $\mathfrak{A}'$. It takes P-time in its size to check that every formula of the type from eq. (3.1) is true in $\mathfrak{A}'$. If it is not true, then $\mathfrak{A}' \not\models \Phi'$, thus we can reduce it to a fixed NO instance of Φ. If any such formula is satisfied, then there exists a τ-structure $\mathfrak{A}$ with the same domain $A = A'$ such that the condition from eq. (3.2) holds. This structure $\mathfrak{A}$ is constructible from $\mathfrak{A}'$ in P-time in its size. And these two structures are equivalent w.r.t. the problems $\mathrm{SAT}(\Phi)$ and $\mathrm{SAT}(\Phi')$, as shown previously.

This concludes that $\mathrm{SAT}(\Phi) \equiv_p \mathrm{SAT}(\Phi')$.

□

3.2 MMSNP$_2$ and Guarded Monotone Strict NP

Let $\tau = \{\mathsf{R}_1, \ldots, \mathsf{R}_t\}$ be the input relational signature and $\sigma = \{\mathsf{X}_1, \ldots, \mathsf{X}_s\}$ be the existential relational signature. A sentence of *guarded monotone strict NP (GMSNP)* has the form

$$\exists \mathsf{X}_1, \ldots, \mathsf{X}_s \; \forall \mathbf{x} \; \phi(\mathbf{x}),$$

where ϕ is a conjunction of formulae

$$\neg\phi_i := \neg(\alpha_1 \wedge \cdots \wedge \alpha_n \wedge \neg\beta_1 \wedge \cdots \wedge \neg\beta_m) \text{ with } n, m \geq 0,$$

where each α_i is either

- an atom $\mathsf{X}_i(\mathbf{y})$, for X_i in σ, or
- an atom $\mathsf{R}(\mathbf{y})$, for R in τ, or
- an equality $x = y$;

and each β_i is of the form $\mathsf{X}_i(\mathbf{y})$, for X_i in σ. Additionally, we require that for every negated atom $\neg\beta_i$ there is an atom α_j such that α_j contains all variables from β_i. It is said that α_j *guards* β_i.

Example 3.2.1. Suppose that we have a ternary input τ-relation $\mathsf{C}_3(x, y, z)$ that encodes the directed cycle $\mathfrak{C}_3$. We want to check that a given τ-structure does not have a 2-cycle. In order to check this property, we introduce two binary σ-relations:

- $\mathsf{E}(x, y)$, which means that (x, y) is an arc, and
- $\mathsf{C}_2(x, y)$, which means that x and y make a directed cycle of length 2.

The query can be expressed by the following GMSNP sentence:

$$\exists \mathsf{E}, \mathsf{C}_2\ \forall x, y \begin{pmatrix} \neg\big(\mathsf{C}_3(x,y,z) \wedge \neg\mathsf{E}(x,y)\big)\wedge \\ \neg\big(\mathsf{C}_3(x,y,z) \wedge \neg\mathsf{E}(y,z)\big)\wedge \\ \neg\big(\mathsf{C}_3(x,y,z) \wedge \neg\mathsf{E}(z,x)\big) \end{pmatrix} \wedge \neg\Big(\mathsf{E}(x,y) \wedge \mathsf{E}(y,x) \wedge \neg\mathsf{C}_2(x,y)\Big) \wedge \neg\mathsf{C}_2(x,y).$$

△

Suppose now that the existential signature σ can be represented in the following form: $\sigma = \sigma_0 \biguplus_{1\leq i\leq t} \sigma_i$, where σ_0 consists of unary relation symbols, and each σ_i consists of relation symbols of the same arity as the corresponding input relation R_i. A sentence of MMSNP_2 logic has the form

$$\exists \mathsf{X}_1, \ldots, \mathsf{X}_s\ \forall \mathbf{x}\ \phi(\mathbf{x}),$$

where ϕ is a conjunction of formulae

$$\neg\phi_i := \neg(\alpha_i \wedge \beta_i),$$

where each α_i is a conjunction of non-negated τ-atoms, each β_i is a conjunction of σ-atoms or negated σ-atoms. And, for each atom in β_i of the form $\mathsf{X}(\mathbf{y})$ or $\neg\mathsf{X}(\mathbf{y})$, where X is in σ_i, for $1 \leq i \leq t$, α_i contains a τ-atom of the form $\mathsf{R}_i(\mathbf{y})$.

Example 3.2.2. (No-Monochromatic-Arc-Triangle) Let Φ describe the following problem. That is, a directed graph $\mathfrak{G}$ satisfies Φ if and only if one can colour its arcs in two colours B and W such that $\mathfrak{G}$ does not contain a 3-cycle as a subgraph, where all 3 arcs have the same colour. Φ can be written as follows:

$$\exists \mathsf{B}, W\ \forall x, y, z \begin{pmatrix} \neg\big(\mathsf{E}(x,y) \wedge \mathsf{B}(x,y) \wedge W(x,y)\big)\wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \neg\mathsf{B}(x,y) \wedge \neg W(x,y)\big)\wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{E}(y,z) \wedge \mathsf{E}(z,x) \wedge \mathsf{B}(x,y) \wedge \mathsf{B}(y,z) \wedge \mathsf{B}(z,x)\big)\wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{E}(y,z) \wedge \mathsf{E}(z,x) \wedge W(x,y) \wedge W(y,z) \wedge W(z,x)\big) \end{pmatrix}$$

△

The class GMSNP is studied in the paper [BtCLW14] of Bienvenu et al. The authors investigate its relation with MMSNP and show that GMSNP is strictly more expressive than MMSNP. Moreover, they compare GMSNP to MMSNP_2 (that was introduced before by Madelaine in [Mad09]) and show that MMSNP_2 has the same expressive power as GMSNP. That is, for any GMSNP sentence Φ there exists an MMSNP_2 sentence Φ' that is logically equivalent to Φ and vice versa.

In Section 3.2.1 we first show that for any finite relational signature τ there is a signature τ_1 consisting of just one relation symbol such that for any τ-sentence Φ in GMSNP there is a P-time equivalent τ_1-sentence Φ_1 in GMSNP. Then, in Section 3.2.2, we strengthen the result of Bienvenu et al. by showing that for any τ_1-sentence Φ in GMSNP there is a logically equivalent τ_1-sentence in MMSNP_2, as in their proof the signature contains arbitrarily many relation symbols. So now without loss of generality one can study the dichotomy question for MMSNP_2 over a signature with a unique relation symbol, this makes the notations much simpler. We also rewrite this proof for convenience as the original proof of Bienvenu et al. uses different notations.

3.2.1 Signature simplification for Guarded Monotone Strict NP

Here we show how one can simplify a GMSNP problem such that without loss of generality one can consider the case when the input relational signature consists only of one relation symbol. We do it by showing that for any finite signature $\tau = \{\mathsf{R}_1, \dots, \mathsf{R}_t\}$ there exists a signature $\tau_1 = \{\mathsf{P}\}$ such that for any GMSNP problem over τ there is a P-time equivalent GMSNP problem over τ_1, see Lemma 3.2.10.

An SNP τ-sentence Φ is called *connected* if no negated conjunct $\neg\phi(\mathbf{x})$ of Φ has the form

$$\neg\left(\psi_1(\mathbf{x}_1) \wedge \psi_2(\mathbf{x}_2)\right),$$

where the tuples $\mathbf{x}_1$ and $\mathbf{x}_2$ share no variables. Denote by ConnectedGMSNP the set of all connected GMSNP sentences over all finite relational signatures.

We first show that GMSNP and ConnectedGMSNP are P-time equivalent. Then we show that any ConnectedGMSNP sentence is equivalent to a ConnectedGMSNP sentence over a one-element relational signature.

Recall that the family of structures that satisfy some connected SNP sentence is closed under taking disjoint unions.

Proposition 3.2.1 ([Bod21])**.** *Let Φ be an SNP sentence. Then the class of structures that satisfy Φ is closed under disjoint unions if and only if Φ is logically equivalent to a connected SNP sentence.*

It is known that one can rewrite any GMSNP sentence as a disjunction of connected sentences.

Proposition 3.2.2 ([BKS20])**.** *Every GMSNP sentence Φ is logically equivalent to a finite disjunction $\Phi_1 \vee \cdots \vee \Phi_k$ of connected GMSNP sentences.*

The following result is similar to a well-know similar result for MMSNP sentences. One can find a proof for the case of MMSNP in [BMM18]. We provide a similar proof for the class GMSNP.

Proposition 3.2.3. *Let Φ be a GMSNP τ-sentence that is logically equivalent to a disjunction $\Phi_1 \vee \cdots \vee \Phi_k$ of ConnectedGMSNP sentences. Then SAT(Φ) is P-time solvable if all SAT(Φ_i) are in P. If one SAT(Φ_i) is NP-hard, then so is SAT(Φ).*

Proof. If all SAT(Φ_i) are P-time solvable, then, for any τ-structure $\mathfrak{A}$, we can check in P-time in its size if $\mathfrak{A}$ satisfies one of Φ_is. If yes, then $\mathfrak{A} \models \Phi$; if no, then $\mathfrak{A} \not\models \Phi$. So SAT($\Phi$) is P-time solvable.

Suppose that, for some i in $[k]$, SAT(Φ_i) is NP-hard. Let k be the smallest number such that Φ is represented as a disjunction of k connected sentences. Then there exists a structure $\mathfrak{B}$ such that, for all j in $[k] \setminus \{i\}$, $\mathfrak{B} \not\models \Phi_j$, and $\mathfrak{B} \models \Phi_i$. If such structure does not exist, then we can delete Φ_i from the disjunction and keep it logically equivalent to Φ, this contradicts the minimality of k. We reduce SAT(Φ_i) to SAT(Φ) as follows: for any τ-structure $\mathfrak{A}$ we construct a τ-structure $\mathfrak{A}' := \mathfrak{A} \uplus \mathfrak{B}$ and show that $\mathfrak{A} \models \Phi_i$ if and only if $\mathfrak{A}' \models \Phi$.

Suppose $\mathfrak{A} \models \Phi_i$; then, as $\mathfrak{B}$ also satisfies Φ_i, by Proposition 3.2.1, their disjoint union $\mathfrak{A}'$ also satisfies Φ_i, thus, $\mathfrak{A}' \models \Phi$.

Suppose $\mathfrak{A}' \models \Phi$. Then, for some j in $[k]$, $\mathfrak{A}' \models \Phi_j$. Lemma 2.2.1 states that any SNP sentence is closed under taking induced substructures, thus, we have $\mathfrak{B} \models \Phi_j$. This means that $j = i$. As $\mathfrak{A}$ is also an induced substructure of $\mathfrak{A}'$, by Lemma 2.2.1, we have $\mathfrak{A} \models \Phi_i$ as well. □

Corollary 3.2.4. *Let τ be a finite relational signature. The class of GMSNP problems over τ has a dichotomy if and only if the class of connected GMSNP problems over τ has a dichotomy.*

We shall prove that from the perspective of dichotomy we can assume that a GMSNP sentence has only one input symbol, that is we prove the following.

Theorem 3.2.5. *For any finite relational signature τ there exists a signature τ_1 consisting of just one relational symbol such that, for any GMSNP τ-sentence Φ there exists a GMSNP τ_1-sentence Φ_1 such that the problems SAT(Φ) and SAT(Φ_1) are P-time equivalent.*

Let R in τ have arity k in $\mathbb{N}$. Let $\mathfrak{B}_\mathsf{R}$ be a τ-structure that consists of one R-tuple: $\mathbf{b}_\mathsf{R} = (b_{\mathsf{R},1}, \ldots, b_{\mathsf{R},k})$. That is, $B_\mathsf{R} = \{b_{\mathsf{R},1}, \ldots, b_{\mathsf{R},k}\}$, the relation R contains just one tuple $\mathbf{b}_\mathsf{R}$, and any other relation is interpreted as the empty set.

Recall that GMSNP problems are closed under inverse homomorphisms because they all belong to MonotoneSNP.

Theorem 3.2.6 ([FV03])**.** *Let Φ be an SNP sentence. Then the class of structures that satisfy Φ is closed under inverse homomorphisms if and only if Φ is logically equivalent to a MonotoneSNP sentence.*

We shall first discard input symbols of GMSNP that are implicitly forbidden by the sentence. This will allow us to eventually encode the many – now necessary – input symbols into a single input symbol.

Lemma 3.2.7. *If a connected GMSNP τ-sentence Φ does not satisfy the structure $\mathfrak{B}_\mathsf{R}$, for some R in τ, then the problem SAT(Φ) is P-time equivalent to SAT($\Phi^{\not\mathsf{R}}$), where $\Phi^{\not\mathsf{R}}$ is a connected $(\tau \setminus \{\mathsf{R}\})$-sentence that is obtained by removing from Φ all negated conjuncts that contain R-atoms.*

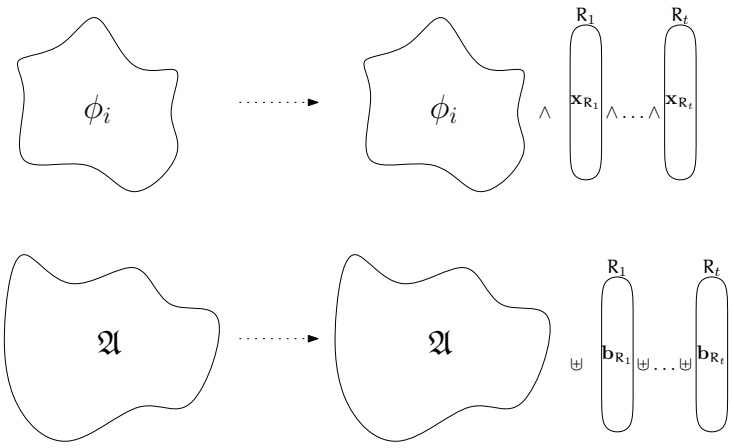


Figure 3.2: Above, one can see the transformation of each negated conjunct of Φ to a negated conjunct of Φ'. Below, it is shown how the structure $\mathfrak{A}'$ is constructed from $\mathfrak{A}$.

Proof. Suppose that $\mathfrak{B}_{\mathsf{R}} \not\models \Phi$. Let a τ-structure $\mathfrak{A}$ be an input instance of SAT(Φ). If $\mathsf{R}^{\mathfrak{A}} \neq \varnothing$, then $\mathfrak{B}_{\mathsf{R}} \to \mathfrak{A}$, so, by Theorem 3.2.6, $\mathfrak{A} \not\models \Phi$, and we reduce it to some NO instance of SAT($\Phi^{\not R}$). Suppose that $\mathsf{R}^{\mathfrak{A}} = \varnothing$; then we reduce $\mathfrak{A}$ to its $(\tau \setminus \{\mathsf{R}\})$-reduct $\mathfrak{A}^{\not R}$. For any σ-expansion of $\mathfrak{A}$, any negated conjunct with an R-atom is always satisfied, as $\mathsf{R}^{\mathfrak{A}} = \varnothing$. Thus, $\mathfrak{A} \models \Phi$ if and only if $\mathfrak{A}^{\not R} \models \Phi^{\not R}$. For τ-structures $\mathfrak{A}$ such that $\mathsf{R}^{\mathfrak{A}} = \varnothing$, the correspondence $\mathfrak{A} \leftrightarrow \mathfrak{A}^{\not R}$ is one-to-one. This means that the two problems are P-time equivalent. □

Suppose that Lemma 3.2.7 does not apply to Φ. Then we shall construct a new sentence Φ' and, for any τ-structure $\mathfrak{A}$, a τ-structure $\mathfrak{A}'$ such that

- any negated conjunct of Φ' contains at least one R-atom of each R in τ;
- $\mathfrak{A} \models \Phi$ if and only if $\mathfrak{A}' \models \Phi'$.

Construction 5. The sentence Φ' is obtained from Φ as follows. Take any negated conjunct $\neg\phi_i$ of Φ and replace it with

$$\neg\phi_i' := \neg\left(\phi_i \wedge \bigwedge_{\mathsf{R}\in\tau} \mathsf{R}(\mathbf{x}_{\mathsf{R}})\right),$$

where each $\mathbf{x}_{\mathsf{R}}$ is a tuple of new variables. Φ' is no longer connected. However, it is still in GMSNP. And Φ is more restrictive than Φ', as we enrich negated conjuncts of Φ with τ-atoms. That is, for any τ-structure $\mathfrak{A}$, we have that $\mathfrak{A} \models \Phi$ implies $\mathfrak{A} \models \Phi'$.

For a τ-structure $\mathfrak{A}$, let $\mathfrak{A}' := \mathfrak{A} \uplus \left(\biguplus_{\mathsf{R}\in\tau} \mathfrak{B}_{\mathsf{R}}\right)$ – be the disjoint union of $\mathfrak{A}$ and singleton tuples, for every relation in τ. The sentence Φ' and the structure $\mathfrak{A}'$ are displayed on Figure 3.2 on page 67. This is the end of Construction 5.

Lemma 3.2.8. *Let $\Phi, \Phi', \mathfrak{A}$ and $\mathfrak{A}'$ be as in Construction 5. Then $\mathfrak{A} \models \Phi$ if and only if $\mathfrak{A}' \models \Phi'$. Consequently, SAT($\Phi$) is P-time reducible to SAT(Φ').*

Proof. Let $\mathfrak{A}$ be an input instance of SAT(Φ), and $\mathfrak{A}'$ be an input instance of SAT(Φ') obtained from $\mathfrak{A}$ as in Construction 5. We need to prove that $\mathfrak{A} \models \Phi$ if and only if $\mathfrak{A}' \models \Phi'$. Suppose that $\mathfrak{A} \models \Phi$. We also know that, for any R in τ, $\mathfrak{B}_{\mathsf{R}} \models \Phi$. Then, by

Proposition 3.2.1, $\mathfrak{A}' \models \Phi$, as $\mathfrak{A}'$ is the disjoint union of structures satisfying Φ. But then, as Φ' is less restrictive than Φ, we have $\mathfrak{A}' \models \Phi'$.

Suppose that $\mathfrak{A}' \models \Phi'$; then, for some σ-expansion $\mathfrak{A}'^{\sigma}$, the first-order part ϕ' of Φ' is satisfied. Choose a σ-expansion $\mathfrak{A}^{\sigma}$ of $\mathfrak{A}$ to be equal to $\mathfrak{A}'^{\sigma}[A]$, the substructure of $\mathfrak{A}'^{\sigma}$ induced on A. Suppose that, for some negated conjunct $\neg\phi_i(\mathbf{x}_i)$ of Φ, there is an assignment that makes it false in $\mathfrak{A}^{\sigma}$. But then we can take the corresponding negated conjunct $\neg\left(\phi_i \wedge \bigwedge_{\mathsf{R}\in\tau} \mathsf{R}(\mathbf{x}_{\mathsf{R}})\right)$ of Φ', and assign $\mathbf{b}_{\mathsf{R}}$ to each new tuple of variables $\mathbf{x}_{\mathsf{R}}$. Under this assignment the conjunct is false in $\mathfrak{A}'^{\sigma}$, this is a contradiction. We have shown that $\mathfrak{A} \models \Phi$ if and only if $\mathfrak{A}' \models \Phi'$. □

Lemma 3.2.9. *Let Φ, Φ' be as in Construction 5. Then SAT(Φ') is P-time reducible to SAT(Φ).*

Proof. Consider a τ-structure $\mathfrak{A}$ that is an input instance of SAT(Φ'). We first suppose that, for some R in τ, $\mathsf{R}^{\mathfrak{A}} = \varnothing$. Then, by Construction 5, $\mathfrak{A} \models \Phi'$. Then we reduce it to some fixed YES instance of SAT(Φ), *e.g.*, to $\mathfrak{B}_{\mathsf{R}}$, as we know that $\mathfrak{B}_{\mathsf{R}} \models \Phi$.

Suppose now that any relation of $\mathfrak{A}$ is interpreted as a non-empty set. Then, $\mathfrak{A} \models \Phi$ implies $\mathfrak{A} \models \Phi'$ because Φ is more restrictive than Φ'. Suppose that $\mathfrak{A} \not\models \Phi$; then, for any σ-expansion $\mathfrak{A}^{\sigma}$ of $\mathfrak{A}$ there is a negated conjunct $\neg\phi_i(\mathbf{x}_i)$ of Φ and a tuple $\mathbf{a}_i$ of elements of $\mathfrak{A}$ such that $\mathfrak{A}^{\sigma} \models \phi_i(\mathbf{a}_i)$. But then, take the negated conjunct $\neg\left(\phi_i(\mathbf{x}_i) \wedge \bigwedge_{\mathsf{R}\in\tau} \mathsf{R}(\mathbf{x}_{\mathsf{R}})\right)$, assign $\mathbf{a}_i$ to $\mathbf{x}_i$ and assign $\mathbf{a}_{\mathsf{R}}$ to $\mathbf{x}_{\mathsf{R}}$ such that, for any R in τ, $\mathbf{a}_{\mathsf{R}} \in \mathsf{R}^{\mathfrak{A}}$. At least one such $\mathbf{a}_{\mathsf{R}}$ exists, by our assumption. But then we have $\mathfrak{A}^{\sigma} \models \left(\phi_i(\mathbf{a}_i) \wedge \bigwedge_{\mathsf{R}\in\tau} \mathsf{R}(\mathbf{a}_{\mathsf{R}})\right)$. This means that the two problems SAT(Φ) $\equiv_p$ SAT(Φ'). □

By Lemma 3.2.7, Lemma 3.2.8, and Lemma 3.2.9, we can assume that every negated conjunct of Φ contains a τ-atom of each relation symbol of τ. Set $\tau_1 := \{\mathsf{P}\}$, where

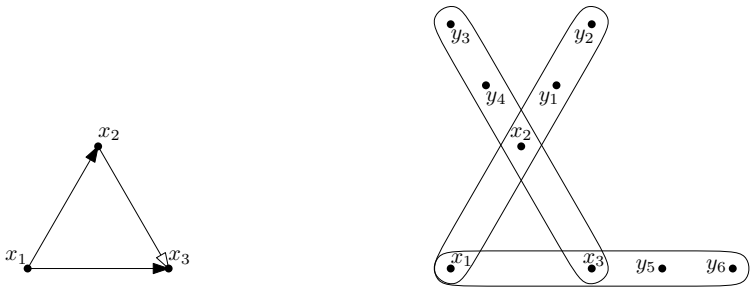
$$arity(\mathsf{P}) := \sum_{\mathsf{R}\in\tau} arity(\mathsf{R}). \tag{3.3}$$

For two tuples $\mathbf{x} = (x_1, \ldots, x_n), \mathbf{y} = (y_1, \ldots, y_m)$, we use the following notation for their concatenation: $(\mathbf{x}, \mathbf{y}) := (x_1, \ldots, x_n, y_1, \ldots, y_m)$.

Now we construct a GMSNP τ_1-sentence Φ^1 and then show that SAT(Φ) with the restricted input is P-time equivalent to SAT(Φ^1). It will be the main result of this subsection.

Construction 6. Consider a negated conjunct $\neg\phi_i$ of Φ and construct the corresponding negated conjunct $\neg\phi_i^1$ as follows. The σ-atoms of ϕ_i^1 are the same as in ϕ_i. Each τ-atom of ϕ_i is replaced by a τ_1-atom as follows. For any j in $[t]$ and a τ-atom $\mathsf{R}_j(\mathbf{x})$, ϕ_i^1 contains a τ_1-atom $\mathsf{P}(\mathbf{y}_1, \ldots, \mathbf{y}_t)$, where $\mathbf{y}_j = \mathbf{x}$ and all other variables of this formula are new and they are used only in this atomic formula. Φ^1 is a GMSNP sentence because the σ-parts are kept the same and all the variables of a τ-atom of Φ are contained in the corresponding τ_1-atom of Φ^1. An example of such construction is displayed on Figure 3.3 on page 69. Observe that Φ^1 is connected. This is the end of Construction 6.

Lemma 3.2.10. *Suppose that Φ^1 is obtained from Φ by Construction 6. Then, for any τ-structure $\mathfrak{A}$ there exists a τ_1-structure $\mathfrak{A}_1$ such that $\mathfrak{A} \models \Phi$ if and only if $\mathfrak{A}_1 \models \Phi^1$; and for any τ_1-structure $\mathfrak{A}_1$ there is a τ-structure $\mathfrak{A}$ such that $\mathfrak{A}_1 \models \Phi^1$ if and only if $\mathfrak{A} \models \Phi$.*



$\mathsf{R}_1(x_1, x_2) \wedge \mathsf{R}_2(x_2, x_3) \wedge \mathsf{R}_1(x_1, x_3)$ $\quad$ $\mathsf{P}(x_1, x_2, y_1, y_2) \wedge \mathsf{P}(y_3, y_4, x_2, x_3) \wedge \mathsf{P}(x_1, x_3, y_5, y_6)$

Figure 3.3: On the left, there is the τ-part of some negated conjunct of Φ. On the right, there is the τ_1-part of the corresponding negated conjunct of Φ^1.

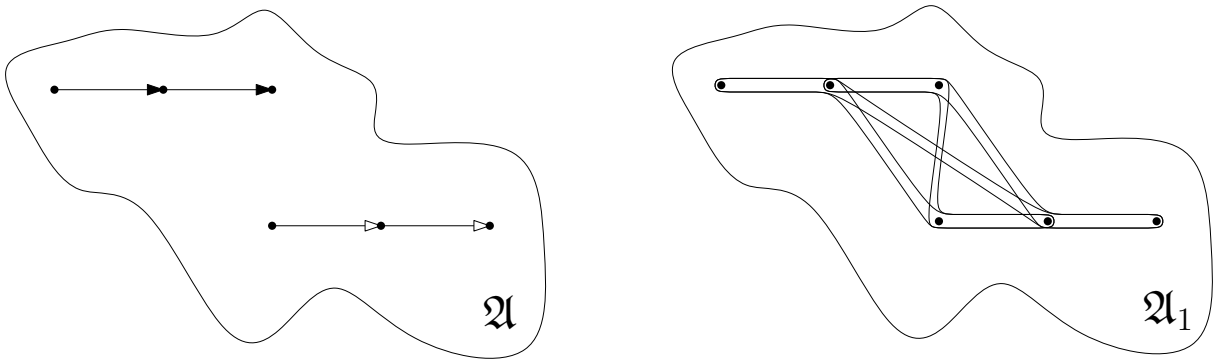


Figure 3.4: The original τ-structure $\mathfrak{A}$ is on the left. The corresponding τ_1-structure $\mathfrak{A}_1$ is on the right.

Proof. For a τ-structure $\mathfrak{A}$ we construct the corresponding τ_1-structure $\mathfrak{A}_1$ as follows. The structures have the same domain: $A_1 = A$. The relation $\mathsf{P}^{\mathfrak{A}_1}$ is defined by the relations of $\mathfrak{A}$ as on eq. (3.4). An example of such construction is displayed on Figure 3.4 on page 69.

$$\mathsf{P}^{\mathfrak{A}_1}(\mathbf{a}_1, \ldots, \mathbf{a}_t) \leftrightarrow \mathsf{R}_1^{\mathfrak{A}}(\mathbf{a}_1) \wedge \cdots \wedge \mathsf{R}_t^{\mathfrak{A}}(\mathbf{a}_t). \tag{3.4}$$

Observe that, if $\mathsf{R}^{\mathfrak{A}} = \varnothing$, for some R in τ, then $\mathsf{P}^{\mathfrak{A}_1} = \varnothing$. In this case, we have $\mathfrak{A} \models \Phi$, as every negated conjunct contains an R-atom. And we also have $\mathfrak{A}_1 \models \Phi_1$, as every negated conjunct of Φ_1 contains at least one P-atom. So, we can now assume that $\mathsf{R}^{\mathfrak{A}} \neq \varnothing$, for any R in τ, and that $\mathsf{P}^{\mathfrak{A}_1} \neq \varnothing$.

Let $\mathfrak{A}^\sigma$ and $\mathfrak{A}_1^\sigma$ be two σ-expansions of $\mathfrak{A}$ and $\mathfrak{A}'$ such that, for any X in σ, $\mathsf{X}^{\mathfrak{A}^\sigma} = \mathsf{X}^{\mathfrak{A}_1^\sigma}$. It is sufficient to show that $\mathfrak{A}^\sigma \models \forall \mathbf{x}\ \phi(\mathbf{x})$ if and only if $\mathfrak{A}_1^\sigma \models \forall \mathbf{x}, \mathbf{y}\ \phi^1(\mathbf{x}, \mathbf{y})$, where $\mathbf{y}$ represents variables that are added during the construction of Φ^1. Suppose that $\mathfrak{A}^\sigma \not\models \forall \mathbf{x}\ \phi(\mathbf{x})$; then, for some negated conjunct $\neg\phi_i(\mathbf{x}_i)$ of Φ and for some tuple $\mathbf{a}_i$, we have $\mathfrak{A}^\sigma \models \phi_i(\mathbf{a}_i)$. Consider the ith negated conjunct $\neg\phi_i^1(\mathbf{x}_i, \mathbf{y}_i)$ of Φ^1. To the tuple of variables $\mathbf{x}_i$ that are used in ϕ_i as well, we assign the same tuple $\mathbf{a}_i$ of elements of A. Let $\mathsf{P}(\mathbf{y}_1, \ldots, \mathbf{y}_t)$ be a τ_1-atom that is contained in ϕ_i^1. Pick a tuple $\mathbf{y}_j$ of variables of this atomic formula such that nothing is assigned to these variables yet. We know that, for the jth relation symbol R_j in τ, ϕ_i contains at least one τ-atom $\mathsf{R}_j(\mathbf{z}_j)$, thus each variable of $\mathbf{z}_j$ is contained in $\mathbf{x}_i$. As $\mathbf{a}_i$ is assigned to $\mathbf{x}_i$, some tuple $\mathbf{b}_j$ of elements of A is already assigned to $\mathbf{z}_j$. Hence, we assign the tuple $\mathbf{b}_j$ to $\mathbf{y}_j$. Repeat this procedure for all unassigned variables of $\mathsf{P}(\mathbf{y}_1, \ldots, \mathbf{y}_t)$. By eq. (3.4), we have $\mathfrak{A}_1^\sigma \models \mathsf{P}(\mathbf{b}_1, \ldots, \mathbf{b}_t)$. This is true for any τ-atom of ϕ_i^1. Thus, $\mathfrak{A}_1^\sigma \not\models \forall \mathbf{x}, \mathbf{y}\ \phi(\mathbf{x}, \mathbf{y})$. For the other direction, suppose that $\mathfrak{A}_1^\sigma \not\models \forall \mathbf{x}, \mathbf{y}\ \phi(\mathbf{x}, \mathbf{y})$; then there is a negated conjunct $\neg\phi_i^1(\mathbf{x}_i, \mathbf{y}_i)$ and tuples $\mathbf{a}_i, \mathbf{b}_i$

of elements of A such that $\mathfrak{A}_1^\sigma \models \phi_i^1(\mathbf{a}_i, \mathbf{b}_i)$. Take the corresponding negated conjunct $\neg\phi_i(\mathbf{x}_i)$ of Φ and assign $\mathbf{a}_i$ to $\mathbf{x}_i$. We have $\mathfrak{A}^\sigma \models \phi_i(\mathbf{a}_i)$, by eq. (3.4).

For an arbitrary τ_1-structure $\mathfrak{B}_1$, we construct a τ-structure $\mathfrak{B}$. They have the same domain B and, for any j in $[t]$, the relation $\mathsf{R}_j^{\mathfrak{B}}$ is defined as follows:

$$\mathsf{R}_j^{\mathfrak{B}}(\mathbf{x}_j) \longleftrightarrow \exists \mathbf{x}_1, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_t \; \mathsf{P}^{\mathfrak{B}_1}(\mathbf{x}_1, \ldots, \mathbf{x}_t). \tag{3.5}$$

Similarly as above, we consider two σ-expansions $\mathfrak{B}^\sigma$ and $\mathfrak{B}_1^\sigma$ such that, for all X in σ, $\mathsf{X}^{\mathfrak{B}^\sigma} = \mathsf{X}^{\mathfrak{B}_1^\sigma}$. Suppose that $\mathfrak{B}_1^\sigma \not\models \forall \mathbf{x}, \mathbf{y}\ \phi^1(\mathbf{x}, \mathbf{y})$; then there exists a negated conjunct $\neg\phi_i^1(\mathbf{x}_i, \mathbf{y}_i)$ and assignments $\mathbf{a}_i, \mathbf{b}_i$ to $\mathbf{x}_i, \mathbf{y}_i$ correspondingly such that $\mathfrak{B}_1^\sigma \models \phi_i^1(\mathbf{a}_i, \mathbf{b}_i)$. Take the ith negated conjunct $\neg\phi_i(\mathbf{x}_i)$ of Φ and assign $\mathbf{a}_i$ to $\mathbf{x}_i$. As ϕ_i and ϕ_i^1 have the same σ-part, we have $\mathfrak{B}^\sigma$ satisfies any σ-atom and negated σ-atom of ϕ_i. Any τ-atom of ϕ_i is satisfied, by eq. (3.5). So we have $\mathfrak{B}^\sigma \models \phi_i(\mathbf{a}_i)$. For the other direction, suppose that $\mathfrak{B}^\sigma \models \phi_i(\mathbf{a}_i)$, for some negated conjunct $\neg\phi_i(\mathbf{x}_i)$ of Φ and for some assignment $\mathbf{a}_i$ of elements of B to the variables $\mathbf{x}_i$ of ϕ_i. For any atomic formula $\mathsf{R}_j(\mathbf{z})$ of α_i, the conjunction ϕ_i^1 contains an atomic formula $\mathsf{P}(\mathbf{y}_1, \ldots, \mathbf{y}_t)$, where $\mathbf{y}_j = \mathbf{z}$. We know that some tuple $\mathbf{c}$ is assigned to $\mathbf{z}$ such that $\mathfrak{B}^\sigma \models \mathsf{R}_j(\mathbf{c})$. By eq. (3.5), there exist $\mathbf{b}_1, \ldots, \mathbf{b}_{j-1}, \mathbf{b}_{j+1}, \ldots, \mathbf{b}_t$ such that $\mathsf{P}^{\mathfrak{B}_1^\sigma}(\mathbf{b}_1, \ldots, \mathbf{b}_t)$, where $\mathbf{b}_j = \mathbf{c}$. So, for any $\mathbf{y}_1, \ldots, \mathbf{y}_t$ except for $\mathbf{y}_j = \mathbf{z}$, we assign $\mathbf{b}_1, \ldots, \mathbf{b}_t$. Then, $\mathfrak{B}_1^\sigma \models \mathsf{P}(\mathbf{b}_1, \ldots, \mathbf{b}_t)$, this is true for any τ-atom of ϕ_i^1. Any σ-atom and negated σ-atom of ϕ_i is also satisfied by the choice of σ-expansions. This implies that $\mathfrak{B}_1^\sigma \not\models \forall \mathbf{x}, \mathbf{y}\ \phi^1(\mathbf{x}, \mathbf{y})$ and we are done.

□

Proof of Theorem 3.2.5. For a given finite relational signature τ, the sole symbol P of τ_1 has arity as in eq. (3.3). Let Φ be some GMSNP τ-sentence. We can assume without loss of generality that it is connected, by Corollary 3.2.4. If the condition of Lemma 3.2.7 applies to Φ, then $\mathrm{SAT}(\Phi) \equiv_p \mathrm{SAT}(\Psi)$, where Ψ is a GMSNP τ'-sentence, and $\tau' \subsetneq \tau$, and the condition of Lemma 3.2.7 does not apply to Ψ. So, we can assume that it does not apply to Φ. Then, by Lemma 3.2.8 and Lemma 3.2.9, $\mathrm{SAT}(\Phi) \equiv_p \mathrm{SAT}(\Phi')$, where every negated conjunct of Φ' contains at least one R-atom, for every R in τ. Then, by Construction 6, we construct a GMSNP τ_1-sentence Φ^1 such that $\mathrm{SAT}(\Phi') \equiv_p \mathrm{SAT}(\Phi^1)$, by Lemma 3.2.10. This means that $\mathrm{SAT}(\Phi) \equiv_p \mathrm{SAT}(\Phi^1)$. □

3.2.2 Equivalence between Guarded Monotone Strict NP and MMSNP_2

Suppose that $\tau = \{\mathsf{R}\}$ is the input signature that consists of one relation symbol of arity k. Our objective is to show that for every GMSNP τ-sentence there is a logically equivalent MMSNP_2 τ-sentence, and vice versa. This yields that the two classes have the same expressive power.

There are two major differences between GMSNP and MMSNP_2. At first, within a GMSNP sentence a negated existential atom may be guarded by another existential atom, while in MMSNP_2 sentences existential atoms are always guarded by τ-atoms. The second difference is that the existential relations within a GMSNP sentence may have arbitrary arity, while, within MMSNP_2 sentences the arity is determined: either it equals the arity of the input relation, or it is 1. The following definition is helpful with respect to the second difference.

Definition 3. Let $\mathcal{E}_k^m$ denote the family of injective mappings from sets of size at most m to sets with at most k elements, for $m < k$. That is,

$$\mathcal{E}_k^m := \big\{\mathsf{e}\colon [m'] \to [k'] \mid m' \leq m, k' \leq k, \mathsf{e} \text{ is injective}\big\}.$$

Denote $n := |\mathcal{E}_k^m|$ to be the number of these mappings and suppose that all of them are linearly ordered: $\mathcal{E}_k^m = \{\mathsf{e}_1, \ldots, \mathsf{e}_n\}$. Let $\mathbf{p}$ be the function from Definition 2 on page 61 that deletes repeating elements from a tuple. Let $\mathbf{x}$ be an m-tuple, $\mathbf{y}$ be a k-tuple, $\mathbf{x}' := \mathbf{p}(\mathbf{x})$ and $\mathbf{y}' := \mathbf{p}(\mathbf{y})$ be the corresponding m'- and k'-tuples with all repeating elements having been removed. We say that $\mathbf{x}$ has *guarding type* i in $\mathbf{y}$ if, for any j in $[m']$, we have $x'_j = y'_{\mathsf{e}_i(j)}$, where $\mathsf{e}_i\colon [m'] \to [k']$ is the ith element in $\mathcal{E}_k^m$.

The following statement is the main result of this part. It is initially proved in [Bt-CLW14]. However, we reprove it here, for convenience.

Proposition 3.2.11. *GMSNP and MMSNP$_2$ have the same expressive power.*

Proof. It is straightforward to show that any MMSNP_2 sentence is also a GMSNP sentence: if $\neg\beta_i$ is of the form $\neg\mathsf{X}(y)$, for some unary X in σ, then we can enrich the conjunction ϕ_i with the equality $y = y$ that guards β_i, and all other βs are already guarded by τ-atoms. As $y = y$ always holds, the resulting sentence is logically equivalent to the original one.

Now, consider a GMSNP sentence Φ. We want to do the following steps in order to finish the proof:

1. To remove all the equalities from Φ.

2. To enrich the negated conjuncts of Φ such that within any conjunct, for any σ_i-atom there is a τ-atom in the same conjunct that guards this σ-atom. By transitivity, this implies that any negated σ-atom is also guarded by a τ-atom. The modified sentence is logically equivalent to Φ.

3. To replace the existential relational signature σ with a new signature σ' consisting only of k-ary relation symbols. And to replace every σ-atom with a σ'-atom so that the result is an MMSNP_2 sentence Φ' that is logically equivalent to Φ.

Removing equalities. For every equality $x = y$ that is in some negated conjunct $\neg\phi_i$ of Φ, we replace every occurence of y within ϕ_i with x. If this equality guards some negated σ-atoms, then, after replacing ys with xs, there might be a negated σ-atom of the form $\neg\mathsf{X}(x, \ldots, x)$. For every X in σ, we add to σ a unary relation symbol L_X and add to Φ the following negated conjuncts:

$$\neg\big(\mathsf{X}(x, \ldots, x) \wedge \neg\mathsf{L}_\mathsf{X}(x)\big) \wedge \neg\big(\mathsf{L}_\mathsf{X}(x) \wedge \neg\mathsf{X}(x, \ldots, x)\big).$$

These two negated conjuncts require that an element x belongs to the relation L_X if and only if the k-tuple $(x, \ldots, x)$ belongs to X. Then, we replace all the unguarded negated σ-atoms $\neg\mathsf{X}(x, \ldots, x)$ by a unary negated atom $\neg\mathsf{L}_\mathsf{X}(x)$. As this atom is unary, it does not have to be guarded.

Making every σ-atom be guarded by a τ-atom. Consider any negated conjunct $\neg\phi_i(\mathbf{x}_i)$ that contains a σ-atom $\mathsf{X}(\mathbf{z})$ that is guarded by no τ-atom. Let n denote the number of guarding types of $\mathbf{z}$ in a k-tuple. We replace the negated conjunct $\neg\phi_i(\mathbf{x}_i)$ with the following conjunction:

$$\neg\big(\phi_i(\mathbf{x}_i) \wedge \mathsf{R}(\mathbf{y}_1)\big) \wedge \cdots \wedge \neg\big(\phi_i(\mathbf{x}_i) \wedge \mathsf{R}(\mathbf{y}_n)\big),$$

where, for any j in $[n]$, the atom $\mathsf{X}(\mathbf{z})$ has guarding type j in $\mathsf{R}(\mathbf{y}_j)$. There is more than one way to choose $\mathbf{y}_j$, so suppose that any variable of $\mathbf{y}_j$ that is not ought to be in $\mathbf{z}$ is distinct.

The new sentence, denoted by Φ', is less restrictive than the old one. That is, for any τ-structure $\mathfrak{A}$, if $\mathfrak{A} \models \Phi$, then $\mathfrak{A} \models \Phi'$. We need to prove that Φ' is still logically equivalent to Φ. Suppose that there is a τ-structure $\mathfrak{A}$ that satisfies Φ' but does not satisfy the original sentence Φ. Then, let $\mathfrak{A}_0^\sigma$ be a σ-expansion of $\mathfrak{A}$ that satisfies the first-order part $\forall\mathbf{x},\mathbf{y}\ \phi'(\mathbf{x},\mathbf{y})$ of Φ', where $\mathbf{y}$ denote new distinct variables from $\mathbf{y}_1,\ldots,\mathbf{y}_n$. By assumption, $\mathfrak{A}_0^\sigma$ does not satisfy the first-order part $\forall\mathbf{x}\ \phi(\mathbf{x})$ of Φ. Then, as all other negated conjuncts are the same, we have, for some assignment $\mathbf{a}_i$ to $\mathbf{x}_i$, $\mathfrak{A}_0^\sigma \models \phi_i(\mathbf{a}_i)$. Assume that, under this assignment, the tuple $\mathbf{z}$ of the atom $\mathsf{X}(\mathbf{z})$ is mapped to some tuple $\mathbf{c}$ of $\mathfrak{A}$. We have $\mathfrak{A}_0^\sigma \models \mathsf{X}(\mathbf{c})$ and that this tuple $\mathbf{c}$ is not contained in any R-tuple $\mathbf{b}_j$ of $\mathfrak{A}$, as then we would have $\mathfrak{A}_0^\sigma \models \phi_i(\mathbf{a}_i) \wedge \mathsf{R}(\mathbf{b}_j)$, that would contradict one of the n new negated conjuncts of Φ'. Then we modify the σ-expansion $\mathfrak{A}_0^\sigma$ by deleting this tuple $\mathbf{c}$ from the relation $\mathsf{X}^{\mathfrak{A}_0^\sigma}$. We do this procedure for any other X-tuple that does not belong to any R-tuple. If the new σ-expansion, denoted by $\mathfrak{A}_1^\sigma$, does not satisfy $\forall\mathbf{x}\ \phi(\mathbf{x})$, then there is a negated conjunct $\neg\phi_{i_1}(\mathbf{x}_{i_1})$ that contains a negated X-atom $\neg\mathsf{X}(\mathbf{z}_1)$ and there is an assignment $\mathbf{x}_{i_1} \to \mathbf{a}_{i_1}$ that maps $\mathbf{z}_1$ to $\mathbf{c}_1$ such that the tuple $\mathbf{c}_1$ is not contained in $\mathsf{X}^{\mathfrak{A}_1^\sigma}$ and that was contained in $\mathsf{X}^{\mathfrak{A}_0^\sigma}$ – before we deleted those X-tuples that were not guarded by R-tuples. This means that, within ϕ_{i_1}, the negated σ-atom $\neg\mathsf{X}(\mathbf{z}_1)$ is guarded by another σ-atom, say $\mathsf{X}_1(\mathbf{z}_2)$. Let $\mathbf{c}_2$ be the tuple assigned to $\mathbf{z}_2$, then $\mathbf{c}_2$ contains the tuple $\mathbf{c}_1$, and $\mathbf{c}_2$ is in $\mathsf{X}_1^{\mathfrak{A}_1^\sigma} = \mathsf{X}_1^{\mathfrak{A}_0^\sigma}$. We keep modifying the σ-expansion by deleting $\mathbf{c}_2$ from $\mathsf{X}_1^{\mathfrak{A}_1^\sigma}$ and so on until all the negated conjuncts of $\forall\mathbf{x}\ \phi(\mathbf{x})$ are satisfied. At each next step, any negated σ-atom that becomes satisfied is always guarded by another σ-atom, not by a τ-atom because $\mathbf{c} \subseteq \mathbf{c}_1 \subseteq \mathbf{c}_2 \subseteq \ldots$ and, for no R-tuple $\mathbf{b}_j$, we have $\mathbf{c} \subseteq \mathbf{b}_j$. We cannot keep deleting the σ-tuples indefinitely, so this process will halt. This means that $\mathfrak{A} \models \Phi$, which is a contradiction, so the sentences Φ and Φ' are logically equivalent.

We continue modifying Φ' by choosing another σ-atom that is not guarded by a τ-atom and replace the corresponding negated conjunct similarly as above. After every such modification, the result is logically equivalent to the original sentence. We do this modification for every σ-atom of Φ that is not guarded by a τ-atom. At the end, we obtain a sentence that is equivalent to Φ and that any of its σ-atoms is guarded by a τ-atom. We can assume now that Φ already satisfies this property.

Changing the arity of σ-relations. Now we are going to modify the existential signature σ so that the new signature consists of either unary or k-ary relation symbols. That is, they are either unary or have the same arity as R.

Define a function $\mathsf{f}\colon \sigma \to \mathbb{N}$ such that $\mathsf{f}(\mathsf{X})$ is equal to the number of guarding types of an X-tuple in a R-tuple. The new existential signature σ' is defined as follows: $\sigma' = \{\mathsf{X}_i \mid \mathsf{X} \in \sigma \wedge i \in [\mathsf{f}(\mathsf{X})]\}$, each relation symbol has arity k. Now we construct an MMSNP_2 sentence Ψ that is logically equivalent to Φ.

- For any negated conjunct $\neg\phi_i$ of Φ, we add to Ψ its copy denoted by $\neg\psi_i$. The τ-atoms in $\neg\psi_i$ are kept the same as in $\neg\phi_i$.

- As the existential signature is different, we need to replace σ-atoms with σ'-atoms. Let $\mathsf{X}(\mathbf{x})$ be some σ-atom guarded by a τ-atom $\mathsf{R}(\mathbf{y})$. Let $\mathsf{X}(\mathbf{x})$ have guarding type j in $\mathsf{R}(\mathbf{y})$, for some j in $[\mathsf{f}(\mathsf{X})]$. We replace $\mathsf{X}(\mathbf{x})$ with a σ'-atom $\mathsf{X}_j(\mathbf{y})$. Do a similar thing for every other σ-atom.

- Let $\mathbf{y}_1, \mathbf{y}_2$ be two R-tuples that have some variables in common. For l-ary X in σ, let $\mathbf{x}$ be an l-tuple consisting of some of the variables shared between $\mathbf{y}_1$ and $\mathbf{y}_2$. Let $\mathsf{X}(\mathbf{x})$ have guarding type j_1 in $\mathsf{R}(\mathbf{y}_1)$, and guarding type j_2 in $\mathsf{R}(\mathbf{y}_2)$. Then, for any such two intersecting R-tuples $\mathbf{y}_1, \mathbf{y}_2$, for any X in σ, and for any such $\mathbf{x}$ contained in both $\mathbf{y}_1, \mathbf{y}_2$, we add the following negated conjuncts:

$$\neg\big(\mathsf{R}(\mathbf{y}_1)\wedge\mathsf{R}(\mathbf{y}_2)\wedge\mathsf{X}_{j_1}(\mathbf{y}_1)\wedge\neg\mathsf{X}_{j_2}(\mathbf{y}_2)\big)\wedge\neg\big(\mathsf{R}(\mathbf{y}_1)\wedge\mathsf{R}(\mathbf{y}_2)\wedge\neg\mathsf{X}_{j_1}(\mathbf{y}_1)\wedge\mathsf{X}_{j_2}(\mathbf{y}_2)\big). \quad (3.6)$$

  Informally, these two conjuncts mean that, for two R-tuples $\mathbf{y}_1, \mathbf{y}_2$ that intersect, $\mathbf{x}$ having guarding type j_1 in $\mathbf{y}_1$ implies it having guarding type j_2 in $\mathbf{y}_2$, and vice versa.

If a σ'-expansion $\mathfrak{A}^{\sigma'}$ of some τ-structure $\mathfrak{A}$ satisfies all the negated conjuncts of the form as in eq. (3.6), then it is associated with a unique σ-expansion $\mathfrak{A}^{\sigma}$, and this correspondence is one-to-one. The correspondence is defined as follows: for any σ-tuple $\mathsf{X}(\mathbf{a})$ of $\mathfrak{A}^{\sigma}$ that is contained in some τ-tuple $\mathsf{R}(\mathbf{b})$ there is a σ'-tuple $\mathsf{X}_i(\mathbf{b})$ of $\mathfrak{A}^{\sigma'}$ such that $\mathsf{X}(\mathbf{a})$ has guarding type i in $\mathsf{R}(\mathbf{b})$. And backwards, if there is no such σ-tuple, then there is no corresponding σ'-tuple. If we did not add the negated conjuncts of the form as in eq. (3.6) to Ψ, then we would have to consider those σ'-expansions that are associated with no σ-expansion of an input τ-structure. Observe that, given one expansion, the other expansion can be constructed in P-time in the size of the given one.

Let $\mathfrak{A}$ be a τ-structure, let $\mathfrak{A}^{\sigma}$ be a σ-expansion of $\mathfrak{A}$, and let $\mathfrak{A}^{\sigma'}$ be the σ'-expansion that is associated with $\mathfrak{A}^{\sigma}$ by the rule above. By the construction of $\mathfrak{A}^{\sigma'}$, for any negated conjunct $\neg\phi_i(\mathbf{x})$ of Φ and the corresponding negated conjunct $\neg\psi_i(\mathbf{x})$ of Ψ, and for any assignment $\mathbf{a}_i$ to the variables $\mathbf{x}_i$ of the conjuncts, we have $\mathfrak{A}^{\sigma} \models \phi_i(\mathbf{a}_i)$ if and only if $\mathfrak{A}^{\sigma'} \models \psi_i(\mathbf{a}_i)$. Also, $\mathfrak{A}^{\sigma'}$ satisfies all the negated conjuncts of the form as in eq. (3.6). Thus, these two expansions either both satisfy the first-order parts of Φ and Ψ correspondingly or both do not satisfy them. As any σ'-expansion of $\mathfrak{A}$ that is not associated with a σ-expansion violates one of the negated conjuncts from eq. (3.6), we can conclude that $\mathfrak{A} \models \Phi$ if and only if $\mathfrak{A} \models \Psi$. □

Example 3.2.3. Consider the GMSNP sentence Φ given in Example 3.2.1:

$$\exists \mathsf{E}, \mathsf{C}_2\ \forall x, y \begin{pmatrix} \neg\big(\mathsf{C}_3(x,y,z)\wedge\neg\mathsf{E}(x,y)\big)\wedge \\ \neg\big(\mathsf{C}_3(x,y,z)\wedge\neg\mathsf{E}(y,z)\big)\wedge \\ \neg\big(\mathsf{C}_3(x,y,z)\wedge\neg\mathsf{E}(z,x)\big) \end{pmatrix} \wedge \neg\Big(\mathsf{E}(x,y)\wedge\mathsf{E}(y,x)\wedge\neg\mathsf{C}_2(x,y)\Big)\wedge\neg\mathsf{C}_2(x,y).$$

Here, there is one input relation C_3, and two existential relations E and C_2. This query checks if a directed graph encoded by its 3-cycles does not have 2-cycles. We transform Φ to an equivalent MMSNP_2 sentence according to the steps from Proposition 3.2.11. As Φ does not contain equalities, we have no equalities to remove. There are two negated

conjuncts, where σ-atoms are not guarded by τ-atoms: $\neg\Big(\mathsf{E}(x,y)\wedge\mathsf{E}(y,x)\wedge\neg\mathsf{C}_2(x,y)\Big)$ and $\neg\mathsf{C}_2(x,y)$. There are six different injective mappings from a 2-element set to a 3-element set. The negated conjunct $\neg\mathsf{C}_2(x,y)$ is replaced by the following negated conjuncts:

$$\begin{aligned}&\neg\Big(\mathsf{C}_2(x,y)\wedge\mathsf{C}_3(x,y,z)\Big)\wedge\neg\Big(\mathsf{C}_2(x,y)\wedge\mathsf{C}_3(y,x,z)\Big)\wedge\neg\Big(\mathsf{C}_2(x,y)\wedge\mathsf{C}_3(x,z,y)\Big)\wedge\\&\neg\Big(\mathsf{C}_2(x,y)\wedge\mathsf{C}_3(y,z,x)\Big)\wedge\neg\Big(\mathsf{C}_2(x,y)\wedge\mathsf{C}_3(z,x,y)\Big)\wedge\neg\Big(\mathsf{C}_2(x,y)\wedge\mathsf{C}_3(z,y,x)\Big).\end{aligned}\tag{3.7}$$

Similarly, the negated conjunct $\neg\Big(\mathsf{E}(x,y)\wedge\mathsf{E}(y,x)\wedge\neg\mathsf{C}_2(x,y)\Big)$ is replaced by six negated conjuncts. After completing the second step, we modify the existential signature so that all existential relations have arity 1 or 3. Right now, both relations are binary, so, eventually, we will have only ternary relations. For a binary relation C_2 there are 9 ways how a E-tuple can be guarded by a C_3-tuple: 6 ways are as in eq. (3.7), the other three ways are as follows:

$$\neg\Big(\mathsf{C}_2(x,x)\wedge\mathsf{C}_3(x,y,z)\Big),\ \neg\Big(\mathsf{C}_2(x,x)\wedge\mathsf{C}_3(y,x,z)\Big),\ \neg\Big(\mathsf{C}_2(x,x)\wedge\mathsf{C}_3(y,z,x)\Big).$$

Instead of two relations E,C_2 we now use 18 relations: $\mathsf{E}^1,\dots,\mathsf{E}^9,\mathsf{C}_2^1,\dots,\mathsf{C}_2^9$. Every E-atom and C_2-atom is replaced by the corresponding atom of this new signature σ'. Then we add all the conjuncts as in eq. (3.6), therefore, if a σ-tuple is contained in more than one τ-tuple, then each of these τ-tuple must be coloured in the corresponding σ'-colour. $\triangle$

Remark. As the τ-atoms in the MMSNP_2 sentence Ψ are the same as in Φ, we know that if Φ is connected, then Ψ is connected. By Lemma 3.2.10, we consider only connected GMSNP sentences, so further we can consider only connected MMSNP_2 sentences.

Chapter 4

MMSNP_2 on ω-categorical structures

We show that any MMSNP_2 problem can be reduced to an MMSNP problem, this reduction is provided by a functor between the inputs. Then, we consider an infinitary extension of MMSNP that allows a sentence to have countably many negated conjuncts. We show that any MMSNP_2 problem is P-time equivalent to an infinite regular MMSNP, where the word "regular" has a similar meaning as in regular languages. Regularity of an infinite MMSNP sentence provides the existence of an ω-categorical relational structure with an equivalent CSP. For every MMSNP_2 sentence there also exists an ω-categorical structure with an equivalent CSP. Finally, we show that the functorial image of this structure is homomorphically equivalent to the ω-categorical template associated with the infinite MMSNP.

4.1 Reduction from MMSNP_2 to MMSNP

Clearly, any MMSNP problem is also an MMSNP_2 problem, by the definition. In this section we show that any MMSNP_2 problem reduces to some MMSNP problem. Firstly, we describe how to construct the MMSNP sentence for a given one of MMSNP_2. And then, for any relational τ-structure of the MMSNP_2 input, we construct the corresponding relational $\tilde{\tau}$-structure of the MMSNP input. It happens that this construction is functorial. That is, it is described by a functor from the category of τ-structures $\mathbf{Struct}[\tau]$ to the category of $\tilde{\tau}$-structures $\mathbf{Struct}[\tilde{\tau}]$. In the category of τ-structures, the objects are τ-structures themselves and the morphisms are homomorphisms between structures.

By Theorem 3.2.5 and Proposition 3.2.11 from the previous chapter, we can assume without loss of generality that the input signature consists of a unique relation symbol. Let $\tau = \{\mathsf{R}\}$ be the input signature of MMSNP_2 sentences and $\sigma = \sigma_{\mathsf{V}} \uplus \sigma_{\mathsf{T}}$ be the existential signature, where $\sigma_{\mathsf{V}} = \{\mathsf{M}_1, \dots, \mathsf{M}_u\}$ and $\sigma_{\mathsf{T}} = \{\mathsf{X}_1, \dots, \mathsf{X}_s\}$. Let R have arity k, each M in σ_{V} be unary, and each X in σ_{T} be k-ary. We introduce signatures $\tilde{\tau}$ and $\tilde{\sigma}$ and show that any MMSNP_2 problem over τ reduces to some MMSNP problem over $\tilde{\tau}$. Construction 7 describes how we modify the sentence, Construction 8 describes how we modify the input of the corresponding problem. These two transformations are similar: the main idea is to add to every tuple a special vertex that represents the colour of this tuple.

Construction 7. The new input signature $\tilde{\tau}$ has unary relation symbols V and T, and a $(k+1)$-ary relation symbol $\tilde{\mathsf{R}}$. The new existential signature $\tilde{\sigma} := \tilde{\sigma}_{\mathsf{V}} \uplus \tilde{\sigma}_{\mathsf{T}}$ contains a unary relation symbol $\tilde{\mathsf{X}}$, for any X in σ.

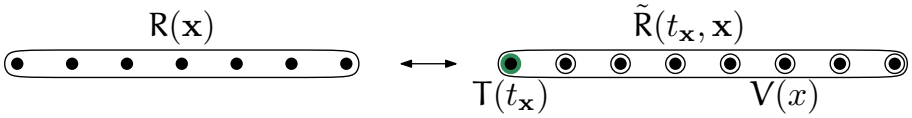


Figure 4.1: The relation between τ-tuples and $\tilde{\tau}$-tuples. The long closed curves are the relations R and $\tilde{\mathsf{R}}$, the green disc is the relation T, the white circles denote the relation V.

Any MMSNP_2 τ-sentence Φ transforms to an MMSNP $\tilde{\tau}$-sentence $\tilde{\Phi}$ as follows. Every τ-atom $\mathsf{R}(\mathbf{x})$ of Φ is replaced by the following conjunction that is displayed on Figure 4.1 on page 76:

$$\tilde{\mathsf{R}}(t, \mathbf{x}) \wedge \mathsf{T}(t) \wedge \bigwedge_{x \text{ in } \mathbf{x}} \mathsf{V}(x).$$

Every unary σ-atom $\mathsf{M}(x)$ is replaced by the following conjunction:

$$\tilde{\mathsf{M}}(x) \wedge \mathsf{V}(x).$$

Every k-ary σ-atom $\mathsf{X}(\mathbf{x})$ that is guarded by a τ-atom $\mathsf{R}(\mathbf{x})$ is replaced by a unary $\tilde{\sigma}_{\mathsf{T}}$-atom $\tilde{\mathsf{X}}(t)$, where t is a new variable that appears when we replace $\mathsf{R}(\mathbf{x})$ with $\tilde{\mathsf{R}}(t, \mathbf{x})$. Finally, we add to $\tilde{\Phi}$ all the negated conjuncts of the two types displayed on Figure 4.2 on page 77. For any $\tilde{\mathsf{X}}$ in $\tilde{\sigma}_{\mathsf{T}}$, we add the negated conjunct $\neg\big(\mathsf{V}(x) \wedge \tilde{\mathsf{X}}(x)\big)$. For any $\tilde{\mathsf{M}}$ in $\tilde{\sigma}_{\mathsf{V}}$, we add the negated conjunct $\neg\big(\tilde{\mathsf{M}}(x) \wedge \mathsf{T}(x)\big)$. This is the end of Construction 7.

Example 4.1.1. Let Φ be the MMSNP_2 sentence that is associated with the problem No-Monochromatic-Arc-Triangle from Example 3.2.2 on page 64:

$$\exists \mathsf{B}, \mathsf{W}\, \forall x, y, z \begin{pmatrix} \neg\big(\mathsf{E}(x,y) \wedge \mathsf{B}(x,y) \wedge \mathsf{W}(x,y)\big) \wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \neg\mathsf{B}(x,y) \wedge \neg\mathsf{W}(x,y)\big) \wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{E}(y,z) \wedge \mathsf{E}(z,x) \wedge \mathsf{B}(x,y) \wedge \mathsf{B}(y,z) \wedge \mathsf{B}(z,x)\big) \wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{E}(y,z) \wedge \mathsf{E}(z,x) \wedge \mathsf{W}(x,y) \wedge \mathsf{W}(y,z) \wedge \mathsf{W}(z,x)\big) \end{pmatrix}$$

Then, the MMSNP sentence $\tilde{\Phi}$ obtained by Construction 7 can be written as follows:

$$\exists \mathsf{B}, \mathsf{W}\, \forall t_{xy}, t_{yz}, t_{zx}, x, y, z$$

$$\begin{pmatrix} \neg\big(\mathsf{V}(x) \wedge \mathsf{B}(x)\big) \wedge \neg\big(\mathsf{V}(x) \wedge \mathsf{W}(x)\big) \wedge \\ \neg\big(\tilde{\mathsf{E}}(t_{xy}, x, y) \wedge \mathsf{T}(t_{xy}) \wedge \mathsf{V}(x) \wedge \mathsf{V}(y) \wedge \mathsf{B}(t_{xy}) \wedge \mathsf{W}(t_{xy})\big) \wedge \\ \neg\big(\tilde{\mathsf{E}}(t_{xy}, x, y) \wedge \mathsf{T}(t_{xy}) \wedge \mathsf{V}(x) \wedge \mathsf{V}(y) \wedge \neg\mathsf{B}(t_{xy}) \wedge \neg\mathsf{W}(t_{xy})\big) \wedge \\ \neg\Big(\tilde{\mathsf{E}}(t_{xy}, x, y) \wedge \tilde{\mathsf{E}}(t_{yz}, y, z) \wedge \tilde{\mathsf{E}}(t_{zx}, z, x) \wedge \\ \wedge \mathsf{T}(t_{xy}) \wedge \mathsf{T}(t_{yz}) \wedge \mathsf{T}(t_{zx}) \wedge \mathsf{V}(x) \wedge \mathsf{V}(y) \wedge \mathsf{V}(z) \wedge \mathsf{B}(t_{xy}) \wedge \mathsf{B}(t_{yz}) \wedge \mathsf{B}(t_{zx})\Big) \wedge \\ \neg\Big(\tilde{\mathsf{E}}(t_{xy}, x, y) \wedge \tilde{\mathsf{E}}(t_{yz}, y, z) \wedge \tilde{\mathsf{E}}(t_{zx}, z, x) \wedge \\ \wedge \mathsf{T}(t_{xy}) \wedge \mathsf{T}(t_{yz}) \wedge \mathsf{T}(t_{zx}) \wedge \mathsf{V}(x) \wedge \mathsf{V}(y) \wedge \mathsf{V}(z) \wedge \mathsf{W}(t_{xy}) \wedge \mathsf{W}(t_{yz}) \wedge \mathsf{W}(t_{zx})\Big) \end{pmatrix}$$

We first add two negated conjuncts from Figure 4.2, and then modify the other conjuncts of Φ as described in the construction. △

Now we explain how to transform an input structure $\mathfrak{A}$ of MMSNP_2 to an input structure $\tilde{\mathfrak{A}}$ of MMSNP.

$\neg(\mathsf{V}(x) \wedge \tilde{\mathsf{X}}(x))$ $\qquad$ $\neg(\mathsf{T}(x) \wedge \tilde{\mathsf{M}}(x))$

Figure 4.2: The two additional negated conjunct that are added to $\tilde{\Phi}$. The green disc is the T relation. The white circle is the V relation. The blue disc is the $\tilde{\mathsf{X}}$ relation corresponding to some X in σ_{T}. The grey disc with the black boundary is the $\tilde{\mathsf{M}}$ relation corresponding to some M in σ_{V}.

Construction 8. The domain $\tilde{A}$ of $\tilde{\mathfrak{A}}$ is the disjoint union $\tilde{A}_{\mathsf{T}} \uplus \tilde{A}_{\mathsf{V}}$. Here, $\tilde{A}_{\mathsf{V}}$ is equal to the domain A of $\mathfrak{A}$. And, for any tuple $\mathbf{a}$ of $\mathsf{R}^{\mathfrak{A}}$ there exists a vertex $t_{\mathbf{a}}$ in $\tilde{A}_{\mathsf{T}}$. That is, $\tilde{A}_{\mathsf{T}} := \{t_{\mathbf{a}} \mid \mathbf{a} \in \mathsf{R}^{\mathfrak{A}}\}$. The relation $\mathsf{V}^{\tilde{\mathfrak{A}}}$ contains all the elements that are associated with the elements of A, that is, $\mathsf{V}^{\tilde{\mathfrak{A}}} := \tilde{A}_{\mathsf{V}}$. The relation T contains all the vertices of the type $t_{\mathbf{a}}$: $\mathsf{T}^{\tilde{\mathfrak{A}}} := \tilde{A}_{\mathsf{T}}$. For any tuple $\mathbf{a}$ of $\mathsf{R}^{\mathfrak{A}}$, the relation $\tilde{\mathsf{R}}^{\tilde{\mathfrak{A}}}$ contains a tuple $(t_{\mathbf{a}}, \mathbf{a})$. This construction is similar to the one from Figure 4.1 on page 76. Transforming a finite τ-structure $\mathfrak{A}$ to a $\tilde{\tau}$-structure $\tilde{\mathfrak{A}}$ can be achieved in P-time in $|A|$. This finishes the Construction 8.

Example 4.1.2. See Figure 4.3. Let $\mathfrak{A}$ be a directed 3-cycle: $A = \{x, y, z\}$, $\mathsf{E}^{\mathfrak{A}} = \{(x, y), (y, z), (z, x)\}$. Then the domain $\tilde{A}$ of $\tilde{\mathfrak{A}}$ is equal to the disjoint union $\tilde{A}_{\mathsf{T}} \uplus \tilde{A}_{\mathsf{V}}$, where $\tilde{A}_{\mathsf{V}} := A = \{x, y, z\}$, and $\tilde{A}_{\mathsf{T}} := \{t_{\mathbf{x}} \mid \mathbf{x} \in \mathsf{E}^{\mathfrak{A}}\} = \{t_{xy}, t_{yz}, t_{zx}\}$. The relations of $\tilde{\mathfrak{A}}$ are defined as follows: $\mathsf{V}^{\tilde{\mathfrak{A}}} := \tilde{A}_{\mathsf{V}}$, $\mathsf{T}^{\tilde{\mathfrak{A}}} := \tilde{A}_{\mathsf{T}}$, and $\tilde{\mathsf{E}}^{\tilde{\mathfrak{A}}} := \{(t_{xy}, x, y), (t_{yz}, y, z), (t_{zx}, z, x)\}$. $\triangle$

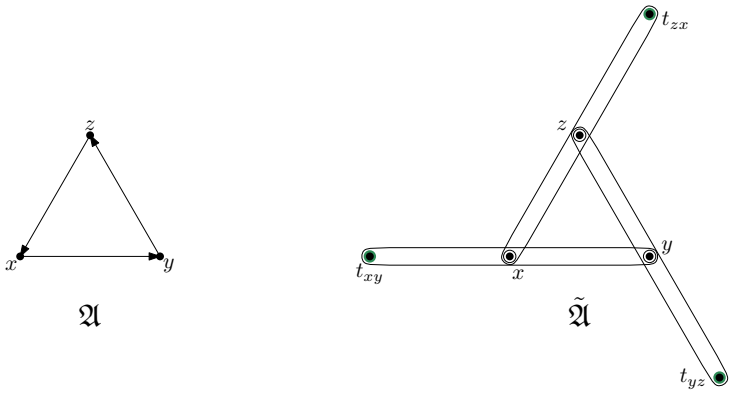


Figure 4.3: An example of Construction 8.

Remark. Further, we will usually call any element x a V-*vertex* if $\mathsf{V}(x)$ is satisfied and a T-*vertex* if $\mathsf{T}(x)$ is satisfied. Also, the relations V and T are not really necessary because they can be defined by the tuple coordinate where the considered element of $\tilde{\mathfrak{A}}$ appears. If x appears on the first coordinate, then it is a T-vertex. If it appears anywhere else except for the first coordinate, then it is a V-vertex.

We prove now that $\mathrm{SAT}(\Phi)$ is reduced in P-time to $\mathrm{SAT}(\tilde{\Phi})$: $\mathrm{SAT}(\Phi) \leq_p \mathrm{SAT}(\tilde{\Phi})$.

Proposition 4.1.1. *Let Φ be in MMSNP_2 and $\tilde{\Phi}$ in MMSNP be obtained from Φ by Construction 7. Then, for any τ-structure $\mathfrak{A}$ there exists a $\tilde{\tau}$-structure $\tilde{\mathfrak{A}}$ constructible from $\mathfrak{A}$ in P-time such that $\mathfrak{A} \models \Phi$ if and only if $\tilde{\mathfrak{A}} \models \tilde{\Phi}$.*

Proof. For any τ-structure $\mathfrak{A}$, we obtain $\tilde{\mathfrak{A}}$ according to Construction 8. We need to show that these two structures are equivalent with respect to being models for Φ and $\tilde{\Phi}$.

Suppose that $\mathfrak{A} \models \Phi$, that is, there is a σ-expansion $\mathfrak{A}^\sigma$ of $\mathfrak{A}$ such that any negated conjunct of Φ is satisfied. We can assume that, for any X in σ_T, $\mathsf{X}^{\mathfrak{A}^\sigma} \subseteq \mathsf{R}^{\mathfrak{A}}$, otherwise, if we delete all the X-tuples not belonging to $\mathsf{R}^{\mathfrak{A}}$, then any negated conjunct will still be satisfied. Then we choose a $\tilde{\sigma}$-expansion $\tilde{\mathfrak{A}}^{\tilde{\sigma}}$ of $\tilde{\mathfrak{A}}$ such that, for any X in σ_T and any k-tuple $\mathbf{a}$ in $\mathsf{R}^{\mathfrak{A}}$, the T-vertex $t_{\mathbf{a}}$ is in $\tilde{\mathsf{X}}^{\tilde{\mathfrak{A}}^{\tilde{\sigma}}}$ if and only if $\mathbf{a}$ is in $\mathsf{X}^{\mathfrak{A}^\sigma}$. And, for any M in σ_V and any a in A, a is in $\tilde{\mathsf{M}}^{\tilde{\mathfrak{A}}^{\tilde{\sigma}}}$ if and only if a is in $\mathsf{M}^{\mathfrak{A}^\sigma}$. By the construction of $\tilde{\Phi}$ and $\tilde{\mathfrak{A}}$, we have that $\tilde{\mathfrak{A}}^{\tilde{\sigma}}$ satisfies any negated conjunct of $\tilde{\Phi}$ as well.

Suppose that $\tilde{\mathfrak{A}} \models \tilde{\Phi}$; then there is a $\tilde{\sigma}$-expansion that satisfies any negated conjunct of $\tilde{\Phi}$. In particular, it satisfies any negated conjunct of one of the types from Figure 4.2 on page 77. This means that no V-vertex is coloured by a $\tilde{\sigma}_\mathsf{T}$-relation and that no T-vertex is coloured by a $\tilde{\sigma}_\mathsf{V}$-relation. Then there exists a σ-expansion $\mathfrak{A}^\sigma$ of $\mathfrak{A}$ such that $\mathfrak{A}^\sigma$ and $\tilde{\mathfrak{A}}^{\tilde{\sigma}}$ are related to each other as in the previous paragraph of this proof. Then, by the constructions of $\tilde{\Phi}$ and $\tilde{\mathfrak{A}}$, if some negated conjunct of Φ is not satisfied, then the corresponding negated conjunct of $\tilde{\Phi}$ is not satisfied in $\tilde{\mathfrak{A}}$. □

Using the method from the proof of Proposition 4.1.1, we can extend Construction 8 onto $(\tau \uplus \sigma)$-structures, where all k-ary existential relations $\mathsf{X} \in \sigma_\mathsf{T}$ are interpreted as subsets of the τ-relation R.

Construction 9. Let $\mathfrak{A}$ be a τ-structure, and $\mathfrak{A}^\sigma$ be one of its σ-expansions such that, for any X in σ_T, $\mathsf{X}^{\mathfrak{A}^\sigma} \subseteq \mathsf{R}^{\mathfrak{A}^\sigma} = \mathsf{R}^{\mathfrak{A}}$. Let $\tilde{\mathfrak{A}}$ be obtained from $\mathfrak{A}$ by Construction 8. The $\tilde{\sigma}$-expansion $\tilde{\mathfrak{A}}^{\tilde{\sigma}}$ of $\tilde{\mathfrak{A}}$ associated with $\mathfrak{A}^\sigma$ is constructed as follows. For any X in σ_T and any R-tuple $\mathbf{a}$ in $\mathsf{R}^{\mathfrak{A}}$, the T-vertex $t_{\mathbf{a}}$ is in $\tilde{\mathsf{X}}^{\tilde{\mathfrak{A}}^{\tilde{\sigma}}}$ if and only if $\mathbf{a}$ is in $\mathsf{X}^{\mathfrak{A}^\sigma}$. And, for any M in σ_V and any a in A, a is in $\tilde{\mathsf{M}}^{\tilde{\mathfrak{A}}^{\tilde{\sigma}}}$ if and only if a is in $\mathsf{M}^{\mathfrak{A}^\sigma}$.

The $(\tilde{\tau} \uplus \tilde{\sigma})$-structures obtained from $(\tau \uplus \sigma)$-structures by Construction 9 agree on homomorphisms.

Proposition 4.1.2. *Let $\tilde{\mathfrak{A}}$ and $\tilde{\mathfrak{B}}$ be constructed from $(\tau \uplus \sigma)$-structures $\mathfrak{A}$ and $\mathfrak{B}$. Then, $\mathfrak{A}$ homomorphically maps to $\mathfrak{B}$ if and only if $\tilde{\mathfrak{A}}$ homomorphically maps to $\tilde{\mathfrak{B}}$.*

Proof. Let $\mathsf{h}\colon \mathfrak{A} \to \mathfrak{B}$ be a homomorphism. Construct a map $\tilde{\mathsf{h}}$ as follows. For any V-vertex v of $\tilde{\mathfrak{A}}$, its image $\tilde{\mathsf{h}}(v)$ equals $\mathsf{h}(v)$. Any T-vertex $t_{\mathbf{x}}$ of $\tilde{\mathfrak{A}}$ is associated with some R-tuple $\mathbf{x}$ of $\mathfrak{A}$. The homomorphism h requires that $\mathsf{h}(\mathbf{x})$ is an R-tuple in $\mathfrak{B}$. Then, $\tilde{\mathfrak{B}}$ contains a T-vertex $t_{\mathsf{h}(\mathbf{x})}$. We set $\tilde{\mathsf{h}}(t_{\mathbf{x}}) := t_{\mathsf{h}(\mathbf{x})}$. For any $\tilde{\mathsf{R}}$-tuple $(t_{\mathbf{x}}, \mathbf{x})$ of $\tilde{\mathfrak{A}}$, its $\tilde{h}$-image is $\big(t_{\mathsf{h}(\mathbf{x})}, \mathsf{h}(\mathbf{x})\big)$, by construction, it is an $\tilde{\mathsf{R}}$-tuple of $\tilde{\mathfrak{B}}$. Any $\tilde{\sigma}_\mathsf{T}$-coloured vertex $t_{\mathbf{x}}$ of $\tilde{\mathfrak{A}}$ is associated with a σ_T-tuple $\mathbf{x}$ of $\mathfrak{A}$. Its image $\mathsf{h}(\mathbf{x})$ is a σ_T-tuple of $\mathfrak{B}$. So, by construction, the image $\tilde{\mathsf{h}}(t_{\mathbf{x}})$ is coloured in the same $\tilde{\sigma}_\mathsf{T}$-colour. $\tilde{\mathsf{h}}$ is the same as h, once restricted on the V-vertices, so any $\tilde{\sigma}_\mathsf{V}$-colour is also preserved.

Let $\tilde{\mathsf{h}}\colon \tilde{\mathfrak{A}} \to \tilde{\mathfrak{B}}$ be a homomorphism. Let $\mathsf{h}\colon A \to B$ be associated with the restriction of $\tilde{\mathsf{h}}$ on the V-vertices of $\tilde{\mathfrak{A}}$. Any σ_V-colour is preserved. Consider an R-tuple $\mathbf{x}$ of $\mathfrak{A}$ such that $\mathbf{x}$ also belongs to $\mathsf{X}^{\mathfrak{A}}$, for some X in σ_T. It is associated with an $\tilde{\mathsf{R}}$-tuple $(t_{\mathbf{x}}, \mathbf{x})$ of $\tilde{\mathfrak{A}}$, where $t_{\mathbf{x}}$ is coloured in $\tilde{\mathsf{X}}$. This $\tilde{\mathsf{R}}$-tuple is mapped to an $\tilde{\mathsf{R}}$-tuple $\big(\tilde{\mathsf{h}}(t_{\mathbf{x}}), \tilde{\mathsf{h}}(\mathbf{x})\big)$, where $\tilde{\mathsf{h}}(t_{\mathbf{x}})$ is coloured in $\tilde{\mathsf{X}}$. This tuple is associated with an R-tuple $\tilde{\mathsf{h}}(\mathbf{x})$ that is coloured in X. As $\tilde{\mathsf{h}}(\mathbf{x}) = \mathsf{h}(\mathbf{x})$, we conclude that h is a homomorphism. □

Construction 9 induces a functor $\tilde{\mathsf{D}}\colon \mathbf{Struct}[\tau \uplus \sigma] \to \mathbf{Struct}[\tilde{\tau} \uplus \tilde{\sigma}]$ between the categories of $(\tau \uplus \sigma)$-structures and $(\tilde{\tau} \uplus \tilde{\sigma})$-structures such that, for any $\mathfrak{A}$ in $\mathbf{Struct}[\tau \uplus \sigma]$, $\tilde{\mathsf{D}}(\mathfrak{A})$ is equal to $\tilde{\mathfrak{A}}$.

Proposition 4.1.3. *There exists a functor* $\tilde{\mathsf{D}}$ *between the categories* $\mathbf{Struct}[\tau \uplus \sigma]$ *and* $\mathbf{Struct}[\tilde{\tau} \uplus \tilde{\sigma}]$.

Proof. For an object $\mathfrak{A}$ in $\mathbf{Struct}[\tau \uplus \sigma]$, we put $\tilde{\mathsf{D}}(\mathfrak{A}) := \tilde{\mathfrak{A}}$. For an arrow $\mathfrak{h}\colon \mathfrak{A} \to \mathfrak{B}$ in $\mathbf{Struct}[\tau \uplus \sigma]$, we put $\tilde{\mathsf{D}}(\mathfrak{h}) := \tilde{\mathfrak{h}}$, as in Proposition 4.1.2.

It is routine to check that $\tilde{\mathsf{D}}$ preserves the identity arrows and the associative property. □

Remark. The functor $\tilde{\mathsf{D}}$ can be naturally "reduced" on the category $\mathbf{Struct}[\tau]$, as any τ-structure is the τ-reduct of a $(\tau \uplus \sigma)$-structure with every σ-relation being interpreted as the empty set.

The following definition describes those structures in the MMSNP input that are associated with some input of MMSNP_2. They are called good because we know how to treat them.

Definition 4. A structure is called *good* if it is isomorphic to $\tilde{\mathsf{D}}(\mathfrak{A})$, for some $\mathfrak{A}$ in $\mathbf{Struct}[\tau] \cup \mathbf{Struct}[\tau \uplus \sigma]$.

4.2 Normal$_1$ form for MMSNP$_2$

We show that MMSNP_2 sentences can be rewritten in a certain form that we use further. The transformed sentence is logically equivalent to the original one. Once a sentence is in this new form, we are able to represent the corresponding MMSNP_2 problem as a forbidden patterns problem (FPP), see Section 4.3 and Section 4.4.

As usual, we consider connected MMSNP_2 sentences over a relational signature with one k-ary symbol: $\tau = \{\mathsf{R}\}$. The existential relations (colours) of Φ are from some relational signature $\sigma = \sigma_{\mathsf{T}} \uplus \sigma_{\mathsf{V}}$, where all symbols of $\sigma_{\mathsf{T}} = \{\mathsf{X}_1, \dots, \mathsf{X}_s\}$ are k-ary, and all symbols of $\sigma_{\mathsf{V}} = \{\mathsf{M}_1, \dots, \mathsf{M}_u\}$ are unary.

Definition 5. An MMSNP_2 sentence Φ is said to be in *normal$_1$ form* if the following conditions hold.

1. Any element and any R-tuple have at least one colour. That is, the first two negated conjuncts of Φ are:
$$\neg\big(\neg\mathsf{M}_1(x) \wedge \dots \wedge \neg\mathsf{M}_u(x)\big) \wedge \neg\big(\mathsf{R}(\mathbf{x}) \wedge \neg\mathsf{X}_1(\mathbf{x}) \wedge \dots \wedge \neg\mathsf{X}_s(\mathbf{x})\big).$$

2. Any element and any R-tuple have at most one colour. That is, for any distinct M, M' in σ_{V} and X, X' in σ_{T}, Φ contains the negated conjuncts:
$$\neg\big(\mathsf{M}(x) \wedge \mathsf{M}'(x)\big), \quad \neg\big(\mathsf{R}(\mathbf{x}) \wedge \mathsf{X}(\mathbf{x}) \wedge \mathsf{X}'(\mathbf{x})\big).$$

3. The clauses of Φ are fully colored. This means that, for any negated conjunct $\neg\phi_i$ of Φ except for the first two ones and for any variable x of ϕ_i, ϕ_i contains an atom $\mathsf{M}(x)$, for some M in σ_{V}. And, similarly, for any negated conjunct $\neg\phi_i$ of Φ except the first two ones and for any R-tuple $\mathbf{x}$ of ϕ_i, ϕ_i contains an atom $\mathsf{X}(\mathbf{x})$, for some X in σ_{T}.

4. Any negated conjunct $\neg\phi_i$ of Φ is *biconnected*, *i.e.*, the conjunction ϕ_i cannot be divided in two parts ψ_1 and ψ_2 such that the negated conjunct $\neg\phi_i$ equals

$$\neg\big(\psi_1(x,\mathbf{y}) \wedge \psi_2(x,\mathbf{z})\big),$$

where x is some variable of ϕ_i and $\mathbf{y},\mathbf{z}$ are two disjoint tuples of variables of ϕ_i.

5. There are no implicit small clauses. That is, any $(\tau \uplus \sigma)$-structure $\mathfrak{A}$ with n elements satisfies the first-order parts of Φ if and only if it satisfies all the negated conjuncts with at most n variables.

If Φ is an MMSNP sentence (notice the absence of the 2 subscript), then this normal form is used by Feder and Vardi to prove the following theorem.

Theorem 4.2.1 ([FV98])**.** *For any MMSNP sentence Φ there is a finite relational structure $\mathfrak{A}$ such that SAT(Φ) and CSP($\mathfrak{A}$) are P-time equivalent under randomized P-time reductions.*

In the rest of this section, we show how to transform a sentence to have normal$_1$ form. We do a similar transformation to the one of Feder and Vardi in [FV98] which is described with more details in [BMM18]. In our case, the only difference is to make the σ_{T}-existential relations form a partition on the R-tuples of an input structure and to ensure that any clause is fully coloured with respect to R-tuples. We proceed as for unary existential relations by Feder and Vardi in [FV98].

Proposition 4.2.2. *For any MMSNP_2 sentence Φ there is an MMSNP_2 sentence Φ' in normal$_1$ form which is logically equivalent to Φ.*

Proof.

Make every conjunct biconnected. If there is a negated conjunct $\neg\phi_i$ that has the form:

$$\neg\big(\psi_1(x,\mathbf{y}) \wedge \psi_2(x,\mathbf{z})\big),$$

then we augment σ_{V} with a new unary relation P and replace $\neg\phi_i$ with the two negated conjuncts:

$$\neg\big(\psi_1(x,\mathbf{y}) \wedge \mathsf{P}(x)\big), \quad \neg\big(\psi_2(x,\mathbf{z}) \wedge \neg\mathsf{P}(x)\big).$$

Make all implicit small clauses explicit. This transformation has the same nature as Construction 3 on page 60. Let $\neg\phi(x_1,\ldots,x_l)$ be some negated conjunct of Φ. Let x be a variable that does not appear among the variables $x_1,\ldots,x_n$ used in Φ. Construct the negated conjunct $\neg\phi(y_1,\ldots,y_l)$, where, for any i in $[l]$, y_i equals either x_i or x, and, for at least two different $i,j \leq l$, we have $y_i = y_j = x$. If this conjunct is biconnected, then we add it to Φ. If it is not biconnected, then we use the procedure from the first step and, instead of $\neg\phi(y_1,\ldots,y_l)$, add the corresponding conjunction of biconnected negated conjuncts. Do this procedure for any negated conjunct of Φ, this process eventually ends as the number of variables reduces each time. Let now $\mathfrak{A}$ be a $(\tau \uplus \sigma)$-structure with n elements. Suppose that it does not satisfy a negated conjunct $\neg\phi(x_1,\ldots,x_l)$ with $l > n$ variables. Then some variables must have a common element of $\mathfrak{A}$ which is

assigned to them. Then we find the corresponding negated conjunct $\neg\phi(y_1, \dots, y_n)$ (if it is biconnected) or one of the biconnected negated conjuncts obtained from it and that is not satisfied by $\mathfrak{A}$. Each of these conjuncts has at most n variables.

Partition and fully coloured clauses. We describe the procedure only for k-ary existential relations of σ_{T}. The procedure for unary relations is similar to the procedure for the case of MMSNP as described in [BMM18]. We check if for any R-atom $\mathsf{R}(\mathbf{x})$ of Φ and for any X in σ_{T} there is either the X-atom $\mathsf{X}(\mathbf{x})$ or the negated X-atom $\neg\mathsf{X}(\mathbf{x})$ in the same conjunct as $\mathsf{R}(\mathbf{x})$. If there is a negated conjunct $\neg\phi_i$ that does not satisfy this condition, then we replace it with

$$\neg\big(\phi_i \wedge \mathsf{X}(\mathbf{x})\big) \wedge \neg\big(\phi_i \wedge \neg\mathsf{X}(\mathbf{x})\big).$$

Once this property is achieved, we replace the existential signature σ_{T} with a signature $\mathbf{2}^{\sigma_{\mathsf{T}}}$, where each relation symbol is associated with a subset of σ_{T}. It is uniquely determined which new relation we should assign to any R-atom $\mathsf{R}(\mathbf{x})$ because now, for any X in σ_{T}, the negated conjunct containing the atom $\mathsf{R}(\mathbf{x})$ also contains either $\mathsf{X}(\mathbf{x})$ or $\neg\mathsf{X}(\mathbf{x})$ but not both at the same time. We demand that the new relations must form a partition of the R-tuples of an input structure by adding the following negated conjuncts:

$$\neg\left(\mathsf{R}(\mathbf{x}) \wedge \bigwedge_{\mathsf{X}\in\mathbf{2}^{\sigma_{\mathsf{T}}}} \neg\mathsf{X}(\mathbf{x})\right) \wedge \bigwedge_{\mathsf{X},\mathsf{Y}\in\mathbf{2}^{\sigma_{\mathsf{T}}}} \neg\big(\mathsf{R}(\mathbf{x}) \wedge \mathsf{X}(\mathbf{x}) \wedge \mathsf{Y}(\mathbf{x})\big).$$

□

Example 4.2.1. Let Φ describe the No-Monochromatic-Arc-Triangle problem. Similarly as in Example 3.2.2. Φ can be written as follows:

$$\exists \mathsf{B}, \mathsf{W}\, \forall x, y, z \begin{pmatrix} \neg\big(\mathsf{E}(x,y) \wedge \neg\mathsf{B}(x,y) \wedge \neg\mathsf{W}(x,y)\big)\wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{B}(x,y) \wedge \mathsf{W}(x,y)\big)\wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{E}(y,z) \wedge \mathsf{E}(z,x) \wedge \mathsf{B}(x,y) \wedge \mathsf{B}(y,z) \wedge \mathsf{B}(z,x)\big)\wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{E}(y,z) \wedge \mathsf{E}(z,x) \wedge \mathsf{W}(x,y) \wedge \mathsf{W}(y,z) \wedge \mathsf{W}(z,x)\big) \end{pmatrix}$$

This sentence is not in normal$_1$ form by several reasons. We do not restrict any element to be coloured, as there are no unary existential relations. Consequently, the clauses are not fully coloured. However, all of the conjuncts are biconnected. But there are implicit small clauses, as the structure consisting of a single vertex with a loop does not satisfy Φ, but, as any negated conjunct has at least 2 variables, this structure satisfies any conjunct with at most 1 variable.

We repeat the steps of the proof of Proposition 4.2.2. We skip the first step, as all conjuncts are already biconnected. Then, we explicitly write all implicit small clauses. Up to isomorphism, there are two digraphs that can be obtained from the directed cycle of length 3. They are displayed on Figure 4.4. Observe that both of them are biconnected.

As only monochromatic-arc triangles are forbidden, for every triangle with coloured arcs, we add the two images with all arcs having the same colour. After this step, the

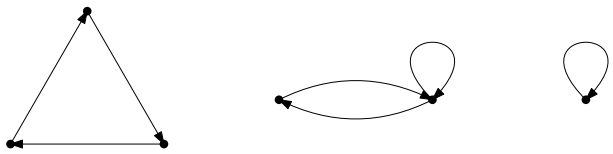


Figure 4.4: The directed 3-cycle and its homomorphic images.

sentence Φ is written as follows:

$$\exists \mathsf{B}, \mathsf{W}\, \forall x, y, z \begin{pmatrix} \neg\big(\mathsf{E}(x,y) \wedge \neg\mathsf{B}(x,y) \wedge \neg\mathsf{W}(x,y)\big)\wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{B}(x,y) \wedge \mathsf{W}(x,y)\big)\wedge \\ \neg\big(\mathsf{E}(x,x) \wedge \mathsf{B}(x,x)\big)\wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{E}(y,x) \wedge \mathsf{E}(x,x) \wedge \mathsf{B}(x,y) \wedge \mathsf{B}(y,x) \wedge \mathsf{B}(x,x)\big)\wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{E}(y,z) \wedge \mathsf{E}(z,x) \wedge \mathsf{B}(x,y) \wedge \mathsf{B}(y,z) \wedge \mathsf{B}(z,x)\big)\wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{W}(x,x)\big)\wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{E}(y,x) \wedge \mathsf{E}(x,x) \wedge \mathsf{W}(x,y) \wedge \mathsf{W}(y,x) \wedge \mathsf{W}(x,x)\big)\wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{E}(y,z) \wedge \mathsf{E}(z,x) \wedge \mathsf{W}(x,y) \wedge \mathsf{W}(y,z) \wedge \mathsf{W}(z,x)\big) \end{pmatrix}$$

Now we only need to introduce one unary existential relation M, require that every element must be coloured with it, and colour every variable of any conjunct with this colour. After this change, the sentence Φ will be in normal$_1$ form. It will be written as follows:

$$\exists \mathsf{M}, \mathsf{B}, \mathsf{W}\, \forall x, y, z$$

$$\begin{pmatrix} \neg\big(\neg\mathsf{M}(x)\big) \wedge \neg\big(\mathsf{E}(x,y) \wedge \neg\mathsf{B}(x,y) \wedge \neg\mathsf{W}(x,y)\big)\wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{M}(x) \wedge \mathsf{M}(y) \wedge \mathsf{B}(x,y) \wedge \mathsf{W}(x,y)\big)\wedge \\ \neg\big(\mathsf{E}(x,x) \wedge \mathsf{M}(x) \wedge \mathsf{B}(x,x)\big)\wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{E}(y,x) \wedge \mathsf{E}(x,x) \wedge \mathsf{M}(x) \wedge \mathsf{M}(y) \wedge \mathsf{B}(x,y) \wedge \mathsf{B}(y,x) \wedge \mathsf{B}(x,x)\big)\wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{E}(y,z) \wedge \mathsf{E}(z,x) \wedge \mathsf{M}(x) \wedge \mathsf{M}(y) \wedge \mathsf{M}(z) \wedge \mathsf{B}(x,y) \wedge \mathsf{B}(y,z) \wedge \mathsf{B}(z,x)\big)\wedge \\ \neg\big(\mathsf{E}(x,x) \wedge \mathsf{M}(x) \wedge \mathsf{W}(x,x)\big)\wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{E}(y,x) \wedge \mathsf{E}(x,x) \wedge \mathsf{M}(x) \wedge \mathsf{M}(y) \wedge \mathsf{W}(x,y) \wedge \mathsf{W}(y,x) \wedge \mathsf{W}(x,x)\big)\wedge \\ \neg\big(\mathsf{E}(x,y) \wedge \mathsf{E}(y,z) \wedge \mathsf{E}(z,x) \wedge \mathsf{M}(x) \wedge \mathsf{M}(y) \wedge \mathsf{M}(z) \wedge \mathsf{W}(x,y) \wedge \mathsf{W}(y,z) \wedge \mathsf{W}(z,x)\big) \end{pmatrix}$$

$\triangle$

4.3 MMSNP$_2$ and infinite MMSNP

In Proposition 4.1.1 of Section 4.1, we have shown that for any MMSNP$_2$ sentence Φ there exists an MMSNP sentence $\tilde{\Phi}$ such that SAT(Φ) $\leq_p$ SAT($\tilde{\Phi}$). We now argue that it is not likely to have the other reduction direction between SAT(Φ) and SAT($\tilde{\Phi}$), even if we consider instead of $\tilde{\Phi}$ another $\tilde{\tau}$-sentence that is logically equivalent to $\tilde{\Phi}$ on the input consisting of good $\tilde{\tau}$-structures. Failing to find an equivalent MMSNP problem, we show that SAT(Φ) is P-time equivalent to a problem that looks like an infinite MMSNP. We turn to the language of forbidden patterns problems that is equivalent to MMSNP when there are finitely many forbidden patterns. Although, in our case, we need infinitely many

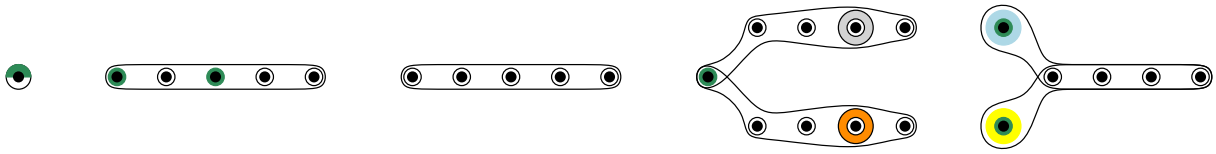


Figure 4.5: The five types of negated conjuncts that are added to Φ. The green disc is the T-relation. The white circle is the V-relation. The closed curves are the $\tilde{\mathsf{R}}$-tuples. The grey and orange discs with black boundaries are any two different $\tilde{\sigma}_{\mathsf{V}}$-relations. The blue and yellow discs are any two different $\tilde{\sigma}_{\mathsf{T}}$-relations.

forbidden patterns, it happens that the family consisting of these patterns is regular, which roughly means "finitely describable". This regularity property will be specified later, in Section 4.4.

By Proposition 4.2.2 from Section 4.2, we can assume that any connected MMSNP_2 sentence that we consider is also in normal$_1$ form. This means that the existential relations always make a partition of both R-tuples and vertices of the input structure.

Recall that $\tau = \{\mathsf{R}\}$ and $\tilde{\tau} = \{\mathsf{T}, \mathsf{V}, \tilde{\mathsf{R}}\}$; and that the arity of the τ-relation symbol R is k, T and V are unary, and the arity of the corresponding $\tilde{\tau}$-relation symbol $\tilde{\mathsf{R}}$ is $k+1$.

Now we modify the MMSNP sentence $\tilde{\Phi}$ so that the resulting sentence rejects more input structures that are not good. To do this, we add to $\tilde{\Phi}$ a family of negated conjuncts that help us reject some non good $\tilde{\tau}$-structures. There are five types of negated conjuncts, all of them are displayed on Figure 4.5 on page 83. Denote the new formula by $\tilde{\Phi}'$. These conjuncts are not fully coloured, therefore we consider every possible way to assign $\tilde{\sigma}$-relations to uncoloured elements.

- The first type contains just one negated conjunct: $\neg\big(\mathsf{T}(x) \wedge \mathsf{V}(x)\big)$. It forbids an element to be both a T-vertex and a V-vertex at the same time.

- The second type forbids a T-vertex to be in an $\tilde{\mathsf{R}}$-tuple on any coordinate except for the first one. For every i in $[k]$, we add the following: $\neg\big(\mathsf{T}(x_i) \wedge \tilde{\mathsf{R}}(t, \mathbf{x})\big)$, where x_i is the ith element of the k-tuple $\mathbf{x}$.

- The third type forbids a V-vertex to be on the first coordinate of some $\tilde{\mathsf{R}}$-tuple. We just add the following negated conjunct: $\neg\big(\mathsf{V}(t) \wedge \tilde{\mathsf{R}}(t, \mathbf{x})\big)$.

- The fourth type is the most interesting. We would like to forbid one T-vertex to be adjacent to more than one $\tilde{\mathsf{R}}$-tuple. Sadly, it is not likely to be expressed unless we use inequalities. However, we are still able to require that, for any two $\tilde{\mathsf{R}}$-tuples $(t, \mathbf{x}), (t, \mathbf{y})$ having a common T-vertex and for any coordinate i in $[k]$, the corresponding elements x_i, y_i must be coloured in the same $\tilde{\sigma}$-relations.

- The fifth type requires that any two $\tilde{\mathsf{R}}$-tuples that have the same V-vertices must have their T-vertices coloured in the same $\tilde{\sigma}$-relations.

Proposition 4.3.1. *SAT*(Φ) *is reducible in P-time to SAT*$(\tilde{\Phi}')$.

Proof. For any τ-structure $\mathfrak{A}$, its image $\tilde{\mathfrak{A}} := \tilde{\mathsf{D}}(\mathfrak{A})$ satisfies all the negated conjuncts from Figure 4.5 on page 83. Thus, by Proposition 4.1.1, we have $\tilde{\mathfrak{A}} \models \tilde{\Phi}$ if and only if $\tilde{\mathfrak{A}} \models \tilde{\Phi}'$. □

Observe that the sentence $\tilde{\Phi}'$ adds constraints only for those elements that are contained either in T-relation or in V-relation. This means that, for any $\tilde{\tau}$-structure $\mathfrak{A}$, we have $\mathfrak{A} \models \tilde{\Phi}'$ if and only if $\mathfrak{A}[A_{\mathsf{TV}}] \models \tilde{\Phi}'$, where $A_{\mathsf{TV}} = \{a \in A \mid a \in \mathsf{T}^{\mathfrak{A}} \cup \mathsf{V}^{\mathfrak{A}}\}$.

We introduce two useful notions that describe the situation when a $\tilde{\tau}$-structure is not good.

Definition 6. Let us call a $\tilde{\tau}$-structure $\mathfrak{A}$ *ugly* if it does not satisfy some negated conjuncts of the first three types from Figure 4.5.

In particular, it is FO-definable to check if a structure is ugly, as the negated conjuncts of the first three types do not have $\tilde{\sigma}$-atoms.

Definition 7. If, for a $\tilde{\tau}$-structure $\mathfrak{A}$, there are two $\tilde{\mathsf{R}}$-tuples $(t, \mathbf{x})$ and $(t, \mathbf{y})$ such that $\mathbf{x}$ is distinct from $\mathbf{y}$, then we say that that $\mathfrak{A}$ has *duplicated tuples.*

The current goal is to argue that $\tilde{\tau}$-structures with duplicated tuples is the only obstacle for having $\mathrm{SAT}(\Phi) \equiv_p \mathrm{SAT}(\tilde{\Phi})$.

Proposition 4.3.2. *Let $\mathfrak{A}$ be a $\tilde{\tau}$-structure. Then one of the following statements holds.*

- *$\mathfrak{A}$ has duplicated tuples.*
- *$\mathfrak{A}$ is an ugly structure and is rejected by $\tilde{\Phi}'$.*
- *One can construct in P-time in $|A|$ a τ-structure $\mathfrak{B}$ such that $\mathfrak{B} \models \Phi$ if and only if $\mathfrak{A} \models \tilde{\Phi}'$.*

Proof. Suppose that $\mathfrak{A}$ is neither ugly nor has duplicated tuples. Recall that without loss of generality we can assume that, for any a in A, a belongs to $\mathsf{T}^{\mathfrak{A}} \uplus \mathsf{V}^{\mathfrak{A}}$. This union is disjoint as $\mathfrak{A}$ is not ugly.

Suppose that there are two tuples $(t, \mathbf{x}), (t', \mathbf{x})$ in $\tilde{\mathsf{R}}^{\mathfrak{A}}$ with distinct T-vertices t and t'. If we prove that $\mathfrak{A} \models \tilde{\Phi}'$ if and only if $\mathfrak{A}[A \setminus \{t'\}] \models \tilde{\Phi}'$, then we are done, because we can delete all such T-vertices one-by-one, and, at the end, we get a good structure. And for any good structure there exists an equivalent τ-structure, by Proposition 4.1.2.

By Lemma 2.2.1, we know that any SNP sentence is closed under taking induced substructures. Thus, $\mathfrak{A} \models \tilde{\Phi}'$ implies $\mathfrak{A}[A \setminus \{t'\}] \models \tilde{\Phi}'$. For the other direction, suppose that $\mathfrak{A}[A \setminus \{t'\}] \models \tilde{\Phi}'$. Let $\mathfrak{A}[A \setminus \{t'\}]^{\tilde{\sigma}}$ be a valid $\tilde{\sigma}$-expansion *i.e.*, it satisfies every negated conjunct of $\tilde{\Phi}'$. Pick a $\tilde{\sigma}$-expansion $\mathfrak{A}^{\tilde{\sigma}}$ such that $\mathfrak{A}^{\tilde{\sigma}}[A \setminus \{t'\}] = \mathfrak{A}[A \setminus \{t'\}]^{\tilde{\sigma}}$ and that the T-vertex t' is in $\tilde{\mathsf{X}}^{\mathfrak{A}^{\tilde{\sigma}}}$ if and only if t is in $\tilde{\mathsf{X}}^{\mathfrak{A}^{\tilde{\sigma}}}$, for any $\tilde{\mathsf{X}}$ in $\tilde{\sigma}_{\mathsf{T}}$. That is, the $\tilde{\sigma}_{\mathsf{T}}$ colours for t and t' have to be the same. Otherwise, it would not satisfy some negated conjunct of the fifth type from Figure 4.5 on page 83. Suppose that, for some negated conjunct $\neg\tilde{\phi}_i(\mathbf{x}_i)$ of $\tilde{\Phi}'$ and for some tuple $\mathbf{a}_i$, we have $\mathfrak{A}^{\tilde{\sigma}} \models \tilde{\phi}_i(\mathbf{a}_i)$. Let $\mathbf{a}'_i$ be obtained from $\mathbf{a}_i$ by replacing each occurence of t' with t. As the T-vertices t and t' agree on the $\tilde{\sigma}_{\mathsf{T}}$ relations, as they do not belong to any $\tilde{\sigma}_{\mathsf{V}}$-relation, by the construction of $\tilde{\Phi}$, and as $\mathfrak{A}$ contains no duplicated tuples, we can conclude that $\mathfrak{A}^{\tilde{\sigma}} \models \tilde{\phi}_i(\mathbf{a}'_i)$. But then we have $\mathfrak{A}^{\tilde{\sigma}}[A \setminus \{t'\}] \models \tilde{\phi}_i(\mathbf{a}'_i)$ which contradicts our assumption that $\mathfrak{A}[A \setminus \{t'\}] \models \tilde{\Phi}'$. □

It is not clear which τ-structure should be associated with a $\tilde{\tau}$-structure with duplicated tuples. Such $\tilde{\mathsf{R}}$-tuples represent two distinct R-tuples. But, as they share the T-vertex, the corresponding R-tuples must have the same σ_{T}-colours. And it is not clear how to construct a τ-structure that can express such constraints.

Although we do not know how to show a P-time equivalence between MMSNP_2 and MMSNP, we still can generalise MMSNP sentences by using the language of forbidden patterns problems (FPPs). This approach is introduced by Madelaine in [MS07] in order to describe the MMSNP problems that are not CSP problems. It is shown to be P-time equivalent to MMSNP. This notion has later been extended for the case of the MMSNP_2 logic in [Mad09]. Though, these classes have always been defined to have finitely many forbidden structures. We need more than that. In the rest of this subsection we define the class of problems FPP_1^∞ and show that any MMSNP_2 problem $\mathrm{SAT}(\Phi)$ over a finite signature τ is P-time equivalent to some FPP_1^∞ problem over the corresponding signature $\tilde{\tau}$.

At first, we provide the definitions of *forbidden patterns problems* FPP_1 and FPP_2 as they are defined in [MS07, Mad09]. Let $\mathcal{V}$ be a finite set of vertex colours, $\mathcal{F}_1$ be a finite family of pairs $(\mathfrak{F}^\tau, \nu_{\mathfrak{F}^\tau})$, where $\mathfrak{F}^\tau$ is a τ-structure and $\nu_{\mathfrak{F}^\tau}\colon F \to \mathcal{V}$ is a mapping that assigns a colour of $\mathcal{V}$ to each element of $\mathfrak{F}^\tau$. $\mathrm{FPP}_1(\mathcal{F})$ is a decision problem defined on τ-structures such that a τ-structure $\mathfrak{A}$ belongs to the class $\mathrm{FPP}_1(\mathcal{F})$ if and only if there is a mapping $\nu_{\mathfrak{A}}\colon A \to \mathcal{V}$ such that, for any $(\mathfrak{F}^\tau, \nu_{\mathfrak{F}^\tau})$ in $\mathcal{F}$, if there is a homomorphism $\mathfrak{h}\colon \mathfrak{F}^\tau \to \mathfrak{A}$, then $\nu_{\mathfrak{A}} \circ \mathfrak{h} \neq \nu_{\mathfrak{F}^\tau}$.

Example 4.3.1. Consider the problem No-Monochromatic-Triangle: this is a decision problem defined on directed graphs, it accepts the input graph $\mathfrak{G}$ if one can colour its vertices with 2 colours B, W such that we cannot map to $\mathfrak{G}$ the directed 3-cycle with its vertices having the same colour. The family $\mathcal{F}_1$ consists of 6 structures that are displayed on Figure 4.6. It is the two monochromatic 3-cycles and all the possible homomorphic images of them.

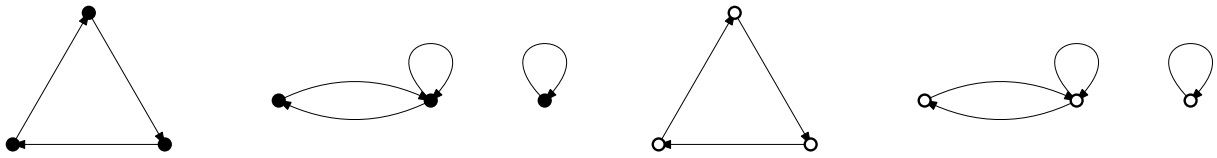


Figure 4.6: The forbidden structures of the No-Monochromatic-Triangle problem.

△

Let $\tau = \{\mathsf{R}\}$. Let $\mathcal{V}$ be a finite set of vertex colours, $\mathcal{E}$ be a finite set of edge (tuple) colours, and $\mathcal{F}$ be a finite set of triples $(\mathfrak{F}^\tau, \nu_{\mathfrak{F}^\tau}, \mathbf{e}_{\mathfrak{F}^\tau})$, where $\mathfrak{F}^\tau$ is a τ-structure, $\nu_{\mathfrak{F}^\tau}\colon F \to \mathcal{V}$ is a mapping that colours the elements of $\mathfrak{F}^\tau$, and $\mathbf{e}_{\mathfrak{F}^\tau}\colon \mathsf{R}^{\mathfrak{F}^\tau} \to \mathcal{E}$ is a mapping that colours the R-tuples of $\mathfrak{F}^\tau$. For a τ-structure $\mathfrak{A}$, we say that $\mathfrak{A}$ belongs to the class $\mathrm{FPP}_2(\mathcal{F})$ if and only if there are mappings $\nu_{\mathfrak{A}}\colon A \to \mathcal{V}$ and $\mathbf{e}_{\mathfrak{A}}\colon \mathsf{R}^{\mathfrak{A}} \to \mathcal{E}$ such that, for any $(\mathfrak{F}^\tau, \nu_{\mathfrak{F}^\tau}, \mathbf{e}_{\mathfrak{F}^\tau})$ in $\mathcal{F}$, there is no homomorphism $\mathfrak{h}\colon \mathfrak{F}^\tau \to \mathfrak{A}$ such that $\nu_{\mathfrak{A}} \circ \mathfrak{h} = \nu_{\mathfrak{F}^\tau}$ and $\mathbf{e}_{\mathfrak{A}}\big(\mathfrak{h}(\mathbf{f})\big) = \mathbf{e}_{\mathfrak{F}^\tau}(\mathbf{f})$, for any tuple $\mathbf{f}$ in $\mathsf{R}^{\mathfrak{F}^\tau}$.

Example 4.3.2. The No-Monochromatic-Arc-Triangle is an MMSNP_2 problem. It was considered before, in Example 3.2.2 and Example 4.2.1. It is also an FPP_2 problem, defined by the family of forbidden structures displayed on Figure 4.7 on page 86. △

Every FPP_2 problem is associated with an MMSNP_2 sentence in normal_1 form.

Proposition 4.3.3. *Let Φ be an MMSNP_2 sentence in normal_1 form. Then there exists a family of triples $\mathcal{F}_2$ such that $\mathrm{SAT}(\Phi)$ and $\mathrm{FPP}_2(\mathcal{F}_2)$ are the same problem.*

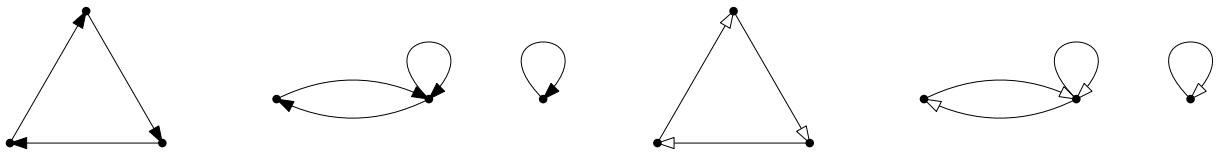


Figure 4.7: The forbidden structures of the No-Monochromatic-Arc-Triangle problem. The heads of arcs highlight the arc colours: B (black) and W (white).

Proof. As Φ is in normal$_1$ form, this sentence contains negated conjuncts that force the choice of the existential relations to be partitions of both vertices and R-tuples. That is, the negated conjuncts of the following forms:

$$\neg\left(\bigwedge_{\mathsf{M}\in\sigma_{\mathsf{V}}}\neg\mathsf{M}(x)\right)\wedge\bigwedge_{\mathsf{M},\mathsf{M}'\in\sigma_{\mathsf{V}}}\neg\big(\mathsf{M}(x)\wedge\mathsf{M}'(x)\big),$$

$$\neg\left(\mathsf{R}(\mathbf{x})\wedge\bigwedge_{\mathsf{X}\in\sigma_{\mathsf{T}}}\neg\mathsf{X}(\mathbf{x})\right)\wedge\bigwedge_{\mathsf{X},\mathsf{X}'\in\sigma_{\mathsf{T}}}\neg\big(\mathsf{R}(\mathbf{x})\wedge\mathsf{X}(\mathbf{x})\wedge\mathsf{X}'(\mathbf{x})\big).$$

For any other negated conjunct of Φ, we know that its canonical database is a connected $(\tau\uplus\sigma)$-structure $\mathfrak{F}$ such that any of its vertices and any of its R-tuples is coloured in exactly one existential colour of σ, where $\sigma=\sigma_{\mathsf{T}}\uplus\sigma_{\mathsf{V}}$. We say that the set of vertex colours $\mathcal{V}$ is equal to the set of monadic existential relation symbols σ_{V}, and that the set of edge colours $\mathcal{E}$ is equal to the set of k-ary existential relation symbols σ_{T}. For a negated conjunct $\neg\phi_i$ of Φ, we construct the triple $(\mathfrak{F}_i^{\tau},\nu_{\mathfrak{F}_i^{\tau}},\mathsf{e}_{\mathfrak{F}_i^{\tau}})$ as follows. The τ-structure $\mathfrak{F}_i^{\tau}$ is the τ-reduct of the canonical database $\mathfrak{F}_i$ of ϕ_i. For any variable x of ϕ_i there is exactly one σ-atom $\mathsf{M}(x)$ in ϕ_i, so, in this case, we set $\nu_{\mathfrak{F}_i^{\tau}}(x):=\mathsf{M}$. For any τ-atom $\mathsf{R}(\mathbf{x})$ of ϕ_i there is exactly one σ-atom $\mathsf{X}(\mathbf{x})$ in ϕ_i, so we set $\mathsf{e}_{\mathfrak{F}_i^{\tau}}(\mathbf{x})=\mathsf{X}$. The mappings $\nu_{\mathfrak{F}_i^{\tau}}$ and $\mathsf{e}_{\mathfrak{F}_i^{\tau}}$ are well-defined for every negated conjunct $\neg\phi_i$ of Φ. The class $\mathcal{F}_2$ is constructed.

Take any τ-structure $\mathfrak{A}$. Suppose that $\mathfrak{A}\models\Phi$; then there exists a σ-expansion $\mathfrak{A}^{\sigma}$ such that any element and any R-tuple of $\mathfrak{A}$ are coloured in exactly one colour of σ and that any negated conjunct of Φ is satisfied. This σ-expansion induces the mappings $\nu_{\mathfrak{A}}\colon A\to\mathcal{V}$ and $\mathsf{e}_{\mathfrak{A}}\colon\mathsf{R}^{\mathfrak{A}}\to\mathcal{E}$. By the construction of $\mathcal{F}_2$, we must have $\mathfrak{A}$ be accepted by $\mathrm{FPP}_2(\mathcal{F}_2)$. Suppose that $\mathfrak{A}$ satisfies $\mathrm{FPP}_2(\mathcal{F}_2)$. Then there exist corresponding mappings $\nu_{\mathfrak{A}}$ and $\mathsf{e}_{\mathfrak{A}}$. These mappings induce a partition of elements of $\mathfrak{A}$ and a partition of R-tuples of $\mathfrak{A}$, thus, they induce a σ-expansion of $\mathfrak{A}$. If this expansion does not satisfy some negated conjunct $\neg\phi_i$, then, by the construction of $\mathcal{F}_2$, there is a homomorphism $\mathsf{h}\colon\mathfrak{F}_i^{\tau}\to\mathfrak{A}$ that violates the property. □

We depart marginally from the way the classes FPP_1 and FPP_2 are defined: this is purely notational. We get rid of the mappings and propose the concept of a colouring expansion. One should think of colourings as of expansions, where we can define σ_{T}-relations only on input tuples.

Definition 8. A σ-expansion $\mathfrak{A}^{\sigma}$ of some τ-structure $\mathfrak{A}$ is called a *σ-colouring* if

- for any element a of A there is a unique unary M in σ_{V} such that $\mathsf{M}^{\mathfrak{A}^{\sigma}}(a)$, and
- for any tuple $\mathbf{a}$ in $\mathsf{R}^{\mathfrak{A}}$ there is a unique X in σ_{T} such that $\mathbf{a}$ belongs to $\mathsf{X}^{\mathfrak{A}^{\sigma}}$, and
- for any relation X in σ_{T}, it is contained in the set of R-tuples of $\mathfrak{A}$: $\mathsf{X}^{\mathfrak{A}^{\sigma}}\subseteq\mathsf{R}^{\mathfrak{A}}$.

Similarly, a $\tilde{\sigma}$-expansion $\mathfrak{B}^{\tilde{\sigma}}$ of some $\tilde{\tau}$-structure $\mathfrak{B}$ is a *$\tilde{\sigma}$-colouring* if

- for any V-vertex v of B there is a unique $\tilde{\sigma}_{\mathsf{V}}$-relation $\tilde{\mathsf{M}}$ such that $\tilde{\mathsf{M}}^{\mathfrak{B}^{\tilde{\sigma}}}(v)$, and
- for any T-vertex t of B there is a unique $\tilde{\sigma}_{\mathsf{T}}$-relation $\tilde{\mathsf{X}}$ such that $\tilde{\mathsf{X}}^{\mathfrak{B}^{\tilde{\sigma}}}(t)$, and
- any $\tilde{\sigma}_{\mathsf{V}}$-relation $\tilde{\mathsf{M}}$ is contained in the set of V-vertices of $\mathfrak{B}$: $\tilde{\mathsf{M}}^{\tilde{\sigma}} \subseteq \mathsf{V}^{\mathfrak{B}}$, and
- any $\tilde{\sigma}_{\mathsf{T}}$-relation $\tilde{\mathsf{X}}$ is contained in the set of T-vertices of $\mathfrak{B}$: $\tilde{\mathsf{X}}^{\tilde{\sigma}} \subseteq \mathsf{T}^{\mathfrak{B}}$.

A $(\tau \uplus \sigma)$-structure $\mathfrak{F}$ is called *coloured* if it is a σ-colouring of its τ-reduct $\mathfrak{F}^{\tau}$. The definition of a *coloured* $(\tilde{\tau} \uplus \tilde{\sigma})$-structure is similar.

Usually, it is more convenient to consider a coloured $(\tau \uplus \sigma)$-structure $\mathfrak{F}$ instead of the corresponding triple $(\mathfrak{F}^{\tau}, \mathsf{v}_{\mathfrak{F}^{\tau}}, \mathsf{e}_{\mathfrak{F}^{\tau}})$. Moreover, for a coloured structure, it is natural to say that its vertices and tuples are *coloured*. The following Proposition 4.3.4 immediately follows from the definition.

Proposition 4.3.4. *A τ-structure $\mathfrak{A}$ is accepted by $FPP_2(\mathcal{F}_2)$ if and only if there is a σ-colouring $\mathfrak{A}^{\sigma}$ of $\mathfrak{A}$ such that for any $\mathfrak{F}$, we have $\mathfrak{F} \not\to \mathfrak{A}^{\sigma}$, where $\mathfrak{F}$ is a coloured $(\tau \uplus \sigma)$-structure corresponding to some triple $(\mathfrak{F}^{\tau}, \mathsf{v}_{\mathfrak{F}^{\tau}}, \mathsf{e}_{\mathfrak{F}^{\tau}})$ in $\mathcal{F}_2$. Similarly, a $\tilde{\tau}$-structure $\mathfrak{B}$ is accepted by $FPP_1(\mathcal{F})$ if and only if there is a $\tilde{\sigma}$-colouring $\mathfrak{B}^{\tilde{\sigma}}$ such that for no $\mathfrak{G}$ corresponding to some $(\mathfrak{G}^{\tilde{\tau}}, \mathsf{v}_{\mathfrak{G}^{\tilde{\tau}}})$ in $\mathcal{F}$ there is a homomorphism $\mathfrak{G} \to \mathfrak{B}^{\tilde{\sigma}}$.*

Because of duplicated tuples, we fail to find an MMSNP sentence that produces a problem P-time equivalent to SAT(Φ). For any finite family of forbidden structures there exists a structure with duplicated tuples such that it is not known what it should be reduced to. However, later in this section, we manage to resolve this issue by extending the set of forbidden structures infinitary long. The resulting family provides a P-time equivalent problem to a given MMSNP_2 problem. As now we have infinitely many forbidden structures, the problem does not belong to FPP_1 anymore, so we need to use another notation: FPP_1^{∞}.

The definition of the class $\mathrm{FPP}_1^{\infty}(\mathcal{F})$ is similar to the finite case, but now the family $\mathfrak{F}$ can be also countably infinite.

Definition 9. Let τ be a finite relational signature and let σ be a finite relational signature consisting of unary symbols. Let $\mathcal{F}$ be a countable family of coloured $(\tau \uplus \sigma)$-structures. The *infinite forbidden patterns problem* for the family $\mathcal{F}$, denoted by $\mathrm{FPP}_1^{\infty}(\mathcal{F})$, is the following class of τ-structures. For a τ-structure $\mathfrak{A}$, $\mathfrak{A}$ is accepted by $\mathrm{FPP}_1^{\infty}(\mathcal{F})$ if there exists a σ-colouring $\mathfrak{A}^{\sigma}$ such that for any $\mathfrak{F}$ in $\mathcal{F}$ there is no homomorphism $\mathfrak{F} \to \mathfrak{A}^{\sigma}$.

Example 4.3.3. The 2-COLOURABILITY problem that accepts precisely all bipartite directed graphs can be represented as an FPP_1^{∞} problem. In this case, the set of vertex colours σ is empty. The family of forbidden structures $\mathcal{F}$ consists of all possible orientations of odd cycles. The members of $\mathcal{F}$ of length 3, 5, and 7 are displayed on Figure 4.8 on page 88. △

In the next construction and subsequent chapters, duplicated tuples are a niggling challenge, that can be somewhat alleviated by the following equivalence relation.

Definition 10. For any $\tilde{\tau}$-structure, let eq be the minimal by inclusion equivalence relation such that, for any two $\tilde{\mathsf{R}}$-tuples $(t, \mathbf{x}), (t, \mathbf{y})$ that have a common T-vertex t, the pair (x_i, y_i) belongs to $\mathsf{eq}^{\mathfrak{A}}$, for any i in $[k]$. See Figure 4.9 on page 88.

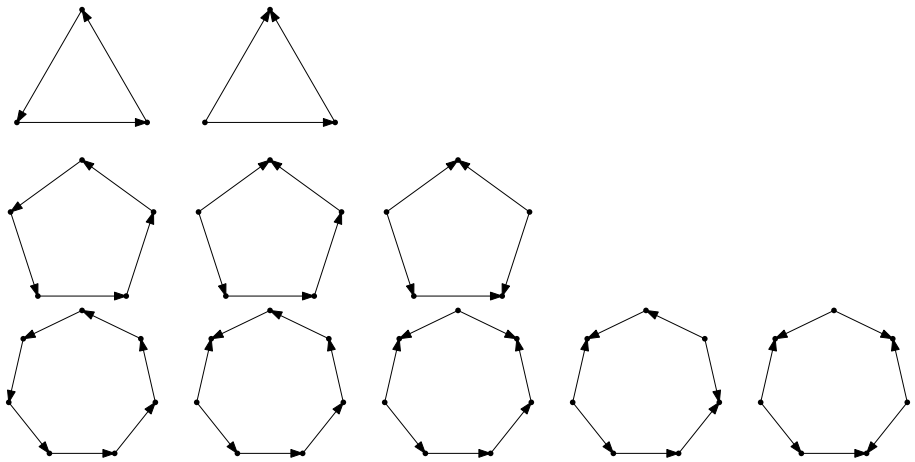


Figure 4.8: All oriented odd cycles of length 3, 5, and 7.

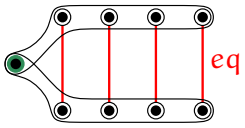


Figure 4.9: The $\mathtt{eq}$ relation (marked with red edges) and a pair of duplicated tuples.

Before we describe how $\mathcal{F}_1$ is constructed, we would like to explain informally why and how we shall proceed. We could use as obstruction set the family $\mathcal{O} = \{\mathfrak{G} \mid \tilde{\mathfrak{F}} \to \mathfrak{G}/\mathtt{eq}, \text{ for some } \mathfrak{F} \in \mathcal{F}_2\}$. Although it is easier to show that $\mathrm{FPP}_1^\infty(\mathcal{O}) \equiv_p \mathrm{FPP}_2(\mathcal{F}_2)$, this family would not have the regularity property we wish for, which is made precise later, in Section 4.4. So we construct a subset $\mathcal{F}_1$ of $\mathcal{O}$ with some care.

We want that for any $\mathfrak{O}$ in $\mathcal{O}$ there is some $\mathfrak{G}$ in $\mathcal{F}_1$ such that $\mathfrak{G}$ has a homomorphism $\mathfrak{h}_1$ to $\mathfrak{O}$. By construction of $\mathfrak{O}$, there is some $\mathfrak{F}$ in $\mathcal{F}_2$ and a homomorphism $\mathfrak{h}_0 \colon \tilde{\mathfrak{F}} \to \mathfrak{O}/\mathtt{eq}$. Let $\mathtt{Old}$ be the set of elements of $\mathfrak{O}$ that are in an equivalence class of $\mathtt{eq}$ in the image of the homomorphism $\mathfrak{h}_0$. We informally refer to such elements as $\mathtt{Old}$ elements.

We want $\mathfrak{G}$ to be "simple" enough so that the class $\mathcal{F}_1$ is regular. Here, "simple" will mean that locally it shall be tree-like. Indeed, viewing the computation of $\mathtt{eq}$ as an inductive process, *i.e.*, two vertices are made $\mathtt{eq}$-equivalent if they appear in the same coordinate of a duplicated tuple, then we can reconstruct a minimal "explanation" as to why $\mathtt{Old}$ elements of $\mathfrak{O}$ are in the same $\mathtt{eq}$-equivalence class. This minimal explanation is by nature tree-like.

What amounts to one image of a tuple of $\tilde{\mathfrak{F}}$ under the homomorphism $\mathfrak{h}_0$ may be associated with several tuples in $\mathfrak{O}$. Because taking the quotient under $\mathtt{eq}$ "squashes" duplicated tuples, we may have in $\mathfrak{O}$ something that amounts to "stretched" tuples in the inverse of the surjective homomorphism from $\mathfrak{O}$ to $\mathfrak{O}/\mathtt{eq}$.

Because we need a simple obstruction $\mathfrak{G}$ to have a homomorphism $\mathfrak{h}_1$ to $\mathfrak{O}$, it means that at the very least it must also have such "stretched tuples". This is made precise in the construction of $\tilde{\mathfrak{F}}_\circ$ below.

This structure $\tilde{\mathfrak{F}}_\circ$ seems not general enough for our purpose as the distance between some $\mathtt{Old}$ elements of $\mathfrak{O}$ may be large unlike the elements of $\tilde{\mathfrak{F}}_\circ$, where there is a unique $\mathtt{Old}$ element within each equivalence class. So we have no hope to guarantee a homomorphism from $\tilde{\mathfrak{F}}_\circ$ to $\mathfrak{O}$. We can however describe a specific structure $\mathfrak{G}$ that is inductively constructed from $\tilde{\mathfrak{F}}_\circ$, where the preimage of $\mathtt{Old}$ elements is connected in a tree-like

fashion that encapsulates minimal explanation of the computation of $\mathtt{eq}$. Figure 4.10 on page 89 illustrates these ideas.

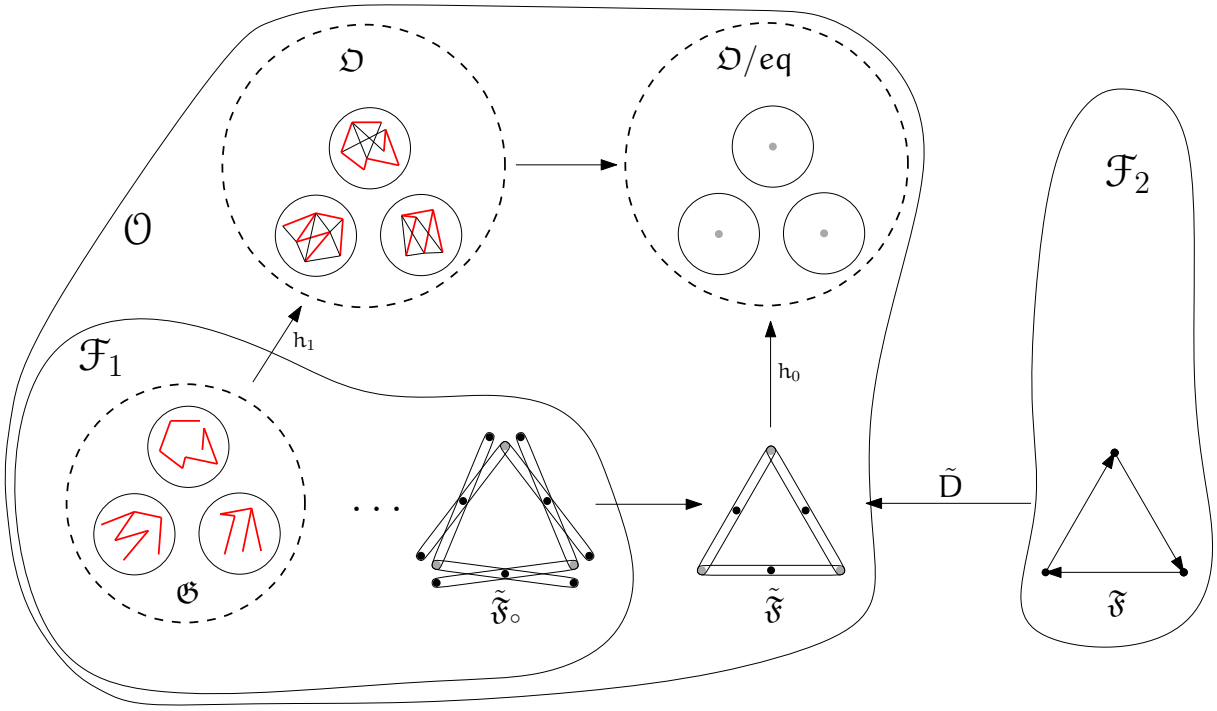
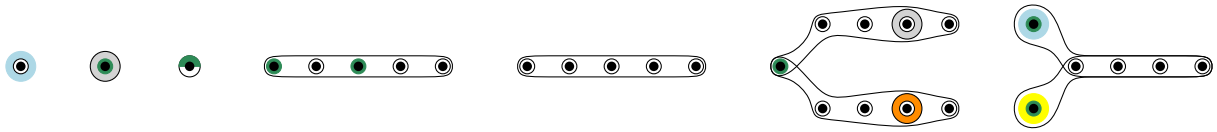


Figure 4.10: An illustration of the construction of $\mathcal{F}_1$. Grey dots are $\mathtt{Old}$ vertices. Small circles are $\mathtt{eq}$-equivalence classes. Red lines are tree edges.

Making the class $\mathcal{F}_1$. Now we are going to construct a countably infinite class $\mathcal{F}_1$ of coloured $(\tilde{\tau} \uplus \tilde{\sigma})$-structures and then show that $\mathrm{SAT}(\Phi) \equiv_p \mathrm{FPP}_1^\infty(\mathcal{F}_1)$. The family $\mathcal{F}_1$ is the disjoint union of the families:

$$\mathcal{F}_1 := \mathcal{F}_{ugly} \uplus \biguplus_{\mathfrak{F} \in \mathcal{F}_2} \mathcal{F}_{\mathfrak{F}}.$$

Here, $\mathcal{F}_{ugly}$ is a family of coloured $(\tilde{\tau} \uplus \tilde{\sigma})$-structures that are canonical databases of the negated conjuncts of the types from Figure 4.2 on page 77 and Figure 4.5 on page 83 reproduced below.



For those negated conjunct variables that are not coloured, we consider all possible $\tilde{\sigma}$-colourings of them.

Every family $\mathcal{F}_{\mathfrak{F}}$ is obtained from some coloured $(\tau \uplus \sigma)$-structure $\mathfrak{F}$ in $\mathcal{F}_2$, where $\mathcal{F}_2$ is constructed from Φ as in the proof of Proposition 4.3.3.

As we require that the V-vertices that are on the same coordinate of two duplicated tuples cannot be coloured with different $\tilde{\sigma}$-relations, we know that no two V-vertices of the same $\mathtt{eq}$-equivalence class can be coloured with different $\tilde{\sigma}$-relations. This relation is

useful because we plan to add to $\mathcal{F}_{\tilde{\mathfrak{F}}}$ all minimal by inclusion coloured $(\tilde{\tau} \uplus \tilde{\sigma})$-structures $\mathfrak{G}$ such that $\mathfrak{G}/\mathsf{eq}$ contains $\tilde{\mathfrak{F}}$ as an induced substructure.

We define a unary relation Old, for any structure that we add to $\mathcal{F}_{\tilde{\mathfrak{F}}}$. For a structure $\tilde{\mathfrak{F}}$, where $\mathfrak{F}$ is in $\mathcal{F}_2$, we set $\mathsf{Old}^{\tilde{\mathfrak{F}}} := \mathsf{V}^{\tilde{\mathfrak{F}}}$. As any structure of $\mathcal{F}_{\tilde{\mathfrak{F}}}$ is constructed from $\tilde{\mathfrak{F}}$ by adding new vertices, the relation Old helps us to remember which vertices are associated with the original ones in $\tilde{\mathfrak{F}}$.

Recall that k is the arity of the relation symbol R in τ. For a coloured $(\tau \uplus \sigma)$-structure $\mathfrak{F}$ in $\mathcal{F}_2$, denote by $\tilde{\mathfrak{F}}_\circ$ the following coloured $(\tilde{\tau} \uplus \tilde{\sigma})$-structure. This structure is obtained from $\tilde{\mathfrak{F}}$ by replacing any $\tilde{\mathsf{R}}$-tuple $(t, \mathbf{x})$ with k duplicated $\tilde{\mathsf{R}}$-tuples $(t, \mathbf{y}_1), \ldots, (t, \mathbf{y}_k)$, where, for i in $[k]$, the ith element $y_{i,i}$ of the tuple $\mathbf{y}_i$ is equal to the ith element x_i of $\mathbf{x}$, and all other elements are new and do not appear anywhere else. The relation $\mathsf{Old}^{\tilde{\mathfrak{F}}_\circ}$ consists of the V-vertices $y_{i,i}$ that are associated with the vertices x_i of $\tilde{F}$. See Figure 4.11 on page 90, for an illustration.

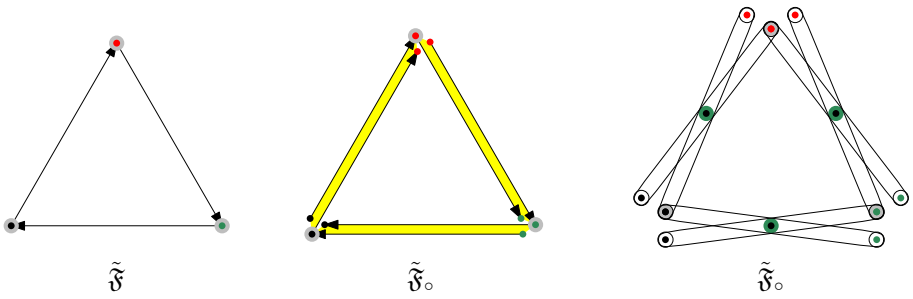


Figure 4.11: Suppose that $\mathfrak{F}$ is a triangle (*e.g.* $\mathfrak{C}_3$). Its $\tilde{\mathsf{D}}$-image $\tilde{\mathfrak{F}}$ is as on the left. There are two ways to represent the structure $\tilde{\mathfrak{F}}_\circ$: as in the middle and as on the right. In the middle, yellow zones between two arcs highlight pairs of duplicated $\tilde{\mathsf{R}}$-tuples, and, for convenience, T-vertices are not displayed in the middle. However, the figure on the right contains T-vertices, they are denoted by green disks. The relation Old is marked on both figures with grey disks. Red, green and black dots inside white circles represent $\tilde{\sigma}_\mathsf{V}$-coloured V-vertices.

The following construction describes the countably infinite family of coloured $(\tilde{\tau} \uplus \tilde{\sigma})$-structures associated with a given coloured $(\tau \uplus \sigma)$-structure $\mathfrak{F}$ in $\mathcal{F}_2$.

Construction 10. The class $\mathcal{F}_{\tilde{\mathfrak{F}}}$ is defined by induction. For the base case, we add $\tilde{\mathfrak{F}}_\circ$ to $\mathcal{F}_{\tilde{\mathfrak{F}}}$.

For the induction step, let a coloured $(\tilde{\tau} \uplus \tilde{\sigma})$-structure $\mathfrak{G}$ belong to $\mathcal{F}_{\tilde{\mathfrak{F}}}$. A new coloured structure $\mathfrak{G}'$ is constructed from $\mathfrak{G}$ as follows. Firstly, let $\mathfrak{G}'$ be a copy of $\mathfrak{G}$, with $\mathsf{Old}^{\mathfrak{G}'} = \mathsf{Old}^{\mathfrak{G}}$.

Let x be an Old vertex of $\mathfrak{G}'$. Suppose that there is more than one $\tilde{\mathsf{R}}$-tuple that contains x: $(t_1, \mathbf{v}_1), \ldots, (t_n, \mathbf{v}_n)$, where the ordering of these tuples is chosen arbitrarily. We split these n tuples in two non-empty sets $\mathcal{X}_1$, $\mathcal{X}_2$, without loss of generality, choose some m in $[n-1]$ and put $(t_i, \mathbf{v}_i)$ in $\mathcal{X}_1$ if $i \leq m$, otherwise put it to $\mathcal{X}_2$.

We introduce two new V-vertices x_1 and x_2 and require that they have the same unary $(\tilde{\tau} \uplus \tilde{\sigma})$-relations as x: they are both Old V-vertices that have the same $\tilde{\sigma}$-colour as x does. Then we introduce a new T-vertex t and $2(k-1)$ new V-vertices, and add to $\tilde{\mathsf{R}}^{\mathfrak{G}'}$ a pair of duplicated tuples $(t, \mathbf{w}_1), (t, \mathbf{w}_2)$, where $\mathbf{w}_1, \mathbf{w}_2$ consist of x_1, x_2 and of the recently added $2(k-1)$ V-vertices. Moreover, for some i in $[k]$, the ith coordinate of $\mathbf{w}_1$ is x_1 and the ith coordinate of $\mathbf{w}_2$ is x_2. We assign some $\tilde{\sigma}_\mathsf{T}$ colour to t and some $\tilde{\sigma}_\mathsf{V}$ colours to the $2(k-1)$ V-vertices such that any two vertices that are on the same coordinate in $\mathbf{w}_1$ and $\mathbf{w}_2$ must have the same $\tilde{\sigma}_\mathsf{V}$ colour.

Finally, delete x from $\mathfrak{G}'$ and replace all its occurences in tuples of $\mathfrak{X}_1$ with x_1 and replace all its occurences in tuples of $\mathfrak{X}_2$ with x_2. The construction of $\mathfrak{G}'$ from $\mathfrak{G}$ is finished. Any structure $\mathfrak{G}'$ obtained by this procedure from some $\mathfrak{G}$ in $\mathcal{F}_{\tilde{\mathfrak{F}}}$ is added to $\mathcal{F}_{\tilde{\mathfrak{F}}}$. This finishes Construction 10.

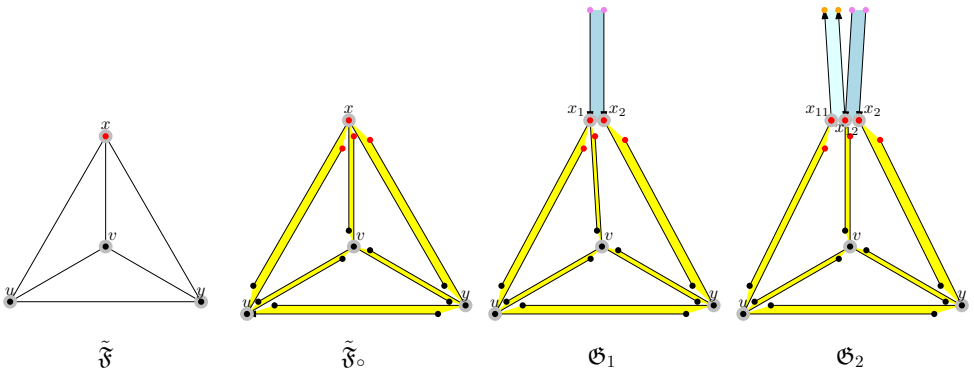


Figure 4.12: An example of new structure creation for $\mathcal{F}_{\tilde{\mathfrak{F}}}$. Yellow and blueish zones highlight pairs of duplicated $\tilde{\mathsf{R}}$-tuples. Coloured dots highlight the sets of V-vertices that have the same $\tilde{\sigma}$-colour.

Example 4.3.4. Consider Figure 4.12 on page 91. Suppose that R is a binary relation and that $\tilde{\mathfrak{F}}$ is as on Figure 4.12. At the first step, we copy the vertex x that is coloured in "red". The new vertices x_1 and x_2 are also $\mathtt{Old}$ and have the same colour as x. The vertex x_1 is not connected to y, and x_2 is not connected to u and v. Then we add a pair of duplicated $\tilde{\mathsf{R}}$-tuples that have x_1 and x_2 on the same coordinate. The pair of newly added V-vertices is coloured with the same $\tilde{\sigma}_{\mathsf{V}}$-colour. The resulting structure is called $\mathfrak{G}_1$. Similarly, $\mathfrak{G}_2$ is obtained from $\mathfrak{G}_1$ by replacing x_1 with x_{11} and x_{12}. △

Tree families. There is a nice way to represent structures of every family $\mathcal{F}_{\tilde{\mathfrak{F}}}$. Any $\mathfrak{G}$ in $\mathcal{F}_{\tilde{\mathfrak{F}}}$ can be encoded by a family of labeled unoriented trees $\mathcal{T}_{\mathfrak{G}} = \{\mathfrak{T}_f \mid f \in F\}$, where F is the domain of $\mathfrak{F} \in \mathcal{F}_2$. Each tree is associated with an element of $\mathfrak{F}$ or, equivalently, with a V-vertex of $\tilde{\mathfrak{F}}$.

Construction 11. The vertex set T_f of a tree $\mathfrak{T}_f$ in $\mathcal{T}_{\mathfrak{G}}$ consists of $\mathtt{Old}$ vertices that belong to the same eq-equivalence class $[f]_{\mathsf{eq}}$ of $\mathfrak{G}$, that is, $T_f := \mathtt{Old}^{\mathfrak{G}} \cap [f]_{\mathsf{eq}}$. For two vertices v, w in T_f there is an edge vw in $\mathfrak{T}_f$ if $\mathfrak{G}$ contains a pair of duplicated $\tilde{\mathsf{R}}$-tuples $(t, \mathbf{y})$ and $(t, \mathbf{z})$ that have v and w on the same coordinate, *i.e.*, for some i in $[k]$, $v = y_i$ and $w = z_i$, see Figure 4.13.

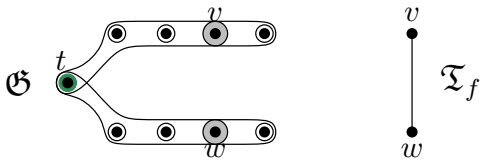


Figure 4.13: Duplicated tuples in $\mathfrak{G}$ are associated with edges of $\mathfrak{T}_f$.

For any tree $\mathfrak{T}_f$ of $\mathcal{T}_{\tilde{\mathfrak{F}}}$ there is a set of labels L_f that are assigned to vertices of $\mathfrak{T}_f$. Each L_f is associated with the set of R-tuples of $\mathfrak{F}$ that contain the element f. Let $l_{\mathbf{f}}$ in L_f be associated with some tuple $\mathbf{f} = (f_1, \ldots, f_k)$ in $\mathsf{R}^{\mathfrak{F}}$, where $f = f_i$, for some i in $[k]$. The label $l_{\mathbf{f}}$ is assigned to a vertex y of T_f if there is an $\tilde{\mathsf{R}}$-tuple $(t, y_1, \ldots, y_k)$ such that,

for any j in $[k]$, y_j belongs to the $\mathtt{eq}$-equivalence class $[f_j]_{\mathtt{eq}}$. By the construction of $\mathcal{F}_{\mathfrak{F}}$, for any label of L_f there is a unique vertex of T_f that has this label.

Any edge of any tree $\mathfrak{T}_f$ of $\mathcal{T}_{\mathfrak{G}}$ has a number i in $[k]$, a relation symbol from $\tilde{\sigma}_{\mathsf{T}}$, and a $(k-1)$-tuple of relation symbols of $\tilde{\sigma}_{\mathsf{V}}$ assigned to it. Suppose that v and w are two $\mathtt{Old}$ vertices that are contained on the same coordinate within a pair of duplicated tuples $\tilde{\mathsf{R}}(t,\mathbf{y})$ and $\tilde{\mathsf{R}}(t,\mathbf{z})$, that is, for some i in $[k]$, $v = y_i$ and $w = z_i$. Then the corresponding edge vw has a number i assigned to it. We require that any element of $\mathfrak{G}$ is coloured with exactly one monadic relation of $\tilde{\sigma}$, so we need to assign a $\tilde{\sigma}_{\mathsf{T}}$-relation to the T-vertex t and $\tilde{\sigma}_{\mathsf{V}}$-relations to the other $2(k-1)$ V-vertices of the tuples $\tilde{\mathsf{R}}(t,\mathbf{x})$ and $\tilde{\mathsf{R}}(t,\mathbf{y})$. As every two V-vertices on the same coordinate must have the same colour, we can express it by a $(k-1)$-tuple of relation symbols of $\tilde{\sigma}_{\mathsf{V}}$. See Figure 4.14.

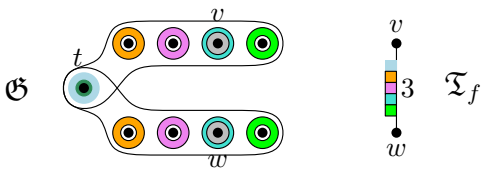


Figure 4.14: We assign to every edge of $\mathfrak{T}_f$ the number i of $[k]$ representing the position of v and w in $\mathbf{y}$ and $\mathbf{z}$, and we also assign the tuple of $\tilde{\sigma}$-relation symbols that colour $(t,\mathbf{y})$ and $(t,\mathbf{z})$.

Example 4.3.5. Look at Figure 4.15. It has three families of labeled trees that are associated with the structures from Figure 4.12 on page 91. Those elements that are connected by a pair of duplicated tuples are adjacent within the corresponding tree. Every tree of the first added structure $\tilde{\mathfrak{F}}_\circ$ consists of one vertex that has all the labels. After that, trees grow in size. In the tree family $\mathcal{T}_{\mathfrak{G}_1}$, the edge has label 2 because x_1 and x_2 are both on the second coordinate within the pair of duplicated tuples in $\mathfrak{G}_1$. Similarly, the edges of $\mathcal{T}_{\mathfrak{G}_2}$ have labels 1 and 2, which is associated with the structure $\mathfrak{G}_2$ of Figure 4.12. The coloured squares represent the monadic relations of $\tilde{\sigma}$ that are assigned to the T-vertex and to the pair of V-vertices, for each pair of duplicated tuples.

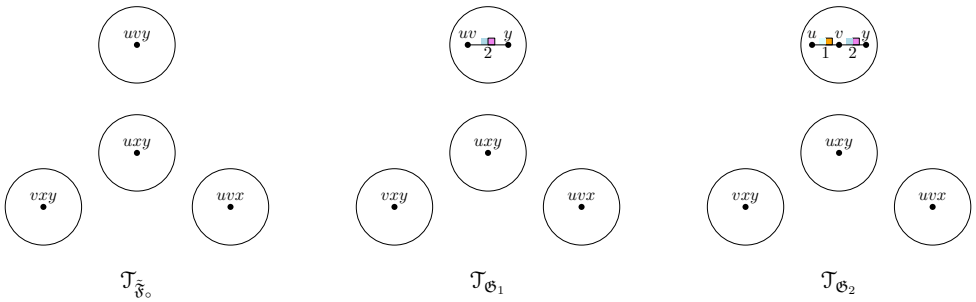


Figure 4.15: Tree families $\mathcal{T}_{\tilde{\mathfrak{F}}_\circ}$, $\mathcal{T}_{\mathfrak{G}_1}$ and $\mathcal{T}_{\mathfrak{G}_2}$. For simplicity, the labels are expressed with single letters u, v, x, y. In particular, a vertex in $[x]_{\mathtt{eq}}$ that has an edge to a vertex in $[y]_{\mathtt{eq}}$ has label y instead of xy.

△

Recall that a vertex of a tree is called a *leaf* if its degree is equal to 1.

Proposition 4.3.5. *Let $\mathfrak{G}$ be in $\mathcal{F}_{\mathfrak{F}}$, and $\mathcal{T}_{\mathfrak{G}}$ be the corresponding labeled tree family. Then, for any tree $\mathfrak{T}_f$ of $\mathcal{T}_{\mathfrak{G}}$ and for any leaf of $\mathfrak{T}_f$ there is a label in L_f assigned to this leaf.*

Proof. This is proved by induction. Consider the structure $\tilde{\mathfrak{F}}_\circ$. Any tree of $\mathcal{T}_{\tilde{\mathfrak{F}}_\circ}$ has only one vertex. Thus, there are no leaves.

Suppose that any leaf of any tree $\mathfrak{T}_f$ of $\mathcal{T}_{\mathfrak{G}}$ has a label in L_f assigned to it. Consider a structure $\mathfrak{G}'$ that is obtained from $\mathfrak{G}$ by Construction 10. All the trees of $\mathcal{T}_{\mathfrak{G}'}$ but one are the same as the ones of $\mathcal{T}_{\mathfrak{G}}$. Let $\mathfrak{T}_f$ in $\mathcal{T}_{\mathfrak{G}}$ and $\mathfrak{T}'_f$ in $\mathcal{T}_{\mathfrak{G}'}$ be the pair of different trees. The tree $\mathfrak{T}'_f$ is obtained from $\mathfrak{T}_f$ by taking some vertex x of $\mathfrak{T}_f$ and replacing it with a pair of adjacent vertices x_1 and x_2 such that either none of x_1, x_2 are leaves or some of them is a leaf but with a label from L_f. This is provided by the collections of tuples $\mathfrak{X}_1, \mathfrak{X}_2$ being non empty. Every tuple of each of these collections either provides an edge or provides a label from L_f to the corresponding vertex among x_1 and x_2. □

Any structure $\mathfrak{G}$ in $\mathcal{F}_{\mathfrak{F}}$ is associated with a family of labeled trees. We show now that this correspondence is one-to-one.

Proposition 4.3.6. *Let $\mathfrak{F}$ be in $\mathcal{F}_2$. For any f in F, let $\mathfrak{T}_f$ be a tree and L_f be a set of labels for vertices, and $[k] \times \tilde{\sigma}_{\mathsf{T}} \times (\tilde{\sigma}_{\mathsf{V}})^{k-1}$ be a set of labels for edges. Each element of L_f is associated with some R-tuple of $\mathfrak{F}$ that contains f. Suppose the following:*

- *for any l in L_f there is a unique vertex v in T_f such that l is assigned to v;*
- *any leaf of $\mathfrak{T}_f$ has some label of L_f assigned to it;*
- *any edge of any tree $\mathfrak{T}_f$ has a number i in $[k]$ and a k-tuple of $\tilde{\sigma}$-relation symbols assigned to it, where the first element of the tuple is a $\tilde{\sigma}_{\mathsf{T}}$-relation and the other $k-1$ elements are $\tilde{\sigma}_{\mathsf{V}}$-relations.*

Denote this family of labeled trees by $\mathcal{T}'$. Then there is a structure $\mathfrak{G}$ in $\mathcal{F}_{\mathfrak{F}}$ such that $\mathcal{T}_{\mathfrak{G}} = \mathcal{T}'$.

Proof. Recall that the structures of $\mathcal{F}_{\mathfrak{F}}$ are constructed inductively. Any new structure is obtained from some structure that has already been added to $\mathcal{F}_{\mathfrak{F}}$ by picking an $\mathtt{Old}$ vertex x and replacing it with two copies x_1 and x_2 that also belong to the relation $\mathtt{Old}$. Any $\tilde{\mathsf{R}}$-tuple that contained x now contains exactly one of x_1 and x_2.

Similarly, as any structure $\mathfrak{G}$ of $\mathcal{F}_{\mathfrak{F}}$ is obtained from $\tilde{\mathfrak{F}}_\circ$, any tree family $\mathcal{T}_{\mathfrak{G}}$ is obtained from a collection of one-vertex trees, where the vertex has all the labels. Each transformation, where x is replaced by x_1 and x_2, is associated with the following transformation of the labeled tree family. Suppose that we are replacing some $\mathtt{Old}$ vertex x that belongs to an equivalence class $[f]_{\mathsf{eq}}$ by two vertices x_1, x_2. We pick the corresponding vertex x in T_f. We replace x with two adjacent vertices x_1 and x_2 by assigning a number i and a k-tuple of $\tilde{\sigma}$-relations to the edge x_1x_2. The edge label i depends on the coordinate of the pair of duplicated $\tilde{\mathsf{R}}$-tuples where x_1 and x_2 are. And the k-tuple defines the way how the vertices of the duplicated tuples are coloured. Any vertex y that was adjacent to x is now adjacent to exactly one of $\{x_1, x_2\}$, we can choose it independently. The set of labels assigned to x is now split between x_1 and x_2, we can also choose it independently, as well as the labels for the edge between x_1 and x_2. Any other tree $\mathfrak{T}_{f'}$, for f' in $F \setminus \{f\}$, is kept the same during the modification of $\mathfrak{T}_f$. Clearly, any collection of labeled trees can be obtained by repeating this transformation sufficiently many times. □

We can prove two useful statements: Lemma 4.3.7 states that any homomorphism from $\mathfrak{A}$ to a good structure factors through $\mathfrak{A}/\mathsf{eq}$, Lemma 4.3.8 states that the family $\mathcal{F}_{\mathfrak{F}}$ is sufficient to reject any structure $\mathfrak{G}$ whose eq-quotient contains a homomorphic image of $\tilde{\mathfrak{F}}$. They help to prove then that $\mathrm{FPP}_1^\infty(\mathcal{F}_1) \equiv_p \mathrm{SAT}(\Phi_2)$.

Lemma 4.3.7. *Let $\mathfrak{A}$ be a finite $(\tilde{\tau} \uplus \tilde{\sigma})$-structure, $\mathfrak{B}$ be a good $(\tilde{\tau} \uplus \tilde{\sigma})$-structure (in particular, $\mathfrak{B} \cong \mathfrak{B}/\mathtt{eq}$). Then $\mathfrak{A} \to \mathfrak{B}$ if and only if $\mathfrak{A}/\mathtt{eq} \to \mathfrak{B}$.*

Proof. Denote by $[x]_{\mathtt{eq}}$ an element of $\mathfrak{A}/\mathtt{eq}$ that is associated with an equivalence class of $\mathfrak{A}$ containing x.

Suppose that $\mathfrak{A}/\mathtt{eq} \to \mathfrak{B}$. By the definition of quotient, we have $\mathfrak{A} \to \mathfrak{A}/\mathtt{eq}$. By transitivity of homomorphism, we have $\mathfrak{A} \to \mathfrak{B}$.

Suppose that $\mathfrak{A} \to \mathfrak{B}$. Let $\mathtt{h}\colon \mathfrak{A} \to \mathfrak{B}$ be a homomorphism. Consider two $\tilde{\mathtt{R}}$-tuples of $\mathfrak{A}$: $\tilde{\mathtt{R}}(t, \mathbf{a}_1), \tilde{\mathtt{R}}(t, \mathbf{a}')$. As $\mathfrak{B}$ is good, for the image $\mathtt{h}(t)$ of t there is only one $\tilde{\mathtt{R}}$-tuple of $\mathfrak{B}$: $\tilde{\mathtt{R}}(\mathtt{h}(t), \mathbf{b})$, that is adjacent to $\mathtt{h}(t)$. Then we must have $\mathtt{h}(\mathbf{a}) = \mathtt{h}(\mathbf{a}') = \mathbf{b}$. By transitivity, for any two a_1 and a_2 that belong to the same $\mathtt{eq}$-equivalence class of $\mathfrak{A}/\mathtt{eq}$, we have $\mathtt{h}(a_1) = \mathtt{h}(a_2)$. So we can construct a homomorphism $\mathtt{h}_{\mathtt{eq}}\colon \mathfrak{A}/\mathtt{eq} \to \mathfrak{B}$ such that $\mathtt{h}_{\mathtt{eq}}([a]_{\mathtt{eq}}) = \mathtt{h}(a)$. □

Lemma 4.3.8. *Let $\mathfrak{A}$ be a finite $(\tilde{\tau} \uplus \tilde{\sigma})$-structure such that there is a homomorphism $\mathtt{h}\colon \tilde{\mathfrak{F}} \to \mathfrak{A}/\mathtt{eq}$, for some $\mathfrak{F}$ in $\mathcal{F}_2$. Then, either there exists $\mathfrak{G}$ in $\mathcal{F}_{\mathfrak{F}}$ such that $\mathfrak{G} \to \mathfrak{A}$, or there exists a structure in $\mathcal{F}_{ugly}$ that is mapped to $\mathfrak{A}$.*

Proof. Denote by $\mathtt{h}(\tilde{\mathfrak{F}})$ the substructure of $\mathfrak{A}/\mathtt{eq}$ consisting of the elements and of the $\tilde{\mathtt{R}}$-tuples of the $\mathtt{h}$-image of $\tilde{F}$. Pick any V-vertex $[x]_{\mathtt{eq}}$ of $\mathfrak{A}/\mathtt{eq}$ that is contained in $\mathtt{h}(\tilde{\mathfrak{F}})$. Let $F_x := \{f_1, \dots, f_p\}$ be the preimage of $[x]_{\mathtt{eq}}$, that is, the set of V-vertices of $\tilde{\mathfrak{F}}$ such that, for any f in F_x, we have $\mathtt{h}(f) = [x]_{\mathtt{eq}}$. Denote by $T_x := [x]_{\mathtt{eq}}$ the corresponding set of vertices of $\mathfrak{A}$. Define a labeled undirected graph $(T_x; \mathsf{E})$ on the set T_x as follows: if $\mathfrak{A}$ contains a pair of duplicated $\tilde{\mathtt{R}}$-tuples that contain x_1 and x_2 on the ith coordinate and are coloured with $\tilde{\sigma}$-relations $\mathsf{X}, \mathsf{M}_1, \dots, \mathsf{M}_k$, then add to $(T_x; \mathsf{E})$ an edge x_1x_2 labeled with i in $[k]$ and $(\mathsf{X}, \mathsf{M}_1, \dots, \mathsf{M}_k)$. If x_1 and x_2 are coloured differently, then there is a structure in $\mathcal{F}_{ugly}$ that can be mapped to $\mathfrak{A}$, so we assume that they are always coloured by the same $\tilde{\sigma}_{\mathsf{V}}$-relation. Similarly, we assume that any two elements of this pair of tuples that are on the same coordinates have the same colour. Suppose that x is contained in an $\tilde{\mathtt{R}}$-tuple $(t, x_1, \dots, x_k)$ such that $(t, [x_1]_{\mathtt{eq}}, \dots, [x_k]_{\mathtt{eq}})$ is contained in $\mathtt{h}(\tilde{\mathfrak{F}})$; then we assign to x a corresponding label from $\bigcup_{f \in F_x} L_f$. For any label of this set there is at least one element in T_x having this label, because $\mathtt{h}$ is a mapping.

The structure $\mathfrak{G}$ of $\mathcal{F}_{\mathfrak{F}}$ that we want to obtain is defined by its family of trees $\mathcal{T}_{\mathfrak{G}}$, by Proposition 4.3.6. Any V-vertex f of $\tilde{\mathfrak{F}}$ belongs to some F_x. For any label l in L_f there is an element x_l in T_x that has this label. Choose $\mathfrak{T}_f$ to be isomorphic to a minimal by inclusion tree that is a subgraph of $(T_x; \mathsf{E})$ and that contains all these elements x_l. This tree exists because $(T_x; \mathsf{E})$ is a connected graph. And any leaf of this tree has a label from L_f, otherwise it is not minimal by inclusion. The edges of $\mathfrak{T}_f$ have the same labels as the corresponding edges of (T_x, E).

A homomorphism $\mathtt{h}'$ from $\mathfrak{G}$ to $\mathfrak{A}$ is constructed as follows. Any $\mathtt{Old}$ vertex v of $\mathfrak{G}$ is associated with a vertex of some tree $\mathfrak{T}_f$ of $\mathcal{T}_{\mathfrak{G}}$. We map the $\mathtt{Old}$ vertices according to the isomorphisms between the trees. Any other vertex of $\mathfrak{G}$ is adjacent to exactly one $\mathtt{Old}$ vertex of $\mathfrak{G}$ by exactly one $\tilde{\mathtt{R}}$-tuple $(t, \mathbf{x})$. So it suffices to explain how $\mathtt{h}'$ maps any $\tilde{\mathtt{R}}$-tuple. A $\tilde{\mathtt{R}}$-tuple may belong to a pair of duplicated tuples that are added to $\mathfrak{G}$ at some step of its construction. In this case, we know that $\mathfrak{A}$ contains a similar pair of duplicated tuples, this is provided by the coincidence of the corresponding edge labels of $\mathfrak{T}_f$ and $(T_x; \mathsf{E})$. Otherwise, an $\tilde{\mathtt{R}}$-tuple belongs to a collection of k duplicated tuples, this collection is associated with some $\tilde{\mathtt{R}}$-tuple of $\tilde{\mathfrak{F}}$, and this collection is added when we introduce the structure $\tilde{\mathfrak{F}}_\circ$. Any tuple of this collection is adjacent to exactly one $\mathtt{Old}$

vertex. This vertex is associated with some tree vertex $v_{\mathfrak{G}}$ that has a label indicating that it is adjacent to this tuple. The vertex $v_{\mathfrak{G}}$ is mapped by h' to $v_{\mathfrak{A}}$ that also has the same label. Thus, there is a tuple in $\mathfrak{A}$ where we can map it. □

Now we can show that any MMSNP_2 problem is P-time equivalent to some FPP_1^∞ problem.

Theorem 4.3.9. *For any Φ in MMSNP_2 there is a countably infinite family of finite structures $\mathcal{F}_1$ such that SAT(Φ) is P-time equivalent to FPP$_1(\mathcal{F}_1)$.*

Proof. Let $\mathcal{F}_2$ be the family of coloured $(\tau \uplus \sigma)$-structures constructed from Φ as in Proposition 4.3.3. Then, for $\mathcal{F}_2$, we construct a countably infinite family $\mathcal{F}_1$:

$$\mathcal{F}_1 := \mathcal{F}_{ugly} \uplus \biguplus_{\mathfrak{F} \in \mathcal{F}_2} \mathcal{F}_{\mathfrak{F}}.$$

Now we prove that $\mathrm{FPP}_2(\mathcal{F}_2) \leq_p \mathrm{FPP}_1^\infty(\mathcal{F}_1)$. Consider any τ-structure $\mathfrak{A}$ and the $\tilde{\tau}$-structure $\tilde{\mathfrak{A}} := \tilde{\mathsf{D}}(\mathfrak{A})$.

Suppose that there is a σ-colouring $\mathfrak{A}^\sigma$ such that for no $\mathfrak{F}$ in $\mathcal{F}_2$ there is a homomorphism $\mathsf{h}\colon \mathfrak{F} \to \mathfrak{A}^\sigma$. Then, choose the $\tilde{\sigma}$-colouring $\tilde{\mathfrak{A}}^{\tilde{\sigma}}$ of $\tilde{\mathfrak{A}}$ to be equal to $\tilde{\mathsf{D}}(\mathfrak{A}^\sigma)$.

As $\tilde{\mathfrak{A}}^{\tilde{\sigma}}$ is a good structure, there is no homomorphism $\mathfrak{u}\colon \mathfrak{U} \to \tilde{\mathfrak{A}}^{\tilde{\sigma}}$, for $\mathfrak{U}$ in $\mathcal{F}_{ugly}$. If there is a homomorphism $\mathsf{g}\colon \mathfrak{G} \to \tilde{\mathfrak{A}}^{\tilde{\sigma}}$, for some $\mathfrak{G}$ in $\mathcal{F}_{\mathfrak{F}} \subset \mathcal{F}_1$, then, by Lemma 4.3.7, we have $\mathfrak{G}/\mathsf{eq} \to \tilde{\mathfrak{A}}^{\tilde{\sigma}}$. By the construction of $\mathcal{F}_{\mathfrak{F}}$, $\tilde{\mathfrak{F}}$ embeds into $\mathfrak{G}/\mathsf{eq}$, for any $\mathfrak{G}$ in $\mathcal{F}_{\mathfrak{F}}$, then, by transitivity, there is also a homomorphism $\tilde{\mathsf{f}}\colon \tilde{\mathfrak{F}} \to \tilde{\mathfrak{A}}^{\tilde{\sigma}}$. By Proposition 4.1.2, there exists a homomorphism $\mathsf{f}\colon \mathfrak{F} \to \mathfrak{A}^\sigma$. This is a contradiction.

Suppose that, for any σ-colouring $\mathfrak{A}^\sigma$ there always exists some $\mathfrak{F}$ in $\mathcal{F}_2$ such that $\mathfrak{F} \to \mathfrak{A}^\sigma$. Take any $\tilde{\sigma}$-colouring $\tilde{\mathfrak{A}}^{\tilde{\sigma}}$. It is the $\tilde{\mathsf{D}}$-image of some σ-colouring $\mathfrak{A}^\sigma$ such that $\mathfrak{F} \to \mathfrak{A}^\sigma$, for some $\mathfrak{F}$ in $\mathcal{F}_2$. So, by Proposition 4.1.2, there is a homomorphism $\tilde{\mathsf{f}}\colon \tilde{\mathfrak{F}} \to \tilde{\mathfrak{A}}^{\tilde{\sigma}}$. Then, by transitivity, we have $\tilde{\mathfrak{F}}_\circ \to \tilde{\mathfrak{A}}^{\tilde{\sigma}}$. We have proved that $\mathrm{SAT}(\Phi) \leq_p \mathrm{FPP}_1^\infty(\mathcal{F}_1)$.

Now we prove that $\mathrm{FPP}_1^\infty(\mathcal{F}_1) \leq_p \mathrm{SAT}(\Phi)$. Let $\mathfrak{B}$ be some $\tilde{\tau}$-structure. We are going to show that $\mathfrak{B}$ is accepted by $\mathrm{FPP}_1^\infty(\mathcal{F}_1)$ if and only if $\mathfrak{A} \models \Phi$. The structure $\tilde{\mathfrak{A}}$ is obtained from $\mathfrak{B}/\mathsf{eq}$ by removing every T-vertex t such that both tuples $(t, \mathbf{x}), (t', \mathbf{x})$ belong to $\tilde{\mathsf{R}}^{\mathfrak{B}/\mathsf{eq}}$, where $t' \neq t$ and $\mathbf{x}$ is some k-tuple of V-vertices, similarly as in the proof of Proposition 4.3.2. We assume that no structure of $\mathcal{F}_{ugly}$ can be mapped to $\mathfrak{B}$, otherwise we know that $\mathfrak{B}$ is rejected by $\mathrm{FPP}_1^\infty(\mathcal{F}_1)$.

Let $\mathfrak{B}^{\tilde{\sigma}}$ be a $\tilde{\sigma}$-colouring of $\mathfrak{B}$ such that, for any $\mathfrak{U}$ in $\mathcal{F}_{ugly}$, $\mathfrak{U} \not\to \mathfrak{B}^{\tilde{\sigma}}$. Denote by $\tilde{\mathfrak{A}}^{\tilde{\sigma}}$ the substructure of $\mathfrak{B}^{\tilde{\sigma}}/\mathsf{eq}$ induced on $\tilde{A}$. It is a $\tilde{\sigma}$-colouring of $\tilde{\mathfrak{A}}$. Let $\mathfrak{A}^\sigma$ be a σ-colouring of $\mathfrak{A}$ such that $\tilde{\mathsf{D}}(\mathfrak{A}^\sigma) \cong \tilde{\mathfrak{A}}^{\tilde{\sigma}}$.

Suppose that there exists $\mathfrak{G}$ in $\mathcal{F}_{\mathfrak{F}}$ such that $\mathfrak{G} \to \mathfrak{B}^{\tilde{\sigma}}$, then, by Lemma 4.3.8, we have $\tilde{\mathfrak{F}} \to \tilde{\mathfrak{B}}^{\tilde{\sigma}}/\mathsf{eq}$. The structure $\tilde{\mathfrak{F}}$ contains no $\tilde{\mathsf{R}}$-tuples $(t, \mathbf{x}), (t', \mathbf{x})$ such that $t \neq t'$, and no ugly structure can be mapped to $\mathfrak{B}^{\tilde{\sigma}}$, so we also have $\tilde{\mathfrak{F}} \to \tilde{\mathfrak{A}}^{\tilde{\sigma}}$. This means, by Proposition 4.1.2, that $\mathfrak{F} \to \mathfrak{A}^\sigma$. This means that $\mathfrak{A}$ is rejected by $\mathrm{FPP}_2(\mathcal{F}_2)$.

Suppose that there exists $\mathfrak{F}$ in $\mathcal{F}_2$ such that $\mathfrak{F} \to \mathfrak{A}^\sigma$. Then, by Proposition 4.1.2, we have $\tilde{\mathfrak{F}} \to \tilde{\mathfrak{A}}^{\tilde{\sigma}}$, and, by transitivity, $\tilde{\mathfrak{F}} \to \mathfrak{B}^{\tilde{\sigma}}/\mathsf{eq}$. Then, by Lemma 4.3.8, there is $\mathfrak{G}$ in $\mathcal{F}_{\mathfrak{F}}$ such that $\mathfrak{G} \to \mathfrak{B}^{\tilde{\sigma}}$. So $\mathfrak{B}$ is rejected by $\mathrm{FPP}_1^\infty(\mathcal{F}_1)$. We have proved that $\mathrm{FPP}_1^\infty(\mathcal{F}_1) \leq_p \mathrm{SAT}(\Phi)$. □

4.4 ω-categorical templates for MMSNP$_2$ and infinite MMSNP

We show that the class $\mathcal{F}_1$ is regular. By the result of Hubička and Nešetřil, it implies the existence of an ω-categorical $(\tilde{\tau} \uplus \tilde{\sigma})$-structure $\mathfrak{C}_{synt}$ such that a $\tilde{\tau}$-structure $\mathfrak{A}$ is accepted by $\mathrm{FPP}_1^{\infty}(\mathcal{F}_1)$ if and only if $\mathfrak{A}$ homomorphically maps to the $\tilde{\tau}$-reduct $\mathfrak{C}_{synt}^{\tilde{\tau}}$ of $\mathfrak{C}_{synt}$.

We also show that for any MMSNP_2 τ-sentence Φ that is in normal$_1$ form there is an ω-categorical $(\tau \uplus \sigma)$-structure $\mathfrak{C}_\Phi$ such that, for any τ-structure $\mathfrak{A}$, $\mathfrak{A}$ satisfies Φ if and only if there is a homomorphism from $\mathfrak{A}$ to the τ-reduct $\mathfrak{C}_\Phi^{\tau}$ of $\mathfrak{C}_\Phi$.

Finally, we show, for the functorial image $\mathfrak{C}_{sem} := \tilde{\mathsf{D}}(\mathfrak{C}_\Phi)$, that the structures $\mathfrak{C}_{sem}$ and $\mathfrak{C}_{synt}$ are homomorphically equivalent.

Preliminaries

A pair $(\mathfrak{P}, \mathbf{r})$ is called a *rooted structure*, if $\mathfrak{P}$ is a relational structure with domain P and $\mathbf{r}$ is a linearly ordered proper subset of elements of $\mathfrak{P}$. The elements of $\mathbf{r}$ are called the *roots* of $\mathfrak{P}$.

Let $\mathfrak{F}$ be a relational structure. A rooted structure $(\mathfrak{P}, \mathbf{r})$ is called a *piece of* $\mathfrak{F}$, if

- $\mathfrak{P}$ is an induced connected substructure of $\mathfrak{F}$;
- $\mathbf{r} = (r_1, \ldots, r_p)$ is a minimal set of elements of P such that, for any relational tuple $\mathbf{x}$, if $\mathbf{x}$ contains an element x that belongs to the complement $F \setminus P$ and an element x' that belongs to P, then x' must be contained in $\mathbf{r}$.

In this case, we say that $\mathfrak{F}$ *contains the piece* $(\mathfrak{P}, \mathbf{r})$. If $\mathfrak{F}$ is a member of some family of structures $\mathcal{F}$, then we also call $(\mathfrak{P}, \mathbf{r})$ a *piece of* $\mathcal{F}$.

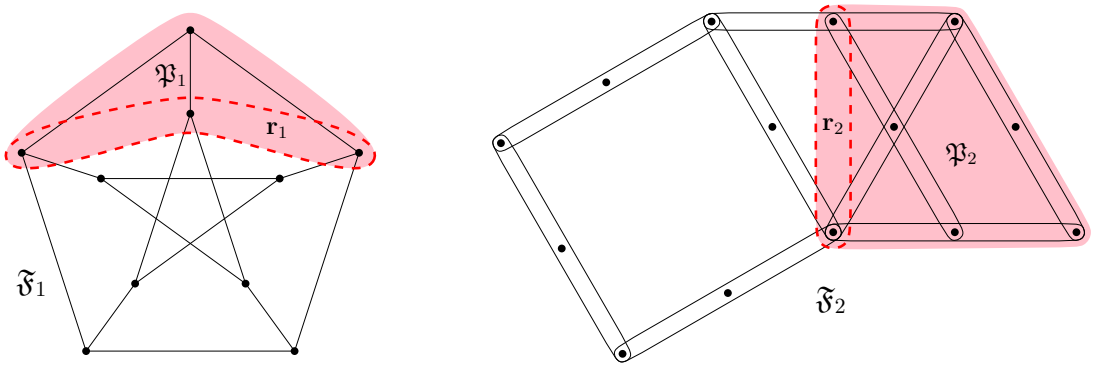


Figure 4.16: Examples of a piece and of not a piece.

Example 4.4.1. If the structure $\mathfrak{F}_1$ is a graph, then any induced connected subgraph $\mathfrak{P}_1$ of $\mathfrak{F}_1$ is a piece, where the set of roots $\mathbf{r}_1$ is a minimal separator, *i.e.*, a minimal by inclusion subset $\mathbf{r}_1$ of P_1 such that $\mathfrak{F}_1[F_1 \setminus \mathbf{r}_1]$ is not connected. However, if relational tuples of $\mathfrak{F}_2$ have arity greater than two, then there might be an induced connected substructure $\mathfrak{P}_2$ and a minimal separator $\mathbf{r}_2$ such that $(\mathfrak{P}_2, \mathbf{r}_2)$ is not a piece of $\mathfrak{F}_2$. These two cases are displayed on Figure 4.16. △

Let $(\mathfrak{A}, \mathbf{a})$ and $(\mathfrak{B}, \mathbf{b})$ be rooted structures. Suppose that the substructures induced on their sets of roots $\mathbf{a}$ and $\mathbf{b}$ are isomorphic: $\mathfrak{A}[\mathbf{a}] \cong \mathfrak{B}[\mathbf{b}]$. Denote by $(\mathfrak{A}, \mathbf{a}) \oplus (\mathfrak{B}, \mathbf{b})$ a structure that is obtained by taking two disjoint copies of $\mathfrak{A}$ and $\mathfrak{B}$ and by pairwisely identifying the roots of $\mathbf{a}$ and $\mathbf{b}$. See Figure 4.17.

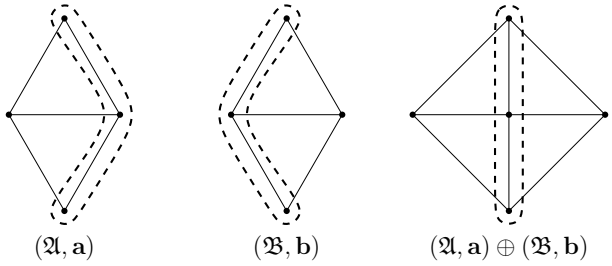


Figure 4.17: An example of the $\oplus$ operation.

Let $\mathcal{F}$ be a family of structures. Suppose that $(\mathfrak{P}, \mathbf{r})$ and $(\mathfrak{P}', \mathbf{r}')$ are pieces of structures of $\mathcal{F}$ such that the pairwise correspondence between their roots induces an isomorphism between the induced substructures $\mathfrak{P}[\mathbf{r}]$ and $\mathfrak{P}'[\mathbf{r}']$. Then, $(\mathfrak{P}, \mathbf{r})$ and $(\mathfrak{P}', \mathbf{r}')$ are called *incompatible with respect to* $\mathcal{F}$ if $(\mathfrak{P}, \mathbf{r}) \oplus (\mathfrak{P}', \mathbf{r}')$ belongs to $\mathcal{F}$.

Let $(\mathfrak{P}, \mathbf{r})$ be a piece of some structure $\mathfrak{F}$ in $\mathcal{F}$. Then, $\mathcal{I}_{(\mathfrak{P}, \mathbf{r})}$ denotes the set of all pieces that are incompatible with $(\mathfrak{P}, \mathbf{r})$ with respect to $\mathcal{F}$.

One can define the following equivalence relation $\sim_{\mathcal{F}}$ on the set of all pieces of $\mathcal{F}$. We say that two pieces $(\mathfrak{P}, \mathbf{r})$ and $(\mathfrak{P}', \mathbf{r}')$ of $\mathcal{F}$ are *equivalent*, denoted $(\mathfrak{P}, \mathbf{r}) \sim_{\mathcal{F}} (\mathfrak{P}', \mathbf{r}')$, if $\mathcal{I}_{(\mathfrak{P}, \mathbf{r})} = \mathcal{I}_{(\mathfrak{P}', \mathbf{r}')}$.

A family of τ-structures $\mathcal{F}$ is called *regular* if the number of the $\sim_{\mathcal{F}}$-equivalence classes is finite.

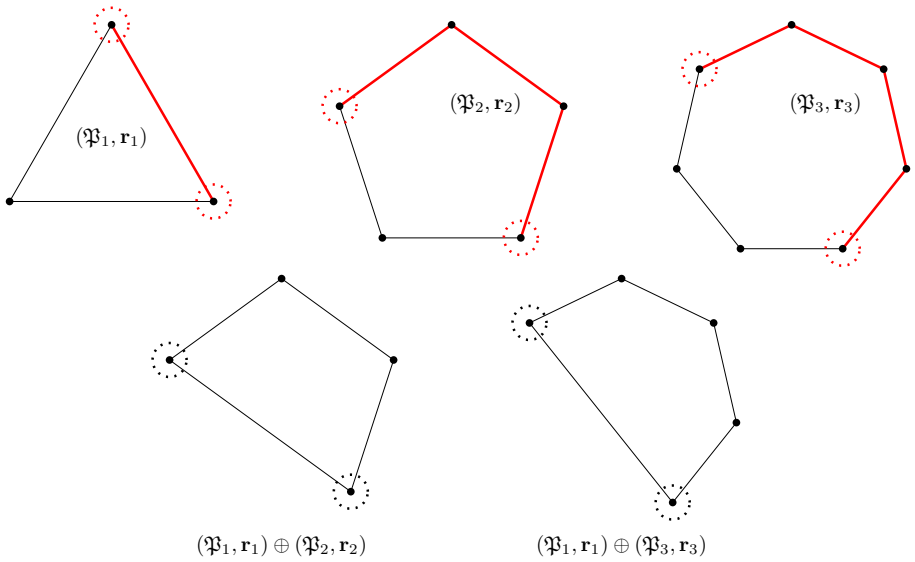


Figure 4.18: Pieces of odd cycles (coloured in red), and graphs obtained from them by the $\oplus$ operation.

Example 4.4.2. Let $\mathcal{F}$ be a family of undirected graphs that consists of odd cycles. Every piece of an odd cycle is a path. There are two equivalence classes of pieces: paths of odd length and paths of even length. As any odd length path is incompatible to any even length path with respect to $\mathcal{F}$. Indeed, if we join them by the $\oplus$ operation, then we will obtain an odd cycle.

Consider three pieces displayed on Figure 4.18. $(\mathfrak{P}_1, \mathbf{r}_1)$ and $(\mathfrak{P}_2, \mathbf{r}_2)$ are paths of odd length. $(\mathfrak{P}_3, \mathbf{r}_3)$ is a path of even length. And, as a result, $(\mathfrak{P}_1, \mathbf{r}_1) \oplus (\mathfrak{P}_2, \mathbf{r}_2)$ is an even cycle that does not belong to $\mathcal{F}$, and $(\mathfrak{P}_1, \mathbf{r}_1) \oplus (\mathfrak{P}_3, \mathbf{r}_3)$ is a member of $\mathcal{F}$. △

A countably infinite relational structure $\mathfrak{A}$ is called *ω-categorical* if all models of the first-order theory of $\mathfrak{A}$ are isomorphic to each other.

The concept of ω-categorical structure has another equivalent definition. An isomorphism $\mathfrak{a}\colon \mathfrak{A} \to \mathfrak{A}$ from a structure to itself is called an *automorphism*. The set $\mathrm{Aut}(\mathfrak{A})$ of automorphisms of $\mathfrak{A}$ is a group with respect to the composition. Consider actions of this group on sets of tuples of elements of A: $A, A^2, A^3, \ldots$, where, for $(a_1, \ldots, a_k)$ in A^k and $\mathfrak{a}$ in $\mathrm{Aut}(\mathfrak{A})$, we put $\mathfrak{a} \cdot (a_1, \ldots, a_k) := \big(\mathfrak{a}(a_1), \ldots, \mathfrak{a}(a_k)\big)$.

For an action of a group G on a set X, the *orbit* of an element x in X is a subset of elements of X where x can be mapped by the elements of G. A group G is called *oligomorphic* if, for any k in $\mathbb{N}$, the action of G on X^k has finitely many orbits.

Proposition 4.4.1. *A countably infinite structure $\mathfrak{A}$ is ω-categorical if and only if $Aut(\mathfrak{A})$ is oligomorphic.*

Example 4.4.3. Consider a structure $(\mathbb{Z}, <)$ – its domain consists of integers, and the only relation is the linear ordering. Any automorphism $\mathfrak{a}_t$ of $(\mathbb{Z}, <)$ is of the form $x \mapsto x + t$, for $t \in \mathbb{Z}$. For any two elements x, y in $\mathbb{Z}$ there exists an automorphism $\mathfrak{a}_{y-x} \in \mathrm{Aut}(\mathbb{Z}, <)$ that maps x to y. This means that all elements are in the same orbit. Thus, $(\mathbb{Z}, <)$ is ω-categorical. △

It is well-known, that, for a finite family of forbidden relational structures there exists a universal ω-categorical structure.

Theorem 4.4.2 (Theorem 4 in [CSS99])**.** *Let $\mathcal{F}$ be a finite set of finite connected relational structures. Then there exists an ω-categorical τ-structure $\mathfrak{B}$ such that, for any finite structure $\mathfrak{A}$, there is an embedding: $\mathfrak{A} \hookrightarrow \mathfrak{B}$ if and only if for any $\mathfrak{F}$ in $\mathcal{F}$ there is no homomorphism from $\mathfrak{F}$ to $\mathfrak{A}$.*

Hubička and Nešetřil extend this result by showing that this property holds not only for finite but also for regular families of forbidden structures.

Theorem 4.4.3 (Theorem 3.1 in [HN15])**.** *Let $\mathcal{F}$ be a regular family of finite connected relational structures. Then there exists an ω-categorical structure $\mathfrak{C}$ such that, for any finite structure $\mathfrak{A}$, there is an embedding: $\mathfrak{A} \hookrightarrow \mathfrak{C}$ if and only if for any $\mathfrak{F}$ in $\mathcal{F}$ there is no homomorphism from $\mathfrak{F}$ to $\mathfrak{A}$.*

Regularity

For showing the regularity of $\mathcal{F}_1$, it suffices to show that the number of $\sim_{\mathcal{F}}$-equivalence classes is finite within any subfamily $\mathcal{F}_{\mathfrak{F}}$, as there are finitely many such subfamilies that make $\mathcal{F}_1$.

Our aim is to prove that the class $\mathcal{F}_1$ is regular, we do it at the end of this part in Theorem 4.4.6. In order to prove this theorem, we represent every piece of a structure of $\mathcal{F}_{\mathfrak{F}}$ in a better-looking form which is based on the labeled tree family construction discussed on page 91. We first describe this form, then justify that we can consider these forms instead of pieces, and finally prove that these forms have the regularity property.

We first define what is the label graph of a structure $\mathfrak{F}$ in $\mathcal{F}_2$.

Definition 11. Take some $(\tau \uplus \sigma)$-structure $\mathcal{F}$ from $\mathcal{F}_2$. Let $\mathfrak{L}_{\mathfrak{F}}$ be an undirected graph with possibly multiple edges that is obtained from $\mathfrak{F}$ as follows.

1. The domain of $\mathfrak{L}_{\mathfrak{F}}$ is the same as the domain F of $\mathfrak{F}$.
2. Edges of $\mathfrak{L}_{\mathfrak{F}}$ have labels from the set $\{l_{\mathbf{x}} \mid \mathbf{x} \in \mathsf{R}^{\mathfrak{F}}\}$.
3. Every k-tuple $\mathbf{x}$ of elements of $\mathfrak{F}$ is replaced in $\mathfrak{L}_{\mathfrak{F}}$ by a k-clique induced on the vertices of $\mathbf{x}$, each edge of this clique has a label $l_{\mathbf{x}}$ that is associated with $\mathbf{x}$.

We call such $\mathfrak{L}_{\mathfrak{F}}$ the *label graph* of $\mathfrak{F}$.

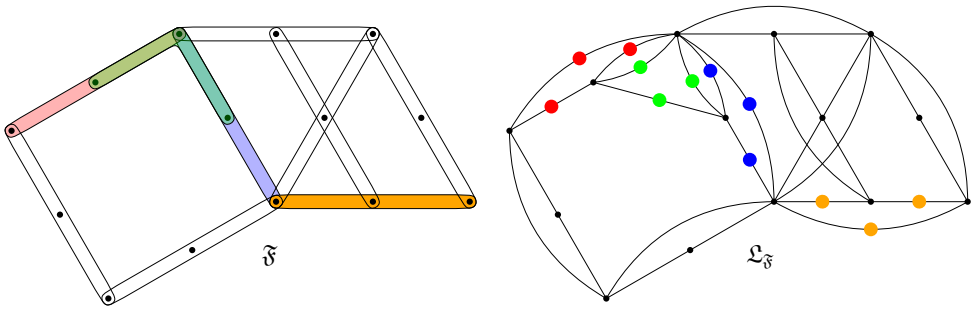


Figure 4.19: An example of a structure and its label graph.

Example 4.4.4. Let $\mathfrak{F}$ be as on Figure 4.19. Every tuple of arity k in $\mathfrak{F}$ is replaced by a k-clique in $\mathfrak{L}_{\mathfrak{F}}$. In our case, every relational tuple of $\mathfrak{F}$ is ternary, so we replace it with a 3-clique K_3. If K_3 substitutes a tuple $\mathbf{t}$, then every edge of it obtains a label associated with $\mathbf{t}$. △

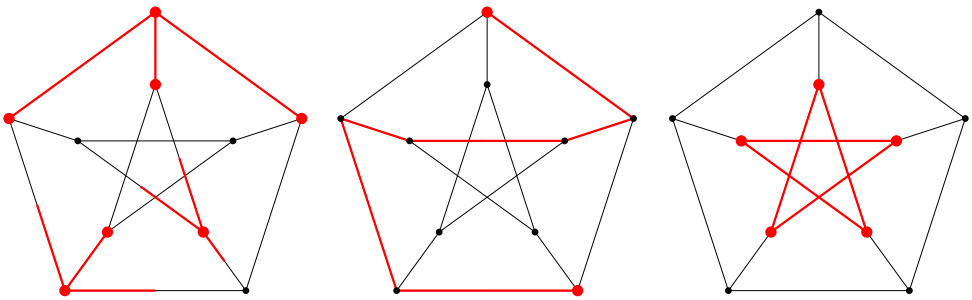
We are going to use the concept of a *semi-edge* that is discussed in [BFJ+22]. Semi-edge is an edge with just one end. Every edge xy contains two semi-edges, one semi-edge is incident to x, the other is incident to y. For every graph $\mathfrak{G} = (G, \mathsf{E}^{\mathfrak{G}})$ there exists a unique set of semi-edges $S_{\mathfrak{G}} = \{(v, e) \mid v \in G, e \in \mathsf{E}^{\mathfrak{G}}, v \text{ is incident to } e\}$.

Remark. We need semi-edges because edges of trees of $\mathcal{T}_{\mathfrak{G}}$ represent pairs of duplicated tuples. Semi-edges help us to describe the case when only one tuple of the pair belongs to a piece.

Definition 12. Let $\mathfrak{G}$ be a graph, $V \subseteq G$ be a subset of vertices, and $S \subset S_{\mathfrak{G}}$ be a subset of semi-edges. We say that a pair (V, S) is *connected* if:

- for any semi-edge s in S there is a vertex v in V such that s is incident to v;
- for any two vertices v and w in V there is a path in $\mathfrak{G}$ such that, for any edge of this path, both its semi-edges are contained in S.

Example 4.4.5. Consider 3 pairs (V_i, S_i), for i in $[3]$, displayed on Figure 4.20. A vertex is red if it belongs to V_i. A semiedge is red if it belongs to S_i. For the pair (V_1, S_1) there exist two vertices of V_1 that are not connected with a path consisting of semi-edges from S_1, so this pair is not connected. For the pair (V_2, S_2) there exist semi-edges in S_2 that are not incident to any vertex of V_2, so this pair is not connected too. The pair (V_3, S_3) satisfies both conditions, so it is connected. △



(V_1, S_1) (V_2, S_2) (V_3, S_3)

Figure 4.20: Two not connected pairs of subsets of vertices and semiedges of the Petersen graph and one connected pair.

Definition 13. Let $\mathfrak{G}$ be in $\mathcal{F}_{\mathfrak{F}}$ and $\mathcal{P}$ be a collection of pairs:

$$\mathcal{P} := \Big\{ \big(V_{\mathfrak{L}_{\mathfrak{F}}}, S_{\mathfrak{L}_{\mathfrak{F}}}\big), \big(V_1, S_1\big), \ldots, \big(V_n, S_n\big) \Big\}.$$

Here, $V_{\mathfrak{L}_{\mathfrak{F}}}$ and $S_{\mathfrak{L}_{\mathfrak{F}}}$ are some subsets of vertices and semi-edges of the graph $\mathfrak{L}_{\mathfrak{F}}$, and $n := |V_{\mathfrak{L}_{\mathfrak{F}}}|$. For i in $[n]$, V_i and S_i are some subsets of vertices and semi-edges of the tree $\mathfrak{T}_{f_i}$ in $\mathcal{T}_{\mathfrak{G}}$, where f_i is the ith element of $V_{\mathfrak{L}_{\mathfrak{F}}} = \{f_1, \ldots, f_n\}$. Such family $\mathcal{P}$ is called *piece alike* if the following conditions hold.

- $\big(V_{\mathfrak{L}_{\mathfrak{F}}}, S_{\mathfrak{L}_{\mathfrak{F}}}\big)$ is connected.
- For any i in $[n]$, for any semi-edge s in S_i there is a vertex v in V_i such that s is incident to v.
- If, for some element f_i of $\mathfrak{L}_{\mathfrak{F}}$, $f_i \notin V_{\mathfrak{L}_{\mathfrak{F}}}$, then the corresponding sets V_i and S_i are empty.
- Suppose that a semi-edge s in $S_{\mathfrak{L}_{\mathfrak{F}}}$ is incident to a vertex v in $V_{\mathfrak{L}_{\mathfrak{F}}}$ and that it belongs to an edge of $\mathfrak{L}_{\mathfrak{F}}$ with a label $l_{\mathbf{x}}$, for some $\mathbf{x}$ in $\mathsf{R}^{\mathfrak{F}}$; then any other semi-edge of $\mathfrak{L}_{\mathfrak{F}}$ which is incident to v and has a label $l_{\mathbf{x}}$ must belong to $S_{\mathfrak{L}_{\mathfrak{F}}}$.
- It is possible to define a connected binary relation P on the set $V_1 \cup \cdots \cup V_n$ such that a pair (x, y) belongs to P if and only if either
  - x and y belong to the same V_i, (x, y) is an edge , and S_i contains both semi-edges of (x, y), or
  - x belongs to some V_i, y belongs to some V_j, (f_i, f_j) is an edge of $\mathfrak{L}_{\mathfrak{F}}$ such that both its semi-edges belong to $S_{\mathfrak{L}_{\mathfrak{F}}}$, and x and y have labels in L_{f_i} and L_{f_j} that are associated with the same R-tuple of $\mathfrak{F}$.

Example 4.4.6. Consider Figure 4.21 on page 101. On the left, there is a structure $\mathfrak{F}$ from $\mathcal{F}_2$. In the middle, there is the label graph $\mathfrak{L}_{\mathfrak{F}}$. As R is ternary, this graph is obtained from $\mathfrak{F}$ by replacing tuples with 3-cliques. Every tuple of $\mathfrak{F}$ is highlighted with its own colour, the edges of every 3-clique of the label graph have the corresponding label. On the right, there is the labeled tree family $\mathcal{T}_{\mathfrak{G}}$, for some structure $\mathfrak{G} \in \mathcal{F}_{\mathfrak{F}}$.

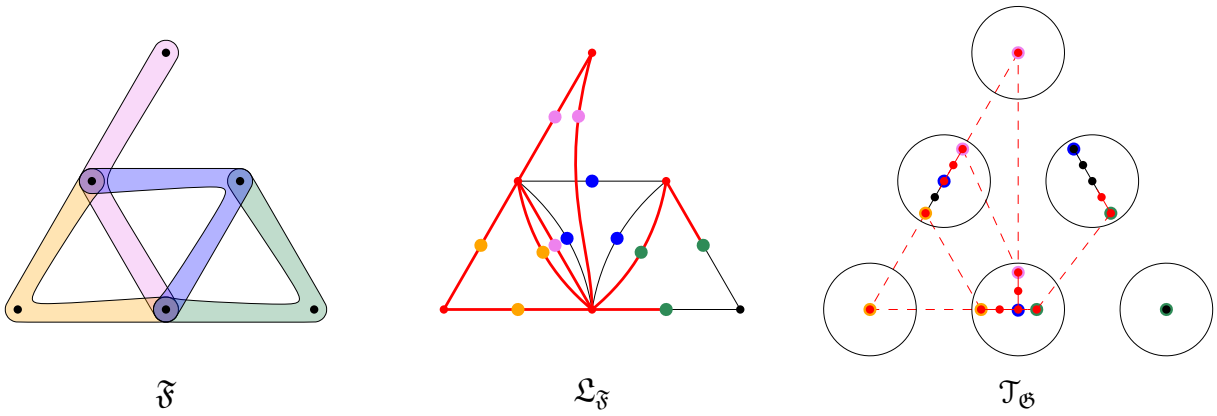


Figure 4.21: An example of a piece alike family for a structure $\mathfrak{G}$ in $\mathcal{F}_{\mathfrak{F}}$.

The vertices and semi-edges $(V_{\mathfrak{L}_{\mathfrak{F}}}, S_{\mathfrak{L}_{\mathfrak{F}}})$ of $\mathfrak{L}_{\mathfrak{F}}$ are highlighted with red. Similarly, for all the trees of $\mathfrak{T}_{\mathfrak{G}}$, these sets of vertices and semi-edges are highlighted with red. The relation P is comprised of red edges and dashed edges between red vertices.

The family on Figure 4.21 is indeed piece alike, it satisfies all the properties. We would also like to point out some details about it.

- The pair (V_i, S_i), for i in $[n]$ does not need to be connected. There is one such tree in the family (the path in the middle of the left side). Both its leaves are connected by P through the tree in the bottom center.

- If the degree of a vertex of $\mathfrak{F}$ is 1, then the corresponding tree of $\mathfrak{T}_{\mathfrak{G}}$ always consists of a single vertex, for any $\mathfrak{G}$ in $\mathcal{F}_1$.

- In the piece alike family definition, it may seem unclear why do we need the condition that requires that all semi-edges incident to the same vertex of $\mathfrak{L}_{\mathfrak{F}}$ must either all belong to $S_{\mathfrak{L}_{\mathfrak{F}}}$ or all not belong to it. This condition arises from the way all $\mathfrak{G}$ in $\mathcal{F}_{\mathfrak{F}}$ are constructed. All of them are obtained from $\tilde{\mathfrak{F}}_\circ$. The construction of $\tilde{\mathfrak{F}}_\circ$ is given on Figure 4.11 on page 90, we recall that it is obtained from $\tilde{\mathfrak{F}}$ by replacing each $(k+1)$-ary $\tilde{\mathsf{R}}$-tuple with k $\tilde{\mathsf{R}}$-tuples. Within $\mathfrak{L}_{\mathfrak{F}}$, each of these k tuples is associated with the set of semi-edges that have the same label and that are incident to the same vertex of $\mathfrak{L}_{\mathfrak{F}}$.

△

Remark. When we just defined a labeled tree family on page 91, we also assigned labels to every edge of every tree of the family. Those edge labels carried the information about the vertex colours of the corresponding pair of duplicated tuples and about the coordinate where the Old vertices were placed. Our current goal is to prove that $\mathcal{F}_1$ is regular and edge-labels are never used here. So, for simplicity, we do not mention them in this part although they always are where they are supposed to be. Sometimes, in particular, in Construction 12 on page 103, we add a new edge to a labeled tree. As we are free to assign any possible label to this edge, we assume that it is assigned arbitrarily.

The next result allows us to represent pieces as piece alike families.

Lemma 4.4.4. *Let $\mathfrak{G}$ be in $\mathcal{F}_{\mathfrak{F}}$. Then, there is a one-to-one correspondence between pieces of $\mathfrak{G}$ and piece alike families constructed from $\mathfrak{G}$.*

Proof. Let $(\mathfrak{P}, \mathbf{r})$ be a piece of $\mathfrak{G}$. The roots $\mathbf{r}$ can contain either $\mathtt{Old}$ elements of $\mathfrak{G}$ or those T-vertices that are in pairs of duplicated $\tilde{\mathsf{R}}$-tuples. As $\mathbf{r}$ is minimal by inclusion, any vertex that is neither $\mathtt{Old}$ nor a T-vertex within a pair of duplicated tuples cannot be contained in $\mathbf{r}$.

There is a one-to-one correspondence between $\mathtt{Old}$ elements of $\mathfrak{G}$ and tree vertices in $\mathcal{T}_{\mathfrak{G}}$. Every $\tilde{\mathsf{R}}$-tuple of $\mathfrak{G}$ is associated either with a set of semi-edges of $\mathfrak{L}_{\mathfrak{F}}$ that are incident to the same vertex or with a semi-edge of some tree of $\mathcal{T}_{\mathfrak{G}}$, this correspondence is also one-to-one.

We construct the piece alike family $\mathcal{P}_{\mathfrak{G}} := \big\{(V_{\mathfrak{L}_{\mathfrak{F}}}, S_{\mathfrak{L}_{\mathfrak{F}}}), (V_1, S_1), \ldots, (V_n, S_n)\big\}$ that is associated with $(\mathfrak{P}, \mathbf{r})$.

- For an $\mathtt{Old}$ element x in G, suppose that it belongs to an $\mathtt{eq}$-equivalence class $[f_i]_{\mathtt{eq}}$; then, if x belongs to the domain P of $\mathfrak{P}$, we add x to V_i and add f_i to $V_{\mathfrak{L}_{\mathfrak{F}}}$.

- Suppose that an $\tilde{\mathsf{R}}$-tuple $(t, \mathbf{x})$ of $\mathfrak{G}$ is in a pair of duplicated tuples and the $\mathtt{Old}$ element x in $\mathbf{x}$ is in some $\mathtt{eq}$-equivalence class $[f_i]_{\mathtt{eq}}$; then we add the semi-edge associated with $(t, \mathbf{x})$ to S_i if $\mathfrak{P}$ has the $\tilde{\mathsf{R}}$-tuple $(t, \mathbf{x})$.

- Suppose that an $\tilde{\mathsf{R}}$-tuple $(t, \mathbf{x})$ belongs to a set of k duplicated tuples that are associated with some R-tuple $\mathbf{f}$ of $\mathfrak{F}$, and that the $\mathtt{Old}$ element x of $\mathbf{x}$ belongs to $[f]_{\mathtt{eq}}$; then, if $(t, \mathbf{x})$ is contained in $\mathfrak{P}$, we add to $S_{\mathfrak{L}_{\mathfrak{F}}}$ all semi-edges incident to f with label $l_{\mathbf{f}}$.

This construction induces an injection, as two distinct pieces have different sets of $\tilde{\mathsf{R}}$-tuples. It is a routine to check that the resulting family $\mathcal{P}$ is piece alike.

For any piece alike family $\mathcal{P}$, we construct a piece $(\mathfrak{P}, \mathbf{r})$ as follows.

- If a set of semi-edges of $\mathfrak{L}_{\mathfrak{F}}$ that have the same label and are incident to the same vertex, and all belong to $\mathcal{P}$, then we add to $\mathfrak{P}$ all the vertices of the corresponding tuple of $\mathfrak{G}$.

- If the semi-edge corresponding to an $\tilde{\mathsf{R}}$-tuple $(t, \mathbf{x})$ is contained in some S_i, then we add to P all the elements of this tuple.

- If a vertex x contained in $V_{\mathfrak{L}_{\mathfrak{F}}}$ or in some V_i is incident to a semi-edge that does not belong to $\mathcal{P}$, then we add x to $\mathbf{r}$.

This construction also induces an injection, as different families are associated with two substructures of $\mathfrak{G}$ that differ either with respect to $\mathtt{Old}$ elements or with respect to $\tilde{\mathsf{R}}$-tuples. We need to show that $(\mathfrak{P}, \mathbf{r})$ is a piece. It is connected because $\mathcal{P}$ is a piece alike family. $\mathbf{r}$ is minimal by inclusion as any vertex of $\mathcal{P}$ incident to a semi-edge not from $\mathcal{P}$ is added to the set of roots $\mathbf{r}$.

This implies that there is a one-to-one correspondence between pieces and piece alike families. Moreover, these two construction mappings are inverses of each other, because there is a one-to-one correspondence between semi-edges of the family and $\tilde{\mathsf{R}}$-tuples of the piece. □

Proposition 4.4.5. *Let $(\mathfrak{P}, \mathbf{r})$ be a piece of some structure $\mathfrak{G}$ in $\mathcal{F}_{\mathfrak{F}}$. Then the size of $\mathbf{r}$ is bounded by $k^2|F|^{2k-1} + k|F|^k$, where k is the arity of R.*

Proof. Let $\mathcal{P}$ be the corresponding piece alike family, provided by Lemma 4.4.4. The binary relation P defined on the set $V_1 \cup \cdots \cup V_n$ is connected. Pick any tree $\mathfrak{T}_f$ of $\mathcal{T}_{\mathfrak{G}}$. Any connected component of P induced on T_f must contain a labeled vertex. The number of labeled vertices is at most the number of distinct R-tuples that can be adjacent to an element f of the structure $\mathfrak{F}$. Thus, it is at most $k|F|^{k-1}$. So, the number of leaves of each tree of $\mathcal{T}_{\mathfrak{G}}$ is at most $k|F|^{k-1}$, as every leaf has a label.

Any vertex of $\mathbf{r}$ is associated either

- with some v in V_i that is incident to a semi-edge not from S_i, for some i in $[n]$, or
- with some v in V_i that has a label $l_{\mathbf{f}}$, for some tuple $\mathbf{f}$ in $\mathsf{R}^{\mathfrak{F}}$, such that the $l_{\mathbf{f}}$-labeled semi-edges incident to f_i are not contained in $S_{\mathfrak{L}_{\mathfrak{F}}}$, or
- with a semi-edge s in S_i such that the other semi-edge of the edge where they belong to is not in S_i, for some i in $[n]$, or
- with a set of semi-edges of $\mathfrak{L}_{\mathfrak{F}}$ that are incident to the same vertex and have the same label.

For any element f_i of $\mathfrak{F}$, and for any connected component of the graph induced by P on $\mathfrak{T}_{f_i}$, the number of vertices and semi-edges corresponding to roots in $\mathbf{r}$ cannot be greater than the number of labeled vertices of $\mathfrak{T}_{f_i}$. This is justified as follows. Pick any vertex x of the component. Call a *branch* every path from x to a labeled vertex, the number of branches is at most to the number of labeled vertices because $\mathfrak{T}_{f_i}$ is a tree. Any branch can contain exactly one vertex y such that either

- y belongs to the same component as x, and the next vertex z of the branch does not belong to this component, or
- y is the end of the branch, *i.e.*, a labeled vertex.

Thus, each $\mathfrak{T}_{f_i}$ contains at most $k^2|F|^{2k-2}$ elements of $\mathbf{r}$, as the number of connected components is at most the number of leaves.

The number of trees in $\mathcal{T}_{\mathfrak{G}}$ equals to $|F|$, the number of sets of semi-edges of $\mathfrak{L}_{\mathfrak{F}}$ that have the same label and are incident to the same vertex is at most $k|F|^k$. So the total number of roots is at most $k^2|F|^{2k-1} + k|F|^k$. □

Consider a structure $\mathfrak{G}$ that belongs to $\mathcal{F}_{\mathfrak{F}}$. Before proving Theorem 4.4.6, we are going to show in Construction 12 how to modify any piece $(\mathfrak{P}, \mathbf{r})$ of $\mathfrak{G}$ so that the resulting piece has bounded size, and the resulting structure still belongs to $\mathcal{F}_{\mathfrak{F}}$.

Construction 12. Consider a piece $(\mathfrak{P}, \mathbf{r})$ of $\mathfrak{G}$. Let $\mathcal{P}$ be a piece alike family that is associated with $(\mathfrak{P}, \mathbf{r})$, by Lemma 4.4.4. Denote by $\overline{\mathfrak{P}}$ the substructure of $\mathfrak{G}$ induced on $G \setminus (P \setminus \mathbf{r})$. By the definition of $\oplus$ operation, $\mathfrak{G}$ is isomorphic to $(\mathfrak{P}, \mathbf{r}) \oplus (\overline{\mathfrak{P}}, \mathbf{r})$. Our goal is to provide a finite family $\mathcal{R}$ of rooted structures such that, for each given $\mathfrak{G}$ of $\mathcal{F}_1$ and any its piece $(\mathfrak{P}, \mathbf{r})$, $\mathcal{R}$ will contain a piece $(\mathfrak{P}', \mathbf{r}')$ such that $\mathfrak{P}[\mathbf{r}]$ is isomorphic to $\mathfrak{P}'[\mathbf{r}']$ and that $\mathfrak{G}' := (\mathfrak{P}', \mathbf{r}') \oplus (\overline{\mathfrak{P}}, \mathbf{r})$ belongs to $\mathcal{F}_1$. If we show that any such $(\mathfrak{P}, \mathbf{r})$ has bounded size, then this will imply that there are finitely many of them.

We describe the process of obtaining $(\mathfrak{P}', \mathbf{r}')$ from $(\mathfrak{P}, \mathbf{r})$ by modifying $\mathcal{P}$ and the trees of $\mathcal{T}_{\mathfrak{G}}$. Let $\mathfrak{T}_{f_i}$ be some labeled tree in $\mathcal{T}_{\mathfrak{G}}$ and V_i and S_i be subsets of its vertices and semi-edges such that the pair (V_i, S_i) is in $\mathcal{P}$. We mentioned before, in Example 4.4.6, that

(V_i, S_i) is not always connected. So we choose any of its connected components induced by the relation P. Let (V_i^c, S_i^c) denote the vertices and semi-edges of some connected component of (V_i, S_i). We highlight the important vertices of V_i^c as follows:

- Labeled vertices. Their number is bounded by $k|F|^k$, where $|F|$ is the domain size of $\mathfrak{F} \in \mathcal{F}_2$.
- Vertices of degree strictly greater than 2. The number of such vertices is as most the number of leaves in $\mathfrak{T}_{f_i}$, which is at most the number of labeled vertices.
- Vertices that are adjacent to a vertex not from V_i. This means that either such a vertex is incident to a semi-edge not from S_i, or this vertex is incident to an edge such that one of its semi-edges is not in S_i. The number of these vertices is at most the number of roots $\mathbf{r}$. So their number is also bounded, by Proposition 4.4.5.

All other vertices of V_i^c have degree 2 and both incident semi-edges belong to S_i (and, consequently, to S_i^c), their number may be arbitrarily large. Our aim is to get rid of them, leaving only vertices of the listed classes. Let v be one such vertex, let u and x be its neighbours. The procedure is as follows, it is displayed on.

1. Delete v and both incident edges uv, vx from $\mathfrak{T}_{f_i}$. This will automatically remove v from V_i and all 4 semi-edges $(u, uv), (v, uv), (v, vx), (x, vx)$ from S_i.
2. Add an edge ux to $\mathfrak{T}_{f_i}$ and both semi-edges $(u, ux), (x, ux)$ to S_i.

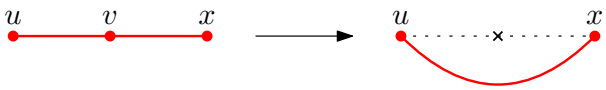


Figure 4.22: The procedure of Construction 12.

We repeat this procedure for every vertex that does not belong to any of the three highlighted classes until there are no such vertices. The result of this procedure applied for one connected component is displayed on Figure 4.23. Do this for every connected component of every tree of $\mathcal{T}_{\mathfrak{G}}$. This finishes Construction 12.

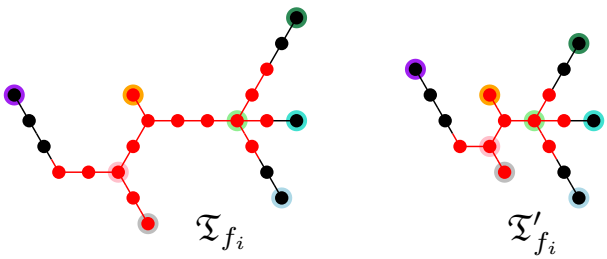


Figure 4.23: On the left, there is the original labeled tree $\mathfrak{T}_{f_i}$. The elements of (V_i, S_i) are coloured in red. The circles are the labels. On the right, there is the modified labeled tree $\mathfrak{T}'_{f_i}$.

Theorem 4.4.6. *The class $\mathcal{F}_1$ is regular.*

Proof. Consider some structure $\mathfrak{G} \in \mathcal{F}_{\mathfrak{F}}$ and one of its pieces $(\mathfrak{P}, \mathbf{r})$ that is represented by a piece alike family $\mathcal{P}$. Denote by $\mathcal{T}'$ and by $\mathcal{P}'$ the tree and piece alike families obtained by Construction 12. By Proposition 4.3.6 from page 93, we know that there is a structure $\mathfrak{G}'$ in $\mathcal{F}_1$ such that $\mathcal{T}_{\mathfrak{G}'} = \mathcal{T}'$. The rest is to show that, for any pair (V_i', S_i') in $\mathcal{P}'$, $|V_i'|$ and $|S_i'|$ are bounded. The maximal degree of a vertex in a tree is at most the number of leaves in this tree. So, as the number of leaves of any tree of $\mathcal{T}_{\mathfrak{G}'}$ is bounded by the number of labels, by Proposition 4.3.5 on page 92, the boundedness of $|S_i'|$ is implied by the boundedness of $|V_i'|$. The set V_i' consists only of vertices of the three types described in Construction 12. The number of vertices of each type is bounded. □

Equivalence of ω-categorical templates

Let Φ in MMSNP_2 be a τ-sentence in normal$_1$ form and σ be the set of existentially quantified relations of Φ. We are going to show that there is an ω-categorical $(\tau \uplus \sigma)$-structure $\mathfrak{C}_\Phi$ such that a τ-structure $\mathfrak{A}$ satisfies Φ if and only if $\mathfrak{A}$ homomorphically maps to the τ-reduct $\mathfrak{C}_\Phi^\tau$.

The existence of a similar structure for MMSNP sentences is shown in [BMM18]. The MMSNP_2 case is treated similarly.

Theorem 4.4.7. *Let $\mathcal{F}_2$ be a finite set of finite connected $(\tau \uplus \sigma)$-structures corresponding to Φ in MMSNP_2. Denote by $\mathfrak{B}_{ind}^{\mathcal{F}_2}$ the universal ω-categorical structure provided by Theorem 4.4.2. That is, for any $(\tau \uplus \sigma)$-structure $\mathfrak{A}$, $\mathfrak{A}$ embeds into $\mathfrak{B}_{ind}^{\mathcal{F}_2}$ if and only if no structure $\mathfrak{F}$ in $\mathcal{F}_2$ homomorphically maps to $\mathfrak{A}$. Then there is an ω-categorical structure $\mathfrak{C}_\Phi$ such that a finite τ-structure $\mathfrak{A}$ satisfies Φ if and only if there is a homomorphism from $\mathfrak{A}$ to $\mathfrak{C}_\Phi^\tau$.*

Proof. Let $\mathfrak{C}_\Phi$ be the substructure of $\mathfrak{B}_{ind}^{\mathcal{F}_2}$ that contains only σ_{V}-coloured vertices of $\mathfrak{B}_{ind}^{\mathcal{F}_2}$ and contains only σ_{T}-coloured R-tuples. By Theorem 4.4.2, $\mathfrak{B}_{ind}^{\mathcal{F}_2}$ is ω-categorical. Thus, its automorphism group is oligomorphic, by Proposition 4.4.1. Any automorphism $\mathfrak{a}$ in $\mathrm{Aut}(\mathfrak{B}_{ind}^{\mathcal{F}_2})$ induces an automorphism on $\mathfrak{C}_\Phi$, so $\mathrm{Aut}(\mathfrak{C}_\Phi)$ is also oligomorphic. Any automorphism of a structure is an automorphism of its reduct, so $\mathrm{Aut}(\mathfrak{C}_\Phi^\tau)$ is also oligomorphic. We conclude that both $\mathfrak{C}_\Phi$ and its τ-reduct are ω-categorical.

Suppose that a finite τ-structure $\mathfrak{A}$ satisfies Φ. Then there exists a σ-expansion $\mathfrak{A}^\sigma$ of $\mathfrak{A}$, where any vertex and any R-tuple is coloured in precisely one colour, such that no $\mathfrak{F}$ in $\mathcal{F}_2$ homomorphically maps to $\mathfrak{A}^\sigma$. Then $\mathfrak{A}^\sigma$ embeds into $\mathfrak{B}_{ind}^{\mathcal{F}_2}$, as it is all coloured, it also embeds into $\mathfrak{C}_\Phi$. Thus, $\mathfrak{A}$ embeds into the τ-reduct $\mathfrak{C}_\Phi^\tau$.

Suppose that there is a homomorphism $\mathfrak{h}\colon \mathfrak{A} \to \mathfrak{C}_\Phi^\tau$. Assign to any a of A the same σ_{V} colour that the image $\mathfrak{h}(a)$ has in $\mathfrak{C}_\Phi$. Assign to any R-tuple $\mathbf{a}$ of $\mathfrak{A}$ the same σ_{T} colour that the tuple of the image $\mathfrak{h}(\mathbf{a})$ has in $\mathfrak{C}_\Phi$. Each element and each R-tuple of $\mathfrak{C}_\Phi$ has precisely one colour because Φ is in normal$_1$ form. Denote the corresponding σ-expansion by $\mathfrak{A}^\sigma$, by its construction, it maps homomorphically to $\mathfrak{C}_\Phi$. If there is $\mathfrak{F}$ in $\mathcal{F}_2$ that maps to $\mathfrak{A}^\sigma$, then, by transitivity, it maps to $\mathfrak{C}_\Phi$, that is a contradiction. So, we have $\mathfrak{A}$ is accepted by $\mathrm{FPP}_2(\mathcal{F}_2)$, thus, $\mathfrak{A}$ satisfies Φ. □

Denote $\mathfrak{C}_{sem} := \tilde{\mathsf{D}}(\mathfrak{C}_\Phi^\tau)$ – the functorial image of the CSP template constructed in Theorem 4.4.7.

Let $\mathfrak{B}_{ind}^{\mathcal{F}_1}$ be the universal ω-categorical $(\tilde{\tau} \uplus \tilde{\sigma})$-structure provided by Theorem 4.4.3 of Hubička and Nešetřil. Denote by $\mathfrak{C}_{synt}$ the τ-reduct of the substructure of $\mathfrak{B}_{ind}^{\mathcal{F}_1}$ induced

by coloured vertices and coloured $\tilde{\mathsf{R}}$-tuples. By arguments similar to the ones in the proof of Theorem 4.4.7, we can obtain the following result.

Theorem 4.4.8. *For any regular family $\mathcal{F}_1$ of $(\tilde{\tau} \uplus \tilde{\sigma})$-structures there exists an ω-categorical structure $\mathfrak{C}_{synt}$ such that, for any finite $\tilde{\tau}$-structure $\mathfrak{A}$, $\mathfrak{A}$ is accepted by the problem $FPP_1^{\infty}(\mathcal{F}_1)$ if and only if $\mathfrak{A}$ homomorphically maps to $\mathfrak{C}_{synt}$.*

Observe that the structure $\mathfrak{C}_{sem}$ is a good structure, as it belongs to the functorial image. And the structure $\mathfrak{C}_{synt}$ contains duplicated tuples. In the rest of this section, we prove that, despite this difference, the two structures are homomorphically equivalent. So they define the same CSP problem.

Lemma 4.4.9. *For any finite induced substructure $\mathfrak{G}$ of $\mathfrak{C}_{sem}$, $\mathfrak{G}$ homomorphically maps to $\mathfrak{C}_{synt}$.*

Proof. Observe that $\mathfrak{G}$ is good as every substructure of $\mathfrak{C}_{sem}$ is good, so there exists $\mathfrak{G}_2$ such that $\mathfrak{G}$ is isomorphic to the functorial image $\tilde{\mathsf{D}}(\mathfrak{G}_2)$. Then, suppose that $\mathfrak{G}$ does not map to $\mathfrak{C}_{synt}$; then there is a forbidden structure $\mathfrak{F}_1$ in $\mathcal{F}_{\mathfrak{F}} \subset \mathcal{F}_1$ that maps to $\mathfrak{G}$. By Lemma 4.3.7, the quotient $\mathfrak{F}_1/\mathtt{eq}$ maps to $\mathfrak{G}$, as $\mathfrak{G}$ is good. Then, by transitivity, we have $\tilde{\mathfrak{F}}$ maps to $\mathfrak{G}$, as it maps to any $\mathtt{eq}$-quotient of $\mathcal{F}_{\mathfrak{F}}$ structures. But $\tilde{\mathfrak{F}}$ is the functorial image of $\mathfrak{F}$ in $\mathcal{F}_2$. By Proposition 4.1.2, we know that $\mathfrak{F}$ maps to $\mathfrak{G}_2$ if and only if $\tilde{\mathfrak{F}}$ maps to $\mathfrak{G}$. We conclude that $\mathfrak{F}$ maps to $\mathfrak{G}_2$, and thus $\mathfrak{G}_2$ does not map to $\mathfrak{C}_2$. So, also by Proposition 4.1.2, $\mathfrak{G}$ does not map to $\mathfrak{C}_{sem}$, that is a contradiction. □

Lemma 4.4.10. *For any finite induced substructure $\mathfrak{G}$ of $\mathfrak{C}_{synt}$, $\mathfrak{G}$ homomorphically maps to $\mathfrak{C}_{sem}$.*

Proof. By Lemma 4.3.7, $\mathfrak{G}$ maps to $\mathfrak{C}_{sem}$ if and only if $\mathfrak{G}/\mathtt{eq}$ maps to $\mathfrak{C}_{sem}$, as $\mathfrak{C}_{sem}$ is a good structure. Suppose that $\mathfrak{G}/\mathtt{eq}$ does not map to $\mathfrak{C}_{sem}$. The structure $\mathfrak{G}/\mathtt{eq}$ is good, so there exists a $(\tau \uplus \sigma)$-structure $\mathfrak{G}_2$ such that $\mathfrak{G}/\mathtt{eq} = \tilde{\mathsf{D}}(\mathfrak{G}_2)$. $\mathfrak{G}_2$ does not map to $\mathfrak{C}_2$, by Proposition 4.1.2. So, there is a $(\tau \uplus \sigma)$-structure $\mathfrak{F}$ in $\mathcal{F}_2$ that maps to $\mathfrak{G}_2$. Then $\tilde{\mathfrak{F}}$ maps to $\mathfrak{G}/\mathtt{eq}$, by Proposition 4.1.2. Then, by Lemma 4.3.8, there is a forbidden structure $\mathfrak{F}'$ in $\mathcal{F}_1$ that maps to $\mathfrak{G}$. So $\mathfrak{G}$ does not map to $\mathfrak{C}_{synt}$, that is a contradiction. □

It is well known that there is a homomorphism (embedding) between two ω-categorical structures $\mathfrak{A}$ and $\mathfrak{B}$ if any finite substructure of $\mathfrak{A}$ maps to $\mathfrak{B}$.

Lemma 4.4.11 (Lemma 4.1.7 in [Bod21])**.** *Let $\mathfrak{B}$ be a finite or countably infinite ω-categorical structure with relational signature τ, and let $\mathfrak{A}$ be a countable τ-structure. If there is no homomorphism (embedding) from $\mathfrak{A}$ to $\mathfrak{B}$, then there is a finite substructure of $\mathfrak{A}$ that does not map homomorphically (embed) to $\mathfrak{B}$.*

Lemma 4.4.9, Lemma 4.4.10 and Lemma 4.4.11 imply that the two ω-categorical templates constructed in this section are homomorphically equivalent.

Corollary 4.4.12. *$\mathfrak{C}_{sem}$ and $\mathfrak{C}_{synt}$ are homomorphically equivalent.*

Chapter 5

MMSNP_2 and expander structures

We reduce MMSNP_2 to finite CSP through MMSNP by Feder and Vardi's method from [FV98]. We provide a new notion of normal form for MMSNP_2 sentences that eases a potential reduction in the other direction by providing new useful tools. And, finally, we study the main obstacle that prevents us from proving a dichotomy for MMSNP_2. This obstacle is to construct, for a given structure, an equivalent (w.r.t. a given CSP) structure that is sufficiently sparse, in a sense. We consider different approaches that try to get over this obstacle, for a better understanding of this problem.

5.1 Introduction to expanders

We know, by Proposition 4.1.1 in Section 4.1, that MMSNP_2 problems can be reduced to MMSNP problems. Moreover, once restricted on the good input, this reduction becomes a P-time equivalence. We can detect ugly input structures and reject them. The only obstacle that stops us from showing a dichotomy for MMSNP_2 is the inability to deal with structures containing duplicated tuples.

It is not known if there is an approach to show a dichotomy for MMSNP_2 other than trying to show that $\mathrm{SAT}(\Phi)$ is equivalent to a CSP problem under (possibly randomized) P-time reductions, where Φ is a connected MMSNP_2 sentence. That is, similarly as it is done for MMSNP by Feder and Vardi in [FV98]. Their approach is as follows: the input is transformed according to Construction 13, and the target structure that defines the CSP problem is obtained from an MMSNP sentence $\tilde{\Phi}$ according to Construction 14.

We first describe how Feder and Vardi reduce MMSNP problems to CSP problems, then we briefly show how they reduce CSP back to MMSNP, and finally we conclude what can be done for our case of MMSNP_2 problems.

In the following construction we explain how to transform the input of an MMSNP problem on $\tilde{\tau}$-structures to the input of the CSP problem.

Construction 13. For a family of $(\tau \uplus \sigma)$-structures $\mathcal{F}$, denote by $\mathcal{F}^\tau$ the set of their τ-reducts: $\mathcal{F}^\tau := \{\mathfrak{F}^\tau \mid \mathfrak{F} \in \mathcal{F}\}$; when $\mathcal{F}$ is a family of $(\tilde{\tau} \uplus \tilde{\sigma})$-structures, denote a similar set by $\mathcal{F}^{\tilde{\tau}}$. Recall the functor $\tilde{\mathsf{D}}$ defined on page 79: it maps a $(\tau \uplus \sigma)$-structure $\mathfrak{F}$ to a $(\tilde{\tau} \uplus \tilde{\sigma})$-structure $\tilde{\mathfrak{F}} := \tilde{\mathsf{D}}(\mathfrak{F})$. Denote by $\tilde{\mathsf{D}}(\mathcal{F}^\tau)$ the family of $\tilde{\mathsf{D}}$-images of the members of $\mathcal{F}^\tau$: $\tilde{\mathsf{D}}(\mathcal{F}^\tau) := \left\{\tilde{\mathsf{D}}(\mathfrak{F}^\tau) \mid \mathfrak{F} \in \mathcal{F}\right\}$. Let ρ be a finite relational signature corresponding to the CSP problem that we construct. For every $\tilde{\tau}$-reduct $\mathfrak{G}$ of $\tilde{\mathsf{D}}(\mathcal{F}^\tau)$, we add to ρ a relation symbol $\mathsf{S}_\mathfrak{G}$ of arity $m_{\mathsf{S}_\mathfrak{G}} := |G|$ equal to the number of elements in the domain of $\mathfrak{G}$.

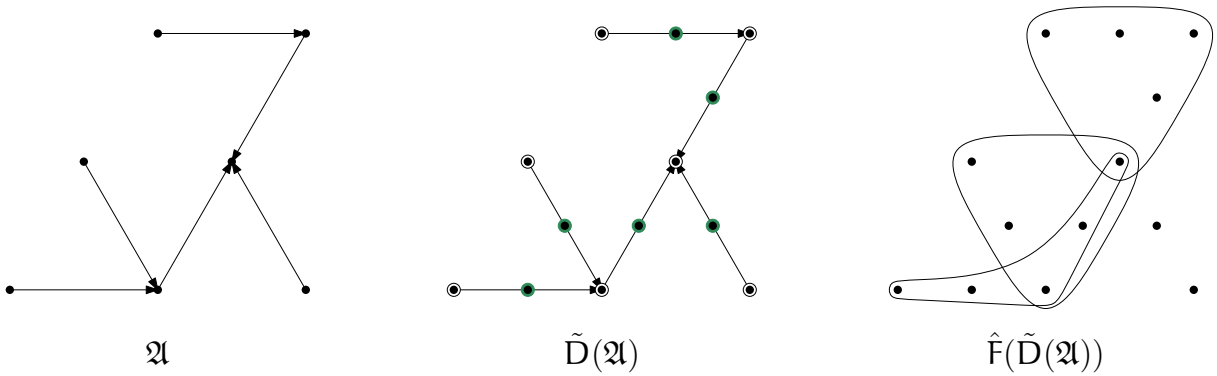


Figure 5.1: On the left, there is the original τ-structure $\mathfrak{A}$. In the middle, there is its $\tilde{\mathsf{D}}$-image $\tilde{\mathfrak{A}}$. On the right, there is the $\hat{\mathsf{F}}$-image of $\tilde{\mathfrak{A}}$. Green disks are T-vertices, white disks are V-vertices. Closed curves are ρ-relational tuples.

We assume that the elements of $\mathfrak{G}$ are linearly ordered, this ordering is defined by a one-to-one correspondence $\mathtt{ord}_{\mathfrak{G}}\colon G \to \big[|G|\big]$. Also assume that the set of T-vertices precedes the set of V-vertices in this ordering, that is, for any T-vertex t and for any V-vertex v in G, we have $\mathtt{ord}_{\mathfrak{G}}(t) < \mathtt{ord}_{\mathfrak{G}}(v)$. For any relation symbol $\mathsf{S}_{\mathfrak{G}}$ in ρ that has arity m, we frequently say that it has arity $(m_{\mathsf{T}}, m_{\mathsf{V}})$: m_{T} denotes the number of T-vertices in $\mathfrak{G}$, they are on the coordinates from 1 to m_{T}; and m_{V} denotes the number of V-vertices in $\mathfrak{G}$, they are on the coordinates from $m_{\mathsf{T}}+1$ to m.

Define a mapping $\hat{\mathsf{F}}\colon \mathbf{Struct}[\tilde{\tau}] \to \mathbf{Struct}[\rho]$ as follows. Let $\mathfrak{A}$ be a $\tilde{\tau}$-structure. The domain of the corresponding ρ-structure $\hat{\mathfrak{A}} := \hat{\mathsf{F}}(\mathfrak{A})$ is the same as the domain of $\mathfrak{A}$. For any $|G|$-ary tuple $\mathbf{x} = (x_1, \ldots, x_{|G|})$ of elements of $\hat{A}$ and for any relation symbol $\mathsf{S}_{\mathfrak{G}}$ in ρ, the tuple $\mathbf{x}$ belongs to the relation $\mathsf{S}_{\mathfrak{G}}^{\hat{\mathfrak{A}}}$ if and only if the mapping $\mathfrak{h}\colon G \to A$ defined by $\mathfrak{h}\colon g \mapsto x_{\mathtt{ord}_{\mathfrak{G}}(g)}$ is a homomorphism from $\mathfrak{G}$ to $\mathfrak{A}$. That is, we find all possible ways to map $\mathfrak{G}$ to $\mathfrak{A}$ and highlight them with the relation $\mathsf{S}_{\mathfrak{G}}^{\hat{\mathfrak{A}}}$. This is the end of Construction 13.

Proposition 5.1.1. *The mapping* $\hat{\mathsf{F}}\colon \mathbf{Struct}[\tilde{\tau}] \to \mathbf{Struct}[\rho]$ *from Construction 13 is a functor.*

Proof. Let $\mathfrak{h}\colon \mathfrak{A} \to \mathfrak{B}$ be a homomorphism between two $\tilde{\tau}$-structures. Let $\hat{\mathfrak{A}}$ and $\hat{\mathfrak{B}}$ be the corresponding $\hat{\mathsf{F}}$-images. We are going to prove that the same mapping $\mathfrak{h}$ is also a homomorphism between $\hat{\mathfrak{A}}$ and $\hat{\mathfrak{B}}$. Pick any ρ-tuple $\mathbf{x} = \big(x_1, \ldots, x_{|G|}\big)$ from $\mathsf{S}_{\mathfrak{G}}^{\hat{\mathfrak{A}}}$, for some $\mathsf{S}_{\mathfrak{G}}$ in ρ. It suffices to prove that the image $\mathfrak{h}(\mathbf{x})$ of this tuple belongs to $\mathsf{S}_{\mathfrak{G}}^{\hat{\mathfrak{B}}}$. By assumption, there is a homomorphism $\mathfrak{g}\colon \mathfrak{G} \to \mathfrak{A}$ such that, for all g in G, $\mathfrak{g}(g) = x_{\mathtt{ord}_{\mathfrak{G}}(g)}$. Then $\mathfrak{h} \circ \mathfrak{g}\colon \mathfrak{G} \to \mathfrak{B}$ is a homomorphism. This implies that, for all g in G, $(\mathfrak{h} \circ \mathfrak{g})(g) = \mathfrak{h}\big(x_{\mathtt{ord}_{\mathfrak{G}}(g)}\big)$, so $\mathfrak{h}(\mathbf{x})$ belongs to $\mathsf{S}_{\mathfrak{G}}^{\hat{\mathfrak{B}}}$. □

Example 5.1.1. Suppose that the family $\mathcal{F}_2$ consists of directed paths of length 3 (that is, containing 2 arcs having the same orientation and 3 vertices). Then ρ contains one $(2,3)$-ary relation symbol corresponding to such a path. On Figure 5.1 on page 108 we show how the functor $\hat{\mathsf{F}}$ works for good structures. There are three ways to map the path to $\mathfrak{A}$. So the structure $\hat{\mathsf{F}}\big(\tilde{\mathsf{D}}(\mathfrak{A})\big)$ contains three ρ-tuples corresponding to paths. As $\tilde{\mathsf{D}}(\mathfrak{A})$ is a good structure, every two ρ-tuples of $\hat{\mathsf{F}}\big(\tilde{\mathsf{D}}(\mathfrak{A})\big)$ that have a common T-vertex also have in common the corresponding V-vertices.

△

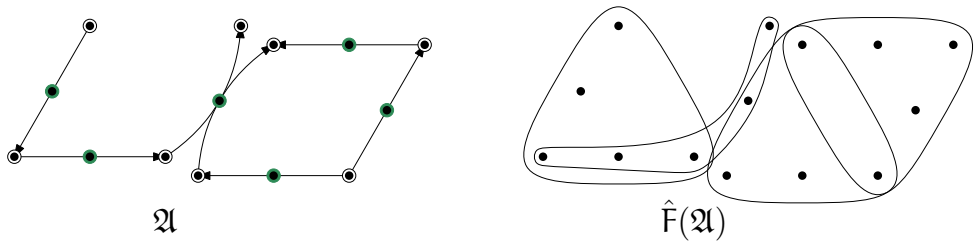


Figure 5.2: On the left, there is a $\tilde{\tau}$-structure $\mathfrak{A}$. On the right, there is the $\hat{\mathsf{F}}$-image of $\mathfrak{A}$. Green disks are T-vertices, white disks are V-vertices. Closed curves are ρ-relational tuples.

Example 5.1.2. Similarly as in Example 5.1.1, suppose that there is only one τ-reduct up to isomorphism, and that it is isomorphic to a directed path of length 3. Consider a $\tilde{\tau}$-structure $\mathfrak{A}$ that is not good, that is, it contains a pair of duplicated tuples. See Figure 5.2 on page 109 for an example of such structure. Then, the ρ-tuples that are associated with directed paths might share a single T-vertex.

△

The following construction explains how to obtain the ρ-structure $\mathfrak{T}_{\tilde{\Phi}}$ from a given MMSNP sentence $\tilde{\Phi}$. It is the target structure of the problem CSP($\mathfrak{T}_{\tilde{\Phi}}$) to which we reduce SAT($\tilde{\Phi}$).

Construction 14. The domain $T_{\tilde{\Phi}}$ of $\mathfrak{T}_{\tilde{\Phi}}$ is the set of existential relations $\tilde{\sigma} = \tilde{\sigma}_{\mathsf{T}} \uplus \tilde{\sigma}_{\mathsf{V}}$. Denote by T_{T} those elements of $T_{\tilde{\Phi}}$ that are associated with $\tilde{\sigma}_{\mathsf{T}}$, and denote by T_{V} those ones that are associated with $\tilde{\sigma}_{\mathsf{V}}$. Consider a relation symbol $\mathsf{S}_{\mathfrak{G}}$ in ρ, suppose that it has arity $(m_{\mathsf{T}}, m_{\mathsf{V}})$. Consider a tuple $(\mathbf{t}, \mathbf{v})$, where $\mathbf{t}$ consists of m_{T} elements of T_{T}, and $\mathbf{v}$ consists of m_{V} elements of T_{V}. We add $(\mathbf{t}, \mathbf{v})$ to $\mathsf{S}_{\mathfrak{G}}^{\mathfrak{T}_{\tilde{\Phi}}}$ if the $\tilde{\sigma}$-expansion of $\mathfrak{G}$, where we assign to the element g_i the colour corresponding to the ith element of the tuple $(\mathbf{t}, \mathbf{v})$, does not belong to $\tilde{\mathsf{D}}(\mathcal{F}_2)$. That is, we add to $\mathsf{S}_{\mathfrak{G}}^{\mathfrak{T}_{\tilde{\Phi}}}$ all tuples associated with $\tilde{\sigma}$-colourings of $\mathfrak{G}$ not from $\tilde{\mathsf{D}}(\mathcal{F}_2)$.

Proposition 5.1.2. *A problem SAT($\tilde{\Phi}$) reduces to CSP($\mathfrak{T}_{\tilde{\Phi}}$) in P-time.*

Proof. For any $\tilde{\tau}$-structure $\mathfrak{A}$, that is an input instance of SAT($\tilde{\Phi}$), we reduce it to a ρ-structure $\hat{\mathfrak{A}} := \hat{\mathsf{F}}(\mathfrak{A})$ – the functorial image of $\mathfrak{A}$.

Suppose that there is a homomorphism $\mathsf{h}\colon \hat{\mathfrak{A}} \to \mathfrak{T}_{\tilde{\Phi}}$. Then, consider a $\tilde{\sigma}$-expansion $\mathfrak{A}^{\tilde{\sigma}}$ of $\mathfrak{A}$ such that we assign to each element a of A a relation of $\tilde{\sigma}$ corresponding to the element $\mathsf{h}(a)$ of T. Suppose that this expansion does not satisfy $\tilde{\Phi}$; then, there is some $\mathfrak{G}$ in $\tilde{\mathsf{D}}(\mathcal{F}_2)$ that homomorphically maps to $\mathfrak{A}^{\tilde{\sigma}}$. Any element of $\mathfrak{G}$ is coloured with exactly one $\tilde{\sigma}$-relation, because we assume that Φ is in normal$_1$ form. Then, there is an $\mathsf{S}_{\mathfrak{G}}$-tuple in $\hat{\mathfrak{A}}$ that is associated with the image of $\mathfrak{G}$ in $\mathfrak{A}$. The h-image of this tuple cannot belong to $\mathsf{S}_{\mathfrak{G}}^{\mathfrak{T}_{\tilde{\Phi}}}$, this contradicts h being a homomorphism.

Suppose that $\mathfrak{A}$ models $\tilde{\Phi}$. Then there is a $\tilde{\sigma}$-expansion $\mathfrak{A}^{\tilde{\sigma}}$ such that no $\mathfrak{G}$ in $\tilde{\mathsf{D}}(\mathcal{F}_2)$ maps to $\mathfrak{A}^{\tilde{\sigma}}$. As Φ is in normal$_1$ form, we assume that, for any element a of A there is a unique $\tilde{\sigma}$-relation assigned to it. Construct a mapping h as follows. Pick an element a in A, suppose without loss of generality that a is a T-vertex that is coloured with a relation $\tilde{\mathsf{M}}$. Then, set $\mathsf{h}(a) := t_{\tilde{\mathsf{M}}}$, where $t_{\tilde{\mathsf{M}}}$ is an element of T that is associated with $\tilde{\mathsf{M}}$. Suppose that h is not a homomorphism, then, for some S in ρ there is a tuple $\mathbf{x}$ that belongs to $\mathsf{S}^{\hat{\mathfrak{A}}}$ and its image $\mathsf{h}(\mathbf{x})$ does not belong to $\mathsf{S}^{\mathfrak{T}_{\tilde{\Phi}}}$. This means that there is a structure $\mathfrak{G}'$ in $\tilde{\mathsf{D}}(\mathcal{F}_2)$ that maps to $\mathfrak{A}^{\tilde{\sigma}}$, it is a contradiction. □

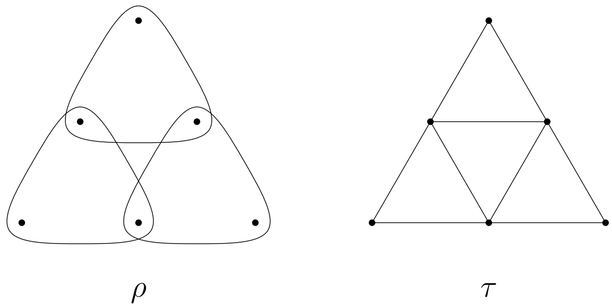
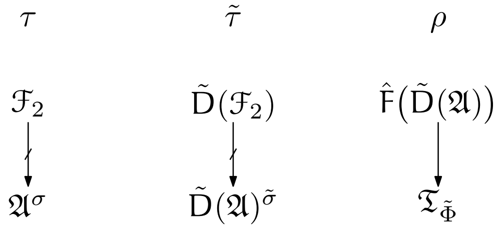


Figure 5.3: On the left there are three ρ-tuples. They are replaced by the corresponding τ-reducts that have here the form of a triangle. The result, on the right, contains (in the center) an implicit triangle.

Let Φ be the MMSNP_2 sentence such that $\tilde{\Phi}$ is obtained from it by Construction 7 from page 75. So Proposition 4.1.1 from page 77 and Proposition 5.1.2 together imply that SAT(Φ) reduces to CSP($\mathfrak{T}_{\tilde{\Phi}}$), where Φ is a sentence in MMSNP_2.

Corollary 5.1.3. *For any sentence Φ in MMSNP_2 in normal$_1$ form there is a finite structure $\mathfrak{T}_{\tilde{\Phi}}$ such that SAT(Φ) reduces in P-time to CSP($\mathfrak{T}_{\tilde{\Phi}}$).*

The diagram below explains how the three problems are related to each other. A τ-structure $\mathfrak{A}$ satisfies Φ if and only if there is a σ-expansion $\mathfrak{A}^\sigma$ such that no structure from $\mathcal{F}_2$ is mapped to $\mathfrak{A}^\sigma$ or, equivalently, iff the ρ-structure $\hat{\mathsf{F}}\big(\tilde{\mathsf{D}}(\mathfrak{A})\big)$ is mapped to $\mathfrak{T}_{\tilde{\Phi}}$.

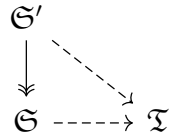


The least trivial part of Feder and Vardi's proof is the reduction from CSP($\mathfrak{T}_{\tilde{\Phi}}$) to SAT($\tilde{\Phi}$), where $\tilde{\Phi}$ is an MMSNP τ-sentence with the existential signature σ, $\mathfrak{T}_{\tilde{\Phi}}$ is the relational ρ-structure constructed from $\tilde{\Phi}$. In order to reduce CSP($\mathfrak{T}_{\tilde{\Phi}}$) to SAT($\tilde{\Phi}$) one has to find, for every given input ρ-structure $\mathfrak{S}$, a τ-structure $\mathfrak{A}_\mathfrak{S}$ such that $\mathfrak{S} \to \mathfrak{T}_{\tilde{\Phi}}$ if and only if $\mathfrak{A}_\mathfrak{S} \models \tilde{\Phi}$. Suppose that we just take $\mathfrak{S}$ and replace each S_i-tuple with the corresponding τ-reduct $\mathfrak{F}_i^\tau$. Then the resulting τ-structure $\mathfrak{B}_\mathfrak{S}$ might have a substructure isomorphic to the hom-image of some τ-reduct $\mathfrak{F}_j^\tau$ and the S_j-tuple corresponding to this substructure is not in $\mathsf{S}_j^\mathfrak{S}$. See Figure 5.3 on page 110, for example.

In order to avoid this issue, Feder and Vardi assume that the given sentence $\tilde{\Phi}$ can be transformed to a logically equivalent sentence that has some special properties. These properties are known now as normal form properties. When $\tilde{\Phi}$ is in normal form, Feder and Vardi use the following adaptation for ρ-structures of the result of Erdős on graphs in [Erd59]. When $\tilde{\Phi}$ is in normal form, all the $\tilde{\tau}$-reducts corresponding to ρ-tuples are biconnected, and if the girth is larger than the maximal arity of a ρ-relation, then a large girth ρ-structure does not contain implicit ρ-tuples. In this case, they are able to replace every ρ-tuple with a $\tilde{\tau}$-structure in order to obtain an equivalent input $\tilde{\tau}$-structure of SAT($\tilde{\Phi}$).

Lemma 5.1.4 ([FV98])**.** *Let $\mathfrak{T}$ be a ρ-structure, and l be a natural number. Then, for every ρ-structure $\mathfrak{S}$ on n vertices there exists a ρ-structure $\mathfrak{S}'$ on n^a vertices such that:*

- *a depends only on l and $|T|$;*
- *there is a surjective homomorphism from $\mathfrak{S}'$ onto $\mathfrak{S}$;*
- *$\mathfrak{S}'$ maps to $\mathfrak{T}$ if and only if $\mathfrak{S}$ maps to $\mathfrak{T}$;*
- *the girth of $\mathfrak{S}'$ at least l.*



Sketch of Proof. The domain of $\mathfrak{S}'$ is the disjoint union $\biguplus_{s\in S} X_s$, where each X_s is a set of copies of s of size $N := n^{a-1}$. For any m-ary S-tuple $\mathbf{s} = (s_1, \ldots, s_n)$ of $\mathfrak{S}$, we impose the relation $\mathsf{S}^{\mathfrak{S}'}$ on a tuple $\mathbf{s}'$ in $X_{s_1} \times \cdots \times X_{s_m}$ with probability $N^{1-m+\epsilon}$. As $|X_{s_1} \times \cdots \times X_{s_m}| = N^m$, the expected number of tuples that are associated with $\mathbf{s}$ is $N^{1+\epsilon}$. If the arity m is equal to 1, then we impose S on all the elements of the corresponding set of copies. There is a surjective homomorphism $\pi\colon \mathfrak{S}' \twoheadrightarrow \mathfrak{S}$ such that $\pi(X_s) = s$, for any s in S. So, by transitivity, if $\mathfrak{S} \to \mathfrak{T}$, then $\mathfrak{S}' \to \mathfrak{T}$.

The expected number of cycles of length less than l is at most $N^{\epsilon'}$, where $\epsilon' = \frac{c_1}{a-1} + \epsilon l$ and c_1 is a constant value. Markov's inequality states that, for any $\delta > 0$,

$$\Pr(X \leq \delta) \geq 1 - \frac{\mathbb{E}[X]}{\delta}.$$

Set $\delta := 2\mathbb{E}[X]$, then the probability that $\mathfrak{S}'$ has at most twice as many such cycles is at least $\frac{1}{2}$.

Suppose that $\mathfrak{S}'$ homomorphically maps to $\mathfrak{T}$ by some $\mathfrak{h}'$, then, for each X_s, at least $\frac{|X_s|}{|T|}$ of its elements map to the same vertex of $\mathfrak{T}$, we denote them by Y_s and define a mapping $\mathfrak{h}\colon S \to T$ such that $\mathfrak{h}(s) := \mathfrak{h}'(Y_s)$. For a S-tuple $\mathbf{s} = (s_1, \ldots, s_m)$ of $\mathfrak{S}$, the expected number of tuples imposed on $Y_{s_1}, \ldots, Y_{s_m}$ is at least $\left(\frac{N}{|T|}\right)^m N^{1-m+\epsilon} = \frac{N^{1+\epsilon}}{|T|^m}$. The multiplicative Chernoff bound states that, for any $\delta > 0$,

$$\Pr\big(X < (1-\delta)\mathbb{E}[X]\big) < \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mathbb{E}[X]},$$

where X is the sum of independent random variables taking values in $\{0, 1\}$. Thus, the probability that the number of such tuples is at most half of the expected number is less than $p := \operatorname{Exp}\left[-\frac{N^{1+\epsilon}}{c_2}\right]$, where c_2 is a constant value.

For N large enough, we have $2N^{\epsilon'} < \frac{1}{2}\frac{N^{1+\epsilon}}{|T|^m}$. With probability at least $\frac{1}{2}$, we can get rid of all the cycles of length less than l by deleting $2N^{\epsilon'}$ tuples from $\mathfrak{S}'$. Denote by n^{c_3} the number of all the tuples $\mathbf{s}$ of $\mathfrak{S}$ that are contained in some relation. The probability that, after removing the tuples, there are no tuples imposed on some $Y_{s_1}, \ldots, Y_{s_m}$ is at most $1 - (1-p)^{n^{c_3}} < n^{c_3}p$. For N large enough, this probability is very small, so there exists a structure $\mathfrak{S}'$. In this case, $\mathfrak{h}$ is a homomorphism, as, for any S-tuple $\mathbf{s}$, $\mathfrak{h}(\mathbf{s}) = \mathfrak{h}'(Y_{s_1}, \ldots, Y_{s_m})$ and there is at least one S-tuple imposed on $Y_{s_1}, \ldots, Y_{s_m}$. □

Example 5.1.3. Have a look at Figure 5.4 on page 112. For simplicity, relations are binary. For every of three vertices of $\mathfrak{S}$, we add N vertices to $\mathfrak{S}'$ and then randomly add edges. There is no cycle of length 3, unlike in $\mathfrak{S}$. However, there are two cycles: red and blue. The red one has length 9, the blue one has the length 4. They have different nature: all edges of the blue cycle are in the preimage of just one edge of $\mathfrak{S}$, and this is not the case for the red cycle. △

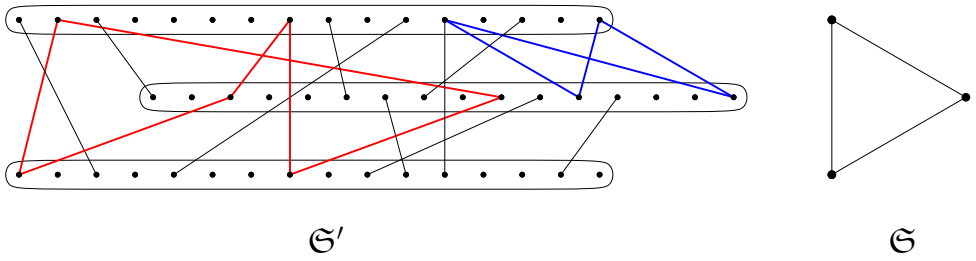


Figure 5.4: An illustration of the construction of the structure $\mathfrak{S}'$ from Lemma 5.1.4. Red and blue edges highlight cycles that appear in $\mathfrak{S}'$.

What can be done for MMSNP_2? If we use Lemma 5.1.4 from page 110 and then replace ρ-tuples with the corresponding $\tilde{\tau}$-structures, then we might obtain a structure with duplicated tuples. Such structure is not good, and it is not clear what τ-structure we can associate with it. In Section 5.2 we show how to transform any MMSNP_2 sentence to a form such that the input of $\mathrm{CSP}(\mathfrak{T}_{\tilde{\Phi}})$ can be restricted to good structures, without loss of generality. That is, to structures without implicit duplicated tuples. An example of a $\tilde{\tau}$-structure with duplicated tuples is displayed on Figure 5.2 on page 109. Apart from that, the new normal form for MMSNP_2 sentences, that we introduce in Section 5.2, allows two ρ-tuples intersect by an implicit $\tilde{\tau}$-tuple without having any implicit ρ-tuples. For example, two ρ-tuples corresponding to triangles could intersect by an edge as well. In the MMSNP case, they could only share a vertex.

On one hand, the large girth structure $\mathfrak{S}'$ must be good – this is more restrictive that the MMSNP case. However, on the other hand, the notion of "large girth" is less restrictive as we can now join two ρ-tuples by an edge, this gives us more tools to build the desired structure.

In Section 5.3, we discuss various approaches to build an appropriate structure $\mathfrak{S}'$ that could be associated with an input instance of $\mathrm{SAT}(\Phi)$. For each of these approaches, we explain why it is problematic to find such a structure.

5.2 Normal_2 form for MMSNP_2

We are going to show that, for MMSNP_2 sentences, the analogue of normal form for MMSNP is not sufficient for providing a reduction similar to the one of Feder and Vardi. We provide some new properties that are necessary for the potential existence of a similar proof for MMSNP_2.

At first, we recall that, by Proposition 4.2.2, for any MMSNP_2 sentence there exists a logically equivalent MMSNP_2 sentence in normal_1 form. We show now the insufficiency of the normal_1 form for the case of MMSNP_2 sentences.

Proposition 5.2.1. *Let Φ in MMSNP$_2$, and $\tilde{\Phi}$ in MMSNP be obtained from Φ by the transformation described in Section 4.1. Then, for some Φ in normal$_1$ form and for some l in $\mathbb{N}$, there is a ρ-structure $\mathfrak{G}$ such that $\mathfrak{G} \not\to \mathfrak{T}_{\tilde{\Phi}}$, and for any ρ-structure $\mathfrak{G}'$ of girth at least l, if $\mathfrak{G}' \to \mathfrak{G}$, then $\mathfrak{G}' \to \mathfrak{T}_{\tilde{\Phi}}$.*

Proof. In order to prove the statement, we are going to provide a counterexample. Let $\tau = \{\mathsf{R}(\cdot,\cdot)\}$, then $\tilde{\tau} = \{\mathsf{T}(\cdot), \mathsf{V}(\cdot), \tilde{\mathsf{R}}(\cdot,\cdot,\cdot)\}$. Suppose that there are four unary relation symbols in $\sigma\colon \mathsf{M}_a, \mathsf{M}_b, \mathsf{M}_c, \mathsf{M}_d$; and eight binary: $\mathsf{X}_{ab}, \mathsf{X}_{bc}, \mathsf{X}_{ca}, \mathsf{X}'_{ca}, \mathsf{X}_{ad}, \mathsf{X}_{dc}, \mathsf{X}_{bd}, \mathsf{X}_{db}$. The τ-reducts of the forbidden $(\tau \uplus \sigma)$-structures induced by Φ belong to one of the three following types: a cycle of length 3, a cycle of length 2, or a loop, they all are displayed on Figure 5.5 on page 113.

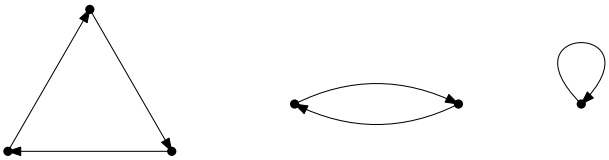


Figure 5.5: The τ-reducts of the forbidden structures corresponding to Φ.

Then, ρ consists of three relations: a binary S_2, a 4-ary S_4, and a 6-ary S_6. Let l denote the maximal arity among the ρ-relations. That is, in our case, $l = 6$. Suppose that

- Φ forbids loops for any possible colouring of their vertices and arcs.
- There is only one valid colouring of a 2-cycle:

$$\mathsf{R}(x,y) \wedge \mathsf{R}(y,x) \wedge \mathsf{M}_b(x) \wedge \mathsf{M}_d(y) \wedge \mathsf{X}_{bd}(x,y) \wedge \mathsf{X}_{db}(y,x).$$

- There are just two valid colourings of 3-cycles. They are associated with these conjunctions:

$$\mathsf{R}(x,y) \wedge \mathsf{R}(y,z) \wedge \mathsf{R}(z,x) \wedge \mathsf{M}_a(x) \wedge \mathsf{M}_b(y) \wedge \mathsf{M}_c(z) \wedge \mathsf{X}_{ab}(x,y) \wedge \mathsf{X}_{bc}(y,z) \wedge \mathsf{X}_{ca}(z,x),$$

$$\mathsf{R}(x,y) \wedge \mathsf{R}(y,z) \wedge \mathsf{R}(z,x) \wedge \mathsf{M}_a(x) \wedge \mathsf{M}_d(y) \wedge \mathsf{M}_c(z) \wedge \mathsf{X}_{ad}(x,y) \wedge \mathsf{X}_{dc}(y,z) \wedge \mathsf{X}'_{ca}(z,x).$$

The resulting ρ-structure $\mathfrak{T}_{\tilde{\Phi}}$ is displayed on the right side of Figure 5.6 on page 114.

Observe that we have chosen Φ such that it is in normal$_1$ form: we can require that the existential relations form a partition of both vertices and R-tuples, all the forbidden structures are biconnected, they are closed under homomorphisms, and we can make them be fully coloured. Our goal is to show that, for some relational ρ-structure $\mathfrak{S}$, any ρ-structure $\mathfrak{S}'$ of girth greater than l that maps to $\mathfrak{S}$ is either not equivalent to $\mathfrak{S}$ with respect to being mapped to $\mathfrak{T}_{\tilde{\Phi}}$ or there is no τ-structure that is associated with $\mathfrak{S}'$ if we try to replace any ρ-tuple with the corresponding τ-structure.

As a counterexample, we choose a structure $\mathfrak{S}$ as on the left side of Figure 5.6. Clearly, $\mathfrak{S} \not\to \mathfrak{T}_{\tilde{\Phi}}$.

Now, consider any ρ-structure $\mathfrak{S}'$ such that $\mathfrak{S}'$ has girth greater than $l = 6$ and that $\mathfrak{S}' \to \mathfrak{S}$. Any two relational tuples of $\mathfrak{S}'$ intersect by at most one point as, otherwise, there would be a cycle of length 2, that is smaller than l. If there are two tuples that intersect by a point that is associated with a T-vertex, then such $\mathfrak{S}'$ does not are associated

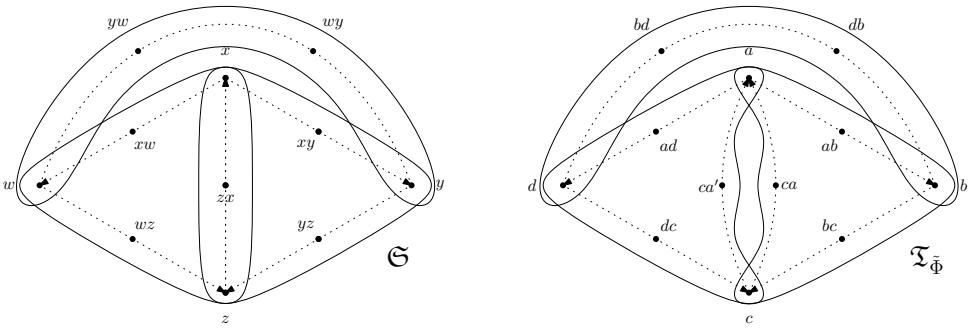


Figure 5.6: On the right, $\mathfrak{T}_{\tilde{\Phi}}$ consists of 12 vertices, two S_6-tuples, that are associated with 3-cycles and a S_4-tuple that is associated with the only allowed 2-cycle. Any vertex t in $T_{\tilde{\Phi}}$ is associated with some M_t in σ, the subscripts are highlighted. On the left, a ρ-structure $\mathfrak{S}$ is similar to $\mathfrak{T}_{\tilde{\Phi}}$, but in $\mathfrak{S}$ two S_6-tuples share a T-vertex zx, when in $\mathfrak{T}_{\tilde{\Phi}}$ there are two distinct ca, ca'. Implicit $\tilde{\mathsf{R}}$-tuples are highlighted with dotted arcs.

with any τ-structure, because, if two $\tilde{\mathsf{R}}$-tuples have a common T-vertex, then they must also share all the V-vertices. Suppose now that all the tuples of $\mathfrak{S}'$ intersect only at V-vertices.

We know that there exists a homomorphism $\mathsf{g}\colon \mathfrak{S}' \to \mathfrak{S}$. Let $\mathsf{f}\colon G \setminus \{zx\} \to T_{\tilde{\Phi}}$ be the obvious mapping, that is, $\mathsf{f}(w) = d, \mathsf{f}(x) = a, \mathsf{f}(y) = b, \mathsf{f}(z) = c$, and so on. We construct a homomorphism $\mathsf{h}\colon \mathfrak{S}' \to \mathfrak{T}_{\tilde{\Phi}}$ as follows. For any v in $\mathfrak{S}'$, if $\mathsf{g}(v) \neq zx$, then set $\mathsf{h}(v) = (\mathsf{f} \circ \mathsf{g})(v)$. If $\mathsf{g}(v) = zx$, then there is at most one S_6-tuple that contains v. This tuple must contain an element v' such that $\mathsf{g}(v')$ is in $\{w, y\}$. If $\mathsf{g}(v') = w$, then set $\mathsf{h}(v) = ca'$; if $\mathsf{g}(v') = y$, then set $\mathsf{h}(v) = ca$. This means that $\mathfrak{S}'$ is not equivalent to $\mathfrak{S}$ regarding having a homomorphism to $\mathfrak{T}_{\tilde{\Phi}}$. □

Definition 14. Let $\mathfrak{S}$ be a ρ-structure, $\mathbf{x} = (x_1, \dots, x_m)$ be a tuple of $\mathfrak{S}$ that belongs to some ρ-relation $\mathsf{S}^{\mathfrak{S}}$. Let $\mathfrak{G}$ be a $\tilde{\tau}$-structure of size m that is associated with S in ρ. For any $\tilde{\mathsf{R}}$-tuple $\mathbf{g} = (g_0, \dots, g_k)$ of $\mathfrak{G}$, a subtuple $\big(x_{\mathtt{ord}_{\mathfrak{G}}(g_0)}, \dots, x_{\mathtt{ord}_{\mathfrak{G}}(g_k)}\big)$ of $\mathbf{x}$ is called an *implicit* $\tilde{\mathsf{R}}$*-tuple* of $\mathfrak{S}$. If g is a T-(V-)vertex of $\mathfrak{G}$, then $x_{\mathtt{ord}_{\mathfrak{G}}(g)}$ is called an *implicit* T-(V-)*vertex*.

Example 5.2.1. Consider ρ-structures $\mathfrak{S}$ and $\mathfrak{T}_{\tilde{\Phi}}$ displayed on Figure 5.6 on page 114. All the dotted arcs stand for implicit $\tilde{\mathsf{R}}$-tuples. They have arity 3 as R is binary. △

Our aim is to modify Φ such that we may allow two relational ρ-tuples from $\mathfrak{S}'$ to share a T-vertex. When two ρ-tuples share a T-vertex t, they must also share all the V-vertices that make an implicit $\tilde{\mathsf{R}}$-tuple together with t. Otherwise, it would be impossible to find a ρ-structure $\mathfrak{S}'$ that is associated with a ρ-structure $\mathfrak{S}$, as in Proposition 5.2.1. Moreover, we want ρ-tuples to be able to share implicit $\tilde{\mathsf{R}}$-tuples so that nothing new appears once we replace ρ-tuples with the corresponding τ-reducts.

Definition 15. Let $\neg\phi$ be a negated conjunct of Φ and k in $\mathbb{N}$ be a natural number. $\neg\phi(\mathbf{x})$ is called *k-separable* if the variables $\mathbf{x}$ of ϕ can be grouped into three disjoint sets $\mathbf{s}, \mathbf{y}, \mathbf{z}$ of variables such that the tuple $\mathbf{s}$ contains k (possibly not pairwise distinct) variables and $\neg\phi(\mathbf{x})$ can be written in the form:

$$\neg\big(\psi_1(\mathbf{s}, \mathbf{y}) \wedge \psi_2(\mathbf{s}, \mathbf{z})\big),$$

where ψ_1 and ψ_2 are subformulae of ϕ obtained by splitting the conjunction ϕ into two parts. If $\neg\phi$ is not k-separable, then it is called *k-inseparable*.

Sometimes, within a k-separable conjunct of Φ, the tuple $\mathbf{s}$ is already contained in a τ-atom $\mathsf{R}(\mathbf{s})$. We want to track all such conjuncts and to get rid of them because they can be replaced by two smaller conjuncts which represent the separation parts. In order to do this, we give the following definition that is analogous to the "biconnected" property.

Definition 16. A negated conjunct $\neg\phi(\mathbf{x})$ of an MMSNP_2 τ-sentence Φ is called *tuple-biconnected* if, for every partition of its variables $\mathbf{x}$ into 3 groups $\mathbf{y}, \mathbf{s}, \mathbf{z}$, it cannot be written in the following form:

$$\neg\big(\psi_1(\mathbf{s}, \mathbf{y}) \wedge \mathsf{R}(\mathbf{s}) \wedge \psi_2(\mathbf{s}, \mathbf{z})\big),$$

where ψ_1, ψ_2 are subformulae of the conjunction ϕ that are obtained from it by removing the atom $\mathsf{R}(\mathbf{s})$ and splitting the rest into two parts.

Example 5.2.2. The canonical conjunctive query of the graph from Figure 5.8 on page 117 is not tuple-biconnected: if we remove the vertical arc together with both incident vertices, then the result will not be connected. On the other side, the canonical conjunctive query of the graph from Figure 5.9 on page 118 is tuple-biconnected. △

We can now give a better definition of normal form for MMSNP_2 sentences.

Definition 17. A τ-sentence Φ in MMSNP_2 is said to be in *normal$_2$ form* if the following conditions hold.

- Φ is in normal$_1$ form.
- Any negated conjunct of Φ is tuple-biconnected.
- For any negated conjunct $\neg\phi$ of Φ that is k-separable by a k-tuple $\mathbf{s}$ such that it can be written in the form:

  $$\neg\big(\psi_1(\mathbf{s}, \mathbf{x}) \wedge \psi_2(\mathbf{s}, \mathbf{y})\big),$$

  and, for any existential k-ary relation X in σ_T, Φ contains precisely one of the two following negated conjuncts:

  $$\neg\big(\psi_1(\mathbf{s}, \mathbf{x}) \wedge \mathsf{R}(\mathbf{s}) \wedge \mathsf{X}(\mathbf{s})\big),$$

  $$\neg\big(\psi_2(\mathbf{s}, \mathbf{y}) \wedge \mathsf{R}(\mathbf{s}) \wedge \mathsf{X}(\mathbf{s})\big).$$

- Any σ_T colour must uniquely define the σ_V colours of the tuple elements. That is, if there are two negated conjuncts

  $$\neg\big(\phi_1 \wedge \mathsf{R}(\mathbf{x}_1) \wedge \mathsf{X}(\mathbf{x}_1)\big) \text{ and } \neg\big(\phi_2 \wedge \mathsf{R}(\mathbf{x}_2) \wedge \mathsf{X}(\mathbf{x}_2)\big)$$

  that contain similarly coloured R-atoms, then, for any i in $[k]$ and for any M in σ_V, ϕ_1 contains the atom $\mathsf{M}(x_1^i)$ if and only if ϕ_2 contains the atom $\mathsf{M}(x_2^i)$. Here, x_1^i and x_2^i are the elements at the ith coordinates of $\mathbf{x}_1$ and $\mathbf{x}_2$ correspondingly.

Proposition 5.2.2. *For any connected τ-sentence Φ in MMSNP_2 there is a connected τ-sentence Ψ in MMSNP_2 such that Ψ is in normal$_2$ form and logically equivalent to Φ.*

Proof. In order to satisfy the fourth condition of normal$_2$ form, we replace any σ_{T}-relation X with $|\sigma_{\mathsf{V}}|^k$ relations of the set

$$\big\{\mathsf{X}_{(\mathsf{M}_1,\dots,\mathsf{M}_k)} \mid i \in [k],\ \mathsf{M}_i \in \sigma_{\mathsf{V}}\big\}.$$

Within every negated conjunct, we replace an atom $\mathsf{X}(\mathbf{x})$ with $\mathsf{X}_{(\mathsf{M}_1,\dots,\mathsf{M}_k)}(\mathbf{x})$ if the conjunct also contains an atom $\mathsf{M}_i(x_i)$, for any i in $[k]$.

Next, we transform Φ to a logically equivalent sentence in normal$_1$ form, this is provided by Proposition 4.2.2. Denote the mapping that transforms a sentence to normal$_1$ form by $\mathfrak{nf}_1 \colon \mathsf{MMSNP}_2 \to \mathsf{MMSNP}_2$. Since now, we can assume that Φ is in normal$_1$ form, as we can set $\Phi := \mathfrak{nf}_1(\Phi)$. Notice that $\mathfrak{nf}_1$ preserves the fourth condition of the normal$_2$ form property.

It remains to achieve the second and the third conditions. Let $\neg\phi_1, \dots, \neg\phi_s$ be negated conjuncts of Φ such that they are k-separable, where k is the arity of R in τ, and that they have the maximal number of variables, denoted by N, and that they violate the third condition of the normal$_2$ form property. Some of them also are not tuple-biconnected. Denote the current sentence Φ by Φ_N. Below, we transform Φ_N to an equivalent sentence in normal$_1$ form, where any k-separable negated conjunct violating the third condition has strictly less than N variables. Hence, we can obtain, in finitely many steps, an equivalent sentence $\Psi := \Phi_0$ that is in normal$_2$ form.

We now describe the procedure that is applied to each of these negated conjuncts until all the non-tuple-biconnected conjuncts are removed and all of the rest satisfy the third condition. Pick any k-separable negated conjunct $\neg\phi_i$ from $\neg\phi_1, \dots, \neg\phi_s$. For any k-tuple $\mathbf{s}$ that separates $\neg\phi_i$ as follows:

$$\neg\big(\psi_1(\mathbf{s}, \mathbf{x}) \wedge \psi_2(\mathbf{s}, \mathbf{y})\big),$$

we add to Φ_N the two following negated conjuncts (each of them contains strictly less than N variables):

$$\neg\big(\psi_1(\mathbf{s}, \mathbf{x}) \wedge \mathsf{R}(\mathbf{s}) \wedge \mathsf{P}(\mathbf{s})\big) \wedge \neg\big(\psi_2(\mathbf{s}, \mathbf{y}) \wedge \mathsf{R}(\mathbf{s}) \wedge \neg\mathsf{P}(\mathbf{s})\big).$$

Here, P is a new k-ary existential relation that we have added at this step. Adding these conjuncts keeps the sentence logically equivalent to the original one. Because if any of these two new negated conjuncts is violated, then either the original negated conjunct $\neg\phi_i$ is also violated or we can choose another interpretation for the relation symbol P and satisfy both negated conjuncts. Also, if $\neg\phi_i$ is connected, then both new negated conjuncts are connected.

If $\neg\phi_i$ is not tuple-biconnected with respect to $\mathbf{s}$, that is, if it can be written as $\neg\big(\psi_1^i(\mathbf{x}_i, \mathbf{s}) \wedge \mathsf{R}(\mathbf{s}) \wedge \psi_2^i(\mathbf{s}, \mathbf{y}_i)\big)$, then we remove it from Φ.

We do this procedure for every k-separating tuple of every negated conjunct $\phi_1, \dots, \phi_s$. Observe that we have extended the existential signature with new relation symbols, *e.g.* P. The number of them is bounded by some constant depending on Φ_N. The fourth normal$_2$ condition is satisfied as, for any new relation symbol P there is a unique non-negated P-atom.

Apply now the transformation $\mathfrak{nf}_1$ to this sentence. Denote the result by $\Phi_{N'}$. This sentence is in normal$_1$ form, it is logically equivalent to Φ, and any negated conjunct that violates the second or the third normal$_2$ form property contains at most N' variables, where N' is strictly smaller than N. We repeat this procedure for $\Phi_{N'}$, and in at most N'

iterations we obtain $\Psi := \Phi_0$ such that any negated conjunct satisfies the second and the third normal$_2$ form conditions.

□

Example 5.2.3. Let $\neg\phi_i$ are associated with a 4-cycle, where the relation R is binary. Then $\neg\phi_i$ is 2-separable, however it is tuple-biconnected. It can be separated by four different tuples that create eight new graphs, however, each of them is 2-inseparable. See Figure 5.7 on page 117. △

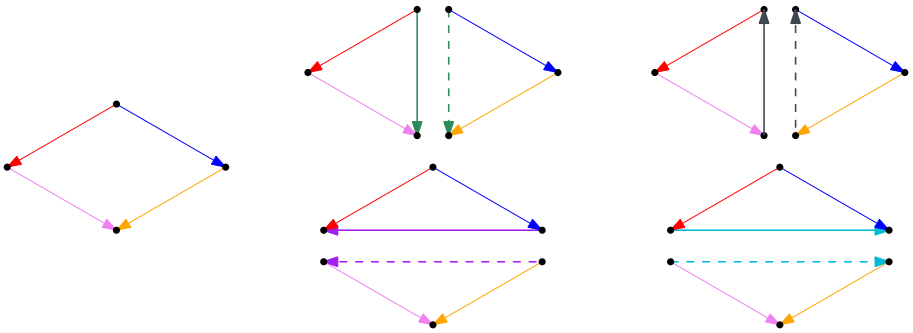


Figure 5.7: $\mathfrak{F}$, corresponding to ϕ_i, is on the left. On the right, the eight structures obtained from $\mathfrak{F}$ by separating it in two in the all possible ways. The corresponding eight negated conjuncts are added to the sentence when we transform it to normal$_2$ form. Dashed arcs mean negated existential atomic formulae.

In the rest of this section we discuss the new opportunities that we have when Φ is in normal$_2$ form. Consider Figure 5.8 on page 117.

Suppose that some of the negated conjuncts of Φ encode triangles (*e.g.*, C_3). Then the relational signature ρ from Proposition 5.2.1 would contain a 6-ary relation symbol △ corresponding to the triangle. In the proof of Proposition 5.2.1, the ρ-structure $\mathfrak{S}'$ cannot contain two △-tuples that share an arc, as on Figure 5.8. Because, in this case, the girth is at most 2. But, if the sentence Φ is in normal$_2$ form, then two tuples are allowed to share a whole arc. If some negated conjunct of Φ encodes a structure from Figure 5.9 on page 118. then ρ must contain a 8-ary relation symbol □ corresponding to this structure. Observe that on Figure 5.8 there is a □-tuple implicitly contained in the union of two triangles. But, by the condition of normal$_2$ form, it is forbidden to assign colours to these two △-tuples such that the implicit □-tuple has a forbidden colouring.

We can conclude that it is not necessary anymore for $\mathfrak{S}'$ to have large *girth*. We need to define a new metric that is more appropriate in our case. Recall that the relation symbols of ρ are associated with $\tilde{\tau}$-reducts of forbidden structures.

Definition 18. Let $\mathfrak{S}$ be a ρ-structure and $\mathbf{x}, \mathbf{y}$ be two relational ρ-tuples. We say that $\mathbf{x}$ and $\mathbf{y}$ are V-*adjacent* if there is an implicit V-vertex z that is contained in both of the

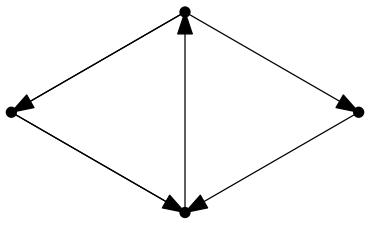


Figure 5.8: Two triangles share an arc.

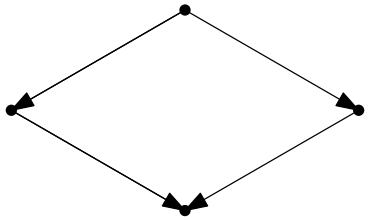


Figure 5.9: An implicit $\square$-tuple contained on Figure 5.8.

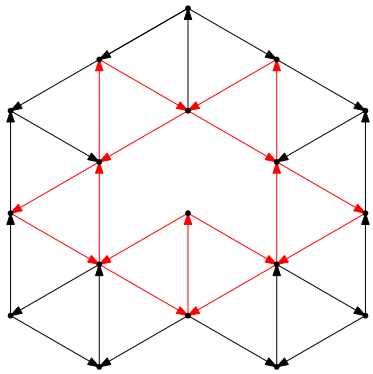


Figure 5.10: An example of a ρ-structure. The red ρ-tuples (triangles) highlight a tuple-cycle of the minimal length.

tuples, and if no other vertex is shared by the tuples. We say that $\mathbf{x}$ and $\mathbf{y}$ are T-*adjacent* if there is an implicit $\tilde{\mathsf{R}}$-tuple $\mathbf{z}$ that is an implicit $\tilde{\mathsf{R}}$-tuple of both ρ-tuples $\mathbf{x}$ and $\mathbf{y}$, and if $\mathbf{x}$ and $\mathbf{y}$ do not share any other vertices.

A sequence of ρ-tuples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is called a *tuple-cycle* if, for i in $[n-1]$, $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$ are either V-adjacent or T-adjacent, and similarly for $\mathbf{x}_n$ and $\mathbf{x}_1$. The *length* of a tuple-cycle is simply the number of tuples in the sequence. A structure $\mathfrak{S}$ is said to have *tuple-girth* be equal to l, for l in $\mathbb{N}$, if $\mathfrak{S}$ has a tuple-cycle of length l and if it has no tuple-cycle of length less than l.

Example 5.2.4. Consider Figure 5.10 on page 118. It is a ρ-structure that contains $\triangle$-tuples. Some of the tuples are T-adjacent, some are V-adjacent. The original girth of this structure is equal to 2. The tuple-girth is equal to 6. A cycle of the minimal length is highlighted with the red triangles. $\triangle$

At last, we explain the fourth condition of the normal$_2$ form property. This condition requires that a σ_{T}-relation X uniquely determines the σ_{V}-relations that are assigned to elements of a tuple from X.

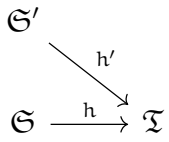
Definition 19. We say that implicit $\tilde{\mathsf{R}}$-tuples $(t, \mathbf{v}_1)$ and $(t, \mathbf{v}_2)$ of a ρ-structure $\mathfrak{A}$ are *duplicated* if $\mathbf{v}_1 \neq \mathbf{v}_2$. Two implicit tuples $(t_1, \mathbf{v})$ and $(t_2, \mathbf{v})$ of $\mathfrak{A}$ are called *multiple* if $t_1 \neq t_2$.

Observation 1. Let Φ be an MMSNP_2 sentence in normal$_2$ form. Let $\mathfrak{T}_{\tilde{\Phi}}$ be the corresponding finite CSP ρ-structure. Then $\mathfrak{T}_{\tilde{\Phi}}$ does not contain implicit duplicated tuples. Though, $\mathfrak{T}_{\tilde{\Phi}}$ still may contain multiple implicit tuples which makes our situation more difficult.

5.3 Expanders

Further, we consider the following problem. Being able to solve it is sufficient to prove the dichotomy property for MMSNP_2.

Problem 1. Let $\mathfrak{T}$ be a ρ-structure without duplicated implicit tuples. Then, for any ρ-structure $\mathfrak{S}$ and for any l in $\mathbb{N}$, one can construct in time polynomial in $|S|$ a ρ-structure $\mathfrak{S}'$ of tuple-girth at least l and without duplicated implicit $\tilde{\mathsf{R}}$-tuples such that there is a homomorphism $\mathfrak{h}\colon \mathfrak{S} \to \mathfrak{T}$ if and only if there is a homomorphism $\mathfrak{h}'\colon \mathfrak{S}' \to \mathfrak{T}$.



Remark. The name of this section contains the word *expander* because of the ε-expander property that is very useful in our situation. This property is given by Kun in [Kun13], where he proves that the structure $\mathfrak{S}'$ of Feder and Vardi's Lemma 5.1.4 can be constructed by a deterministic P-time algorithm. Roughly speaking, this property estimates how closely to a uniform distribution the relational tuples are distributed. More precisely, let $\mathfrak{A}$ be a ρ-structure and let $\varepsilon > 0$. Then $\mathfrak{A}$ is called an *ε-expander* if, for any S in ρ of arity m and for any subsets $A_1, \dots, A_m$ of A, we have

$$\left| |\mathsf{S}(A_1, \dots, A_m)| - \frac{|A_1| \times \dots \times |A_m|}{|A|^m} |\mathsf{S}^{\mathfrak{A}}| \right| < \varepsilon |\mathsf{S}^{\mathfrak{A}}|,$$

where $\mathsf{S}(A_1, \dots, A_m)$ denotes the set of S-tuples imposed on the subsets $A_1, \dots, A_m$. This formula states that the difference between the actual number of tuples imposed on the subsets differs from their expected number under the uniform distribution by a small proportion of the total number of tuples.

Further in this section we consider different approaches that can be used in order to solve the problem. We explain every time why the approach can not be used to solve Problem 1.

V-vertex expander

Let ρ_{V} be a signature obtained from ρ such that, for every S in ρ of arity $(m_{\mathsf{T}}, m_{\mathsf{V}})$ there is a relation symbol S_{V} in ρ_{V} of arity m_{V}. For any ρ-structure $\mathfrak{A}$ with the domain $A = A_{\mathsf{T}} \uplus A_{\mathsf{V}}$, denote by $\mathfrak{A}_{\mathsf{V}}$ a ρ_{V}-structure with the domain A_{V} such that, for any S_{V} in ρ_{V} and for any $\mathbf{v}$ in $A_{\mathsf{V}}^{m_{\mathsf{V}}}$, the tuple $\mathbf{v}$ belongs to $\mathsf{S}_{\mathsf{V}}^{\mathfrak{A}_{\mathsf{V}}}$ if and only if $(\mathbf{t}, \mathbf{v}) \in \mathsf{S}^{\mathfrak{A}}$ for some $\mathbf{t} \in A_{\mathsf{T}}^{m_{\mathsf{T}}}$.

Now we explain how to obtain the structure $\mathfrak{S}'$ for a given input structure $\mathfrak{S}$ and target structure $\mathfrak{T}$. After that, we discuss if it is a good candidate for Problem 1.

Construction 15. For two given ρ-structures $\mathfrak{S}$ and $\mathfrak{T}$, we consider the corresponding ρ_{V}-structures $\mathfrak{S}_{\mathsf{V}}$ and $\mathfrak{T}_{\mathsf{V}}$. In this case, we can find a ρ_{V}-structure $\mathfrak{S}'_{\mathsf{V}}$ of large girth that is equivalent to $\mathfrak{S}_{\mathsf{V}}$ with respect to $\mathrm{CSP}(\mathfrak{T}_{\mathsf{V}})$, by Feder and Vardi's Lemma 5.1.4. And then we return to the signature ρ by adding T-vertices: for every S_{V}-tuple $\mathbf{v}$ of every S_{V} in ρ_{V}, we introduce a tuple $\mathbf{t}$ of pairwise distinct new T-vertices and require that $(\mathbf{t}, \mathbf{v})$ belongs to $\mathsf{S}^{\mathfrak{S}'}$. This finishes Construction 15.

This approach does not work because $\mathfrak{T}$ may contain multiple tuples.

Counterexample 5.3.1. Let $\mathfrak{S}$ and $\mathfrak{T}$ be the same as $\mathfrak{S}$ and $\mathfrak{T}_{\tilde{\Phi}}$ in Proposition 5.2.1 from Section 5.2, see Figure 5.11 on page 120. Then, $\mathfrak{S}$ does not map to $\mathfrak{T}$, but any $\mathfrak{S}'$ obtained from $\mathfrak{S}$ by Construction 15 maps to $\mathfrak{T}$.

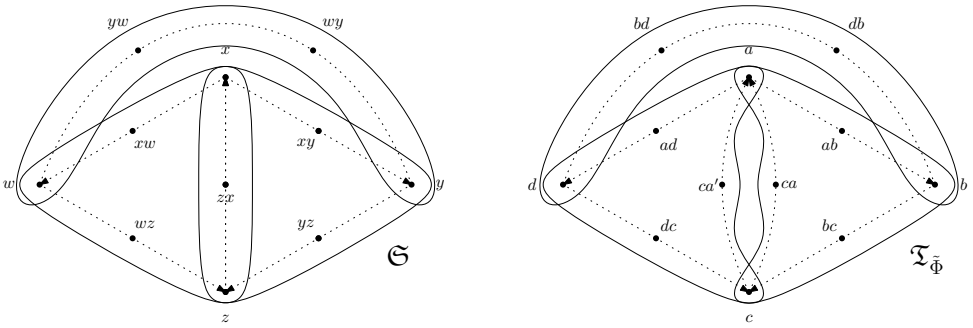


Figure 5.11: A counterexample for the V-vertex expander strategy.

T-vertex expander

Let ρ_{T} be a signature obtained from ρ in a similar way as ρ_{V} is obtained for V-vertex expanders. That is, for every $\mathsf{S} \in \rho$ of arity $k = m_{\mathsf{T}} + m_{\mathsf{V}}$, we introduce a relation symbol $\mathsf{S}_{\mathsf{T}} \in \rho_{\mathsf{T}}$ of arity m_{T}. And for any ρ-structure $\mathfrak{A}$ with the domain $A_{\mathsf{T}} \uplus A_{\mathsf{V}}$, we denote by $\mathfrak{A}_{\mathsf{T}}$ a ρ_{T}-structure with the domain A_{T} such that, for any $\mathsf{S}_{\mathsf{T}} \in \rho_{\mathsf{T}}$ and any $\mathbf{t} \in A_{\mathsf{T}}^{m_{\mathsf{T}}}$, the tuple $\mathbf{t}$ belongs to the relation $\mathsf{S}_{\mathsf{T}}^{\mathfrak{A}_{\mathsf{T}}}$ if and only if $(\mathbf{t}, \mathbf{v})$ is contained in $\mathsf{S}^{\mathfrak{A}}$ for some tuple $\mathbf{v} \in A_{\mathsf{V}}^{m_{\mathsf{V}}}$.

Construction 16. One can treat the structure $\mathfrak{A}_{\mathsf{T}}$ similarly to the structure $\mathfrak{A}_{\mathsf{V}}$. For given ρ-structures $\mathfrak{S}, \mathfrak{T}$ we consider ρ_{T}-structures $\mathfrak{S}_{\mathsf{T}}, \mathfrak{T}_{\mathsf{T}}$ and find a structure $\mathfrak{S}'_{\mathsf{T}}$ such that it satisfies the conditions of Feder and Vardi's Lemma 5.1.4 over the signature ρ_{T}. Then, we add V-vertices to $\mathfrak{S}'_{\mathsf{T}}$ in order to obtain the desired ρ-structure $\mathfrak{S}'$. At first, we add to $\mathfrak{S}'$ all the vertices of $\mathfrak{S}'_{\mathsf{T}}$, they make the set of implicit T-vertices. Let $\mathbf{t}$ be a S_{T}-tuple of $\mathfrak{S}'_{\mathsf{T}}$, for some S_{T} in ρ_{T}. Suppose that the arity of S_{T} is m_{T}, and the arity of the corresponding ρ-relation S is $(m_{\mathsf{T}}, m_{\mathsf{V}})$. For $\mathbf{t}$, we add m_{V} new pairwise distinct vertices $\mathbf{v}$, they become implicit V-vertices, and we add the tuple $(\mathbf{t}, \mathbf{v})$ to the relation $\mathsf{S}^{\mathfrak{S}'}$. Do this procedure for any tuple $\mathbf{t}$ of any ρ_{T}-relation of $\mathfrak{S}'_{\mathsf{T}}$. If an implicit T-vertex t is contained in more than one ρ-tuple, then it is contained in a pair of duplicated implicit $\tilde{\mathsf{R}}$-tuples. For any such pair $(t, \mathbf{v}_1)$ and $(t, \mathbf{v}_2)$, we identify $\mathbf{v}_1$ and $\mathbf{v}_2$ in the coordinate-wise fashion.

Counterexample 5.3.2. Suppose that $\mathfrak{S}$ contains two relational tuples $(\mathbf{t}_1, \mathbf{v}_1), (\mathbf{t}_2, \mathbf{v}_2)$ such that $\mathbf{t}_1 \cap \mathbf{t}_2 = \varnothing$ and $\mathbf{v}_1 \cap \mathbf{v}_2 \neq \varnothing$. Then, during the construction of $\mathfrak{S}_{\mathsf{T}}$, the corresponding tuples share no elements. Thus, none of them intersect in $\mathfrak{S}'$ neither. We can consider a case where any two tuples of $\mathfrak{S}$ can only share some V-vertices but not T-vertices. Then, all $\mathfrak{S}_{\mathsf{T}}, \mathfrak{S}'_{\mathsf{T}}$ and $\mathfrak{S}'$ consist of pairwise disjoint tuples, for any choice of $\mathfrak{T}_{\tilde{\Phi}}$. This means that, for some $\mathfrak{T}_{\tilde{\Phi}}$, $\mathfrak{S}$ and $\mathfrak{S}'$ are not equivalent with respect to having a homomorphism to $\mathfrak{T}_{\tilde{\Phi}}$.

TV-expanders

In this part, we construct the structure $\mathfrak{S}'$ directly from $\mathfrak{S}$ unlike V-vertex expanders and T-vertex expanders. We use the same approach as in Feder and Vardi's Lemma 5.1.4: we

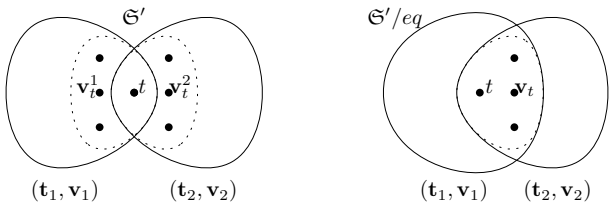


Figure 5.12: To the left, there are duplicated implicit tuples in $\mathfrak{S}'$. Tuples $(\mathbf{t}_1, \mathbf{v}_1)$ and $(\mathbf{t}_2, \mathbf{v}_2)$ share one T-vertex t but not the V-vertices of the corresponding implicit tuples $\mathbf{v}_\mathsf{T}^1, \mathbf{v}_\mathsf{T}^2$. To the right, there is the result of their identification within $\mathfrak{S}'/\mathsf{eq}$ that does not have duplicated implicit tuples.

substitute every vertex s of the domain S of $\mathfrak{S}$ by a set X_s of vertices and then, for any tuple $(t_1, \dots, t_{m_\mathsf{T}}, v_1, \dots, v_{m_\mathsf{V}})$ of $\mathfrak{S}$, we randomly impose tuples on the sets $X_{t_1}, \dots, X_{v_{m_\mathsf{V}}}$.

The resulting structure $\mathfrak{S}'$ has large tuple-girth, as the tuple-girth is always at least as large as the girth. But $\mathfrak{S}'$ does not look exactly as what we want. Particularly, it has duplicated implicit $\tilde{\mathsf{R}}$-tuples, *i.e.*, it contains tuples that have some T-vertex t in common but that do not share the V-vertices $\mathbf{v}_\mathsf{T}$ of the corresponding implicit tuple $(t, \mathbf{v}_\mathsf{T})$, see the left part of Figure 5.12 on page 121. This is a problem because the ρ-structure $\mathfrak{S}'$ should be associated with some $\tilde{\tau}$-structure. If the target structure $\mathfrak{T}_{\tilde{\Phi}}$ does not have duplicated implicit $\tilde{\mathsf{R}}$-tuples, then, for any structure $\mathfrak{S}'$, one can construct in time polynomial in the size of $\mathfrak{S}'$ a ρ-structure $\mathfrak{S}'/\mathsf{eq}$ that is equivalent to $\mathfrak{S}'$ with respect to $\mathrm{CSP}(\mathfrak{T}_{\tilde{\Phi}})$ and that does not have duplicated implicit $\tilde{\mathsf{R}}$-tuples. Here, eq is the minimal by inclusion equivalence relation over S' such that two V-vertices v_1, v_2 are equivalent if there is a T-vertex t and two implicit $\tilde{\mathsf{R}}$-tuples $(t, \mathbf{v}_1), (t, \mathbf{v}_2)$, where v_1 and v_2 are on the same coordinate. The quotient does not have duplicated implicit $\tilde{\mathsf{R}}$-tuples.

On one hand, the structure $\mathfrak{S}'/\mathsf{eq}$ does not have duplicated tuples. But, on the other hand, we still can not derive a $\tilde{\tau}$-structure from it because taking a quotient might make the tuple-girth smaller. In the rest of this section we estimate how many new cycles of short length appear after taking the quotient.

For simplicity, we consider a particular case when $\tilde{\tau} = \{\mathsf{E}(\cdot, \cdot)\}$ and $\rho = \{\mathsf{S}(\cdot, \cdot, \cdot, \cdot, \cdot, \cdot)\}$. Here, S is a $(3, 3)$-ary relation that represents a triangle. We are going to consider a cycle of length n in a ρ-structure $\mathfrak{S}'$ and calculate the probability that, after taking the quotient $\mathfrak{S}'/\mathsf{eq}$, this cycle contains a subcycle of length at most l, for some fixed l. Let the cycle consist of tuples $(\mathbf{t}_1, \mathbf{v}_1), \dots, (\mathbf{t}_n, \mathbf{v}_n)$ such that, for each two neighbour tuples $(\mathbf{t}_i, \mathbf{v}_i), (\mathbf{t}_{i+1}, \mathbf{v}_{i+1})$ there is a unique vertex x that is contained either in both $\mathbf{t}_i, \mathbf{t}_{i+1}$ or in both $\mathbf{v}_i, \mathbf{v}_{i+1}$. This means that tuples intersect correctly with respect to the partition into implicit T- and V-vertices: a vertex cannot be simultaneously an implicit T-vertex and an implicit V-vertex. We also make an assumption that the way how two neighbour tuples intersect is uniformly arbitrary: say that in a tuple $(\mathbf{t}_i, \mathbf{v}_i)$ we randomly choose a vertex (with the same probability for each vertex) and if it appears to be a T-vertex, then we also randomly choose a T-vertex in $(\mathbf{t}_{i+1}, \mathbf{v}_{i+1})$ and identify these two.

It is important how exactly neighbour tuples touch. Suppose that a tuple $(\mathbf{t}_i, \mathbf{v}_i) = (t_a, t_b, t_c, a, b, c)$ of the cycle touches the precedent tuple $(\mathbf{t}_{i-1}, \mathbf{v}_{i-1})$ by a T-vertex, say, by t_a. Then there are 5 different cases, see Figure 5.14 on page 123, how the next tuple $(\mathbf{t}_{i+1}, \mathbf{v}_{i+1})$ can touch $(\mathbf{t}_i, \mathbf{v}_i)$:

- 'tDv': by the V-vertex a, in this case, taking the quotient does not reduce the cycle length;
- 'tLv': by the V-vertex b, in this case, the $(i+1)$th tuple touches the $(i-1)$th tuple

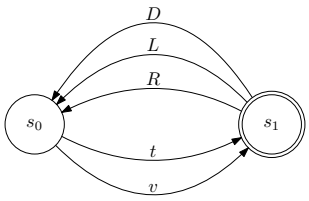


Figure 5.13: A finite automaton that accepts a word if and only if it is associated with a sequence of tuples. The starting state is s_0, the accepting state is s_1.

after taking the quotient;

- 'tRv': by the V-vertex c, in this case, the $(i+1)$th tuple touches the $(i-1)$th tuple after taking the quotient;
- 'tLt': by the T-vertex t_c, in this case, the $(i+1)$th tuple touches the $(i-1)$th tuple after taking the quotient;
- 'tRt': by the T-vertex t_b, in this case, the $(i+1)$th tuple touches the $(i-1)$th tuple after taking the quotient.

Sequences of tuples can be encoded by a regular language $\mathcal{L} \subset \{t, v, D, L, R\}^*$. A finite automaton displayed on Figure 5.13 on page 122 defines this language. A string of symbols of $\{t, v, D, L, R\}$ belongs to $\mathcal{L}$ if and only if any odd symbol is either t or v, and any even symbol is either D, L, or R.

Suppose that a tuple $(\mathbf{t}_i, \mathbf{v}_i) = (t_a, t_b, t_c, a, b, c)$ touches the precedent tuple $(\mathbf{t}_{i-1}, \mathbf{v}_{i-1})$ by a V-vertex, say, by a. Then there are 3 different cases, see Figure 5.15 on page 123, how the next tuple $(\mathbf{t}_{i-1}, \mathbf{v}_{i-1})$ may touch $(\mathbf{t}_i, \mathbf{v}_i)$:

- 'vDv' or 'vDt': by the V-vertex b or c or by the T-vertex t_a correspondingly, in each of these cases, the cycle length is not reduced by taking the quotient;
- 'vLt': by the T-vertex t_c, or
- 'vRt': by the T-vertex t_b, in both of these cases, the $(i+1)$th tuple touches the $(i-1)$th tuple.

Every case has its own label. The first lower-case letter explains how the $(i-1)$th tuple touches the ith one: either by a T-vertex (t) or by a V-vertex (v). The upper-case letter describes which side of the ith triangle touches the $(i+1)$th one: left (L), right (R), or the distant side (D). The last lower-case letter precises if the ith and $(i+1)$th tuples touch either by a T-vertex (t) or by a V-vertex (v).

After giving a label to each triangle, depending on how it touches the precedent and the next triangles, we can represent any such cycle by a word. And we can characterize how each combination of labels reduces the length. Consider an example given on Figure 5.16 on page 123.

Observe that for a sequence of tuples there is a unique word, but for the same word there may be more than one different sequence of tuples. However, they all have the same length even after taking the quotient. We are going to consider all possible cases how the addition of the next tuple changes the length of the sequence in the quotient. And to provide an upper bound for the probability that, after increasing the sequence by one more tuple, the length of the sequence increases once we identify all the duplicated implicit tuples.

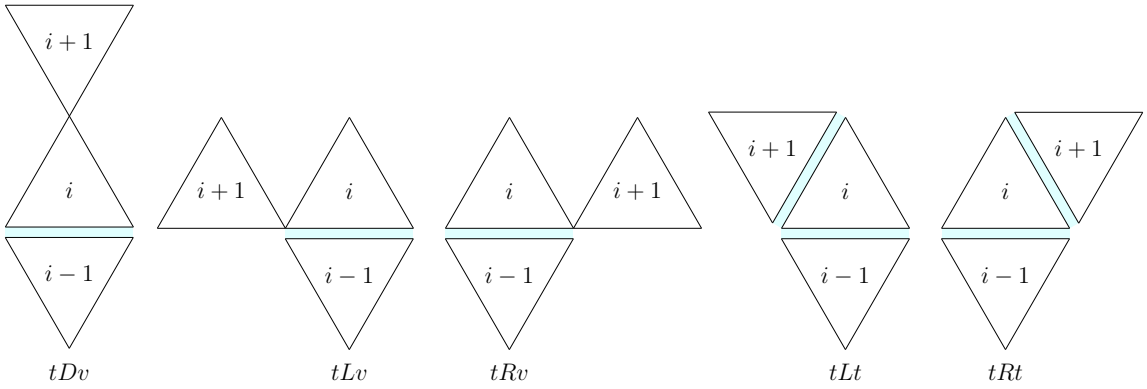


Figure 5.14: Five cases how the $(i+1)$the tuple can touch the ith tuple if the $(i-1)$th and the ith tuples intersect by a T-vertex. Light blue zones between two implicit edges mean that they are duplicated.

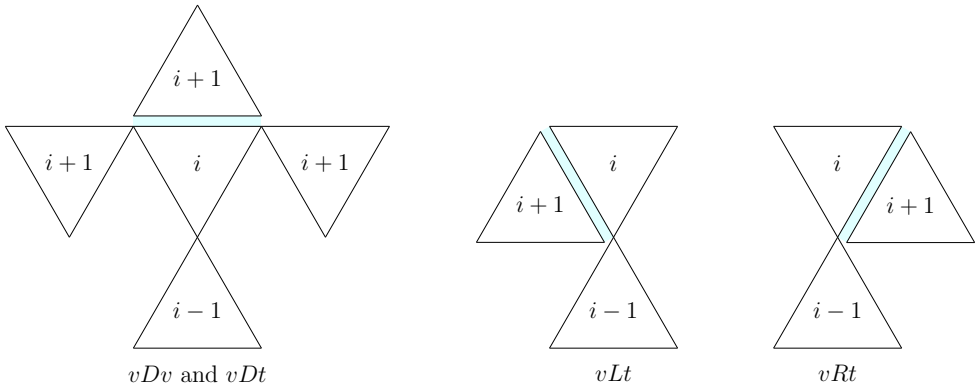


Figure 5.15: Three cases how the $(i+1)$the tuple can touch the ith tuple if the $(i-1)$th and the ith tuples intersect by a V-vertex. Light blue zones between two implicit edges mean that they are duplicated.

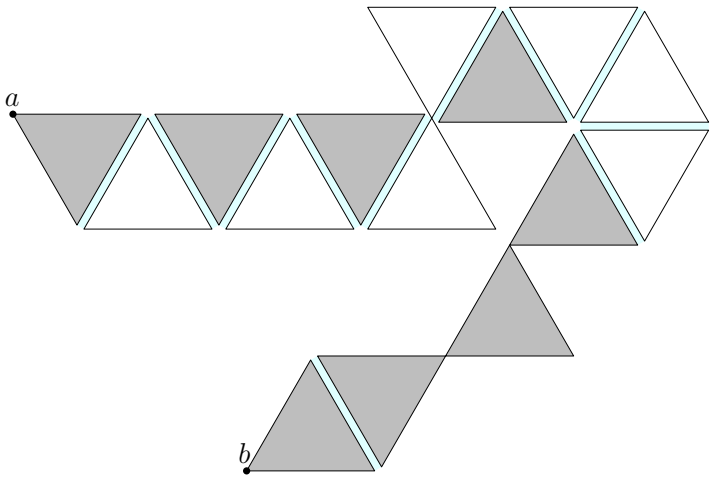


Figure 5.16: A sequence of S-tuples corresponding to the word $tLtRtLtRtLvRtLtRtRtRtDvDvDt$. Grey triangles make the shortest path between a and b after taking the quotient. Light blue zones between two implicit edges mean that they are duplicated.

- If a word ends with tLt, then adding Lt or Lv does not increase the length because the last two tuples intersect the same tuples. If we assume that adding each label has equal probability $\frac{1}{5}$, then the black triangle is added with probability at most $\frac{3}{5}$. The case for tRt is similar.
- If a word ends with tLv, then adding Lt or Rt does not increase the length because the newly added tuple intersects the same tuples as the previous one. Thus, the length increases with probability at most $\frac{3}{5}$. The case for tRv is similar.
- If a word ends with vLt, then adding Lt or Lv does not add a new black triangle. The probability that the length increases is at most $\frac{3}{5}$. The case for vRt is similar.
- If a word ends with tDv or with vDv, then adding Lt or Rt does not add a black triangle. Thus, the probability is again at most $\frac{3}{5}$.
- If a word ends with vDt then adding anything but Dv does not add a black triangle, thus the probability is at most $\frac{1}{5}$ so we can assume that it is at most $\frac{3}{5}$.

Let $p \leq \frac{3}{5}$. Let $X = \sum_{i=0}^{k} X_i$, where k denotes the number of tuples in the sequence, and X_is are independent random values that take values in $\{0, 1\}$ such that, for all i in $[k]$, we have $\Pr(X_i = 1) = p$. Below, we write the probability that a sequence, having length k before taking the quotient, has length at most l after taking the quotient:

$$\Pr(X < l) = \sum_{i=0}^{l} \binom{n}{l} p^i (1-p)^{k-i}.$$

The smaller the value of p, the greater this probability becomes. So this probability is minimal when $p = \frac{3}{5}$. This implies the following lower bound:

$$\Pr(X < l) > \sum_{i=0}^{l} \binom{n}{l} \left(\frac{3}{5}\right)^i \left(\frac{2}{5}\right)^{k-i} > \left(\frac{2}{5}\right)^k.$$

Consider a random ρ-structure $\mathfrak{S}'$, as in the proof of Lemma 5.1.4, with the domain of size n, where, as every relation has 6 coordinates, the probability that for given 3 T-vertices and 3 V-vertices there is a tuple on them is $\frac{n^\varepsilon}{n^5}$, where $\varepsilon > 0$ is fixed. We want to compute the expected number of cycles on k vertices, for some natural number k. We use the following information in order to do this:

- $\binom{n}{k}$ – the number of ways to choose k vertices that make the cycle;
- $\frac{k!}{2k}$ – the number of different cycles consisting of k elements;
- $(n-k)\dots(n-5k+1)$ – the number of ways to choose the vertices that belong to some tuple but do not participate in the cycle;
- $\left(\frac{n^{1+\varepsilon}}{n^6}\right)^k$ – the probability that every of the k tuples of the cycle is present.

The expected number of cycles of length k is the product of these four values.

$$\binom{n}{k} \frac{k!}{2k} (n-k)\dots(n-5k+1) \frac{n^{\varepsilon k}}{n^{5k}} = \frac{n!}{2k(n-5k)!} \cdot \frac{n^{\varepsilon k}}{n^{5k}} \tag{5.1}$$

Stirling's formula states that $n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$, so the expected number of such cycles transforms as follows.

$$\frac{n^n}{2ke^{5k}(n-5k)^{n-5k}} \cdot \frac{n^{\varepsilon k}}{n^{5k}} = \frac{1}{2ke^{5k}} \cdot \left(1 + \frac{5k}{n-5k}\right)^{n-5k} \cdot n^{\varepsilon k} > n^{\varepsilon' k}.$$

Here, ε' in $(0, \varepsilon)$ is a fixed value such that $n^{(\varepsilon-\varepsilon')k} > 2ke^{5k}$, for any $k < \lfloor n/5 \rfloor$ and n large enough.

So, the expected number of cycles that have length k within $\mathfrak{S}'$ and that have length at most l within the quotient $\mathfrak{S}'/\mathtt{eq}$ is at least $\left(\frac{2}{5}\right)^k n^{\varepsilon' k}$, so it belongs to $\Omega(n^{\varepsilon'' k})$, where ε'' in $(0, \varepsilon')$ is a constant such that $n^{(\varepsilon'-\varepsilon'')k} > \left(\frac{5}{2}\right)^k$, for all $k < \lfloor n/5 \rfloor$ and n large enough. As k can take values up to $\lfloor n/5 \rfloor$, we conclude that the number of cycles that become short in the quotient is greater than the number of tuples in $\mathfrak{S}'$, so we cannot guarantee that we can remove all short cycles from $\mathfrak{S}'/\mathtt{eq}$ by removing $\mathcal{O}(n^{\varepsilon l})$ tuples as in the proof of Lemma 5.1.4.

Regular graphs

A graph is *regular* if all the vertices have the same degree, *i.e.*, if there exists $d \in \mathbb{N}$ such that for any vertex there are exactly d edges adjacent to it. If $\mathfrak{G}$ is a graph, then a graph $\mathfrak{L}_{\mathfrak{G}}$ is called the *line graph of* $\mathfrak{G}$ if its vertex set $L_{\mathfrak{G}} = \mathsf{E}^{\mathfrak{G}}$ is the edge set of $\mathfrak{G}$ and two vertices $e_1, e_2 \in L_{\mathfrak{G}}$ are adjacent if the corresponding edges share a common point in $\mathfrak{G}$.

In this part, we introduce two properties such that if a graph satisfies them, then one could use this graph for proving the dichotomy of MMSNP_2. The first property is used in the deterministic proof of the equivalence between CSP and MMSNP of Kun in [Kun13].

Property 1 (ε-expander)**.** Let $\mathfrak{G}$ be a graph. For two subsets $S, T \subseteq G$, denote by $\mathsf{E}(S,T)$ the subset of $\mathsf{E}^{\mathfrak{G}}$ of edges with one end in S and the other in T. Then $\mathfrak{G}$ is called *ε-expander* if it satisfies the following condition, for any subsets $S, T \subseteq G$:

$$\left| |\mathsf{E}(S,T)| - \frac{|S||T|}{|G|^2} |\mathsf{E}^{\mathfrak{G}}| \right| < \varepsilon |\mathsf{E}^{\mathfrak{G}}|. \tag{5.2}$$

The second property appears from the difference between MMSNP and MMSNP_2 and from the construction of an ε-expander structure from a regular graph in [Kun13].

Property 2. Let $\mathfrak{G}$ be a graph. Let $\mathfrak{T}_{\mathfrak{G}}$ be a graph with its vertex set $T_{\mathfrak{G}}$ consisting of all the triangles of $\mathfrak{G}$. Two vertices $t_1, t_2 \in T_{\mathfrak{G}}$ are adjacent if the corresponding two triangles of $\mathfrak{G}$ have two vertices out of three in common. $\mathfrak{T}_{\mathfrak{G}}$ is called the *triangle graph* of $\mathfrak{G}$. The graph $\mathfrak{G}$ is called an *ε-triangle-expander* if $\mathfrak{T}_{\mathfrak{G}}$ is an ε-expander.

Property 1 is well-studied and there are numerous examples of graphs that satisfy it. Despite, Property 2 is introduced here for the first time. We could provide an example of a graph that is an ε-triangle-expander.

Example 5.3.1. Let $\mathfrak{G}$ be a graph with the domain $\{a, b, c_1, \ldots, c_n\}$. The set of edges $\mathsf{E}^{\mathfrak{G}}$ is $\{ab, ac_1, bc_1, \ldots, ac_n, bc_n\}$. Then any triangle is of the form abc_i and every two triangles have the edge ab in common. Thus, $\mathfrak{T}_{\mathfrak{G}} = K_n$ is the clique on n vertices. $\triangle$

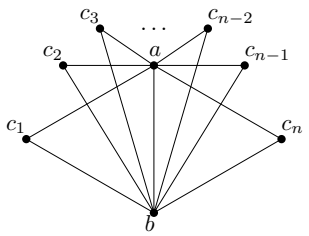


Figure 5.17: An example of a graph $\mathfrak{G}$ whose triangle graph $\mathfrak{T}_{\mathfrak{G}}$ is a clique.

Our goal is to obtain a relational structure over a relational signature with symbols of large arity from a graph, similarly to what is made in Kun's paper [Kun13]. However, he does not consider triangles, he constructs relational tuples of arity k from paths of the length k within a regular graph. And the resulting relational structure satisfies Property 1 which is sufficient for him. This is not sufficient for us as we need to take into the account Property 2 as well.

We are going to show that the approach of Kun fails to satisfy Property 2. For simplicity, we consider a particular case, when the corresponding relational structure has one ternary symbol (that is associated with a triangle). In order to construct an ε-expander, Kun takes some regular graph $\mathfrak{G}$ with a small enough value of $\frac{\lambda_2}{d}$, where λ_2 denotes the second largest eigenvalue (with respect to its absolute value) of the adjacency matrix of $\mathfrak{G}$ and d denotes the degree of any vertex of the graph. The value $\frac{\lambda_2}{d}$ is called the *eigenvalue gap* of $\mathfrak{G}$.

The following proposition shows that if one constructs $\mathfrak{T}_{\mathfrak{G}}$ by considering paths of length 2 instead of triangles, then no graph $\mathfrak{G}$ can satisfy both Property 1 and Property 2 at the same time.

Proposition 5.3.1. *Let $\mathfrak{G}$ be a regular graph of degree d, with second largest eigenvalue λ_2. Also suppose that $\mathfrak{G}$ has girth at least 5 so that saying that two paths of length 2 share two vertices is equivalent to saying that these two paths share an edge. Let $\mathfrak{L}_{\mathfrak{G}}$ be the line graph of $\mathfrak{G}$. And let $\mathfrak{T}_{\mathfrak{G}}$ be a graph with the domain $T_{\mathfrak{G}} = \{abc \mid ab, bc \in \mathsf{E}^{\mathfrak{G}}\}$ consisting of all the paths of length 2 in $\mathfrak{G}$. And there is an edge between two vertices $a_1a_2a_3$ and $b_1b_2b_3$ of $T_{\mathfrak{G}}$ if $|\{a_1, a_2, a_3\} \cap \{b_1, b_2, b_3\}| = 2$. Then, the eigenvalue gap of $\mathfrak{T}_{\mathfrak{G}}$ is at least $\frac{3}{4}$.*

Proof. Recall that, for a graph $\mathfrak{G}$ on n vertices and with m edges, the *adjacency matrix* $\mathsf{A}_{\mathfrak{G}}$ is an $n \times n$ matrix where $a_{ij} = 1$, if $ij \in \mathsf{E}^{\mathfrak{G}}$, and $a_{ij} = 0$, otherwise. And the *incidence matrix* of $\mathfrak{G}$ is an $n \times m$ matrix $\mathsf{X}_{\mathfrak{G}}$ where $x_{ve} = 1$, if the edge e is incident to the vertex v, and $x_{ve} = 0$, otherwise.

It is known that $\mathsf{X}_{\mathfrak{G}}\mathsf{X}_{\mathfrak{G}}^t = \mathsf{A}_{\mathfrak{G}} + d\mathsf{I}_n$ where I_n is the incidence matrix of size n. Also, $\mathsf{X}_{\mathfrak{G}}^t\mathsf{X}_{\mathfrak{G}} = \mathsf{A}_{\mathfrak{L}_{\mathfrak{G}}} + 2(d-1)\mathsf{I}_m$, where $\mathfrak{L}_{\mathfrak{G}}$ is the line graph of $\mathfrak{G}$.

There is a proof of Sachs that shows how the characteristic polynomials of $\mathsf{A}_{\mathfrak{G}}$ and $\mathsf{A}_{\mathfrak{L}_{\mathfrak{G}}}$ are related to each other, in the book "Algebraic Graph Theory" of Biggs [Big74]. Two matrices are considered:

$$\mathsf{U} = \begin{pmatrix} \lambda\mathsf{I}_n & -\mathsf{X}_{\mathfrak{G}} \\ 0 & \mathsf{I}_m \end{pmatrix}, \quad \mathsf{V} = \begin{pmatrix} \mathsf{I}_n & \mathsf{X}_{\mathfrak{G}} \\ \mathsf{X}_{\mathfrak{G}}^t & \lambda\mathsf{I}_m \end{pmatrix}.$$

These matrices provide the following fact about the characteristic polynomials of $\mathsf{X}_{\mathfrak{G}}$ and $\mathsf{X}_{\mathfrak{L}_{\mathfrak{G}}}$:

$$\det(\mathsf{UV}) = \det(\mathsf{VU}) \Leftrightarrow \lambda^m \det(\lambda\mathsf{I}_n - \mathsf{X}_{\mathfrak{G}}\mathsf{X}_{\mathfrak{G}}^t) = \lambda^n \det(\lambda\mathsf{I}_m - \mathsf{X}_{\mathfrak{G}}^t\mathsf{X}_{\mathfrak{G}}) \Leftrightarrow$$

$\Leftrightarrow \lambda^m \det\big((\lambda-d)\mathsf{I}_n - \mathsf{A}_{\mathfrak{G}}\big) = \lambda^n \det\big((\lambda-2)\mathsf{I}_m - \mathsf{A}_{\mathfrak{L}_{\mathfrak{G}}}\big) \Leftrightarrow (\lambda+2)^{m-n}\chi_{\mathfrak{G}}(\lambda-d+2) = \chi_{\mathfrak{L}_{\mathfrak{G}}}(\lambda)$.

Characteristic polynomials are useful if one needs to compute the second largest eigenvalue. So, in order to finish the proof, we need to compute the characteristic polynomial of $\mathfrak{T}_{\mathfrak{G}}$. Firstly, in a d-regular graph the number of edges equals $m = \frac{nd}{2}$, and the number of paths of length 2 equals $m_2 := \frac{nd(d-1)}{2}$. Let X_Δ be an $m \times m_2$ matrix, where every row is associated with an edge of $\mathfrak{G}$ and every column is associated with a path of length 2, $\mathsf{X}_\Delta[e,p] = 1$, if the edge e belongs to the path p, and it equals 0, otherwise. Observe that $\mathsf{A}_{\mathfrak{T}_{\mathfrak{G}}} = \mathsf{X}_\Delta^t\mathsf{X}_\Delta - 2\mathsf{I}_{m_2}$ and $\mathsf{A}_{\mathfrak{L}_{\mathfrak{G}}} = \mathsf{X}_\Delta\mathsf{X}_\Delta^t - 2(d-1)\mathsf{I}_m$ as every edge is contained in exactly $2(d-1)$ paths. For every path there are exactly $(d-1)+2(d-2)+(d-1) = 2(2d-3)$ other paths that share an edge with it. Hence, the degree of every vertex of $\mathfrak{T}_{\mathfrak{G}}$ is $2(2d-3)$. Using the similar approach as above we get:

$$\begin{aligned}\chi_{\mathfrak{T}_{\mathfrak{G}}}(\lambda) = \det(\lambda\mathsf{I}_{m_2} - \mathsf{A}_{\mathfrak{T}_{\mathfrak{G}}}) &= \det\big((\lambda+2)\mathsf{I}_{m_2} - \mathsf{X}_\Delta^t\mathsf{X}_\Delta\big) = \\ &= (\lambda+2)^{m_2-m}\det\big((\lambda+2)\mathsf{I}_m - \mathsf{X}_\Delta\mathsf{X}_\Delta^t\big) = \\ &= (\lambda+2)^{m_2-m}\det\big((\lambda-2d+4)\mathsf{I}_m - \mathsf{A}_{\mathfrak{L}_{\mathfrak{G}}}\big) = \\ &= (\lambda+2)^{m_2-m}\chi_{\mathfrak{L}_{\mathfrak{G}}}(\lambda-2d+4) = (\lambda+2)^{m_2-n}\chi_{\mathfrak{G}}\big(\lambda-3(d-2)\big).\end{aligned}$$

Thus, the second largest eigenvalue of $\mathfrak{T}_{\mathfrak{G}}$ is $\lambda_2 + 3(d-2)$, therefore, the eigenvalue gap of $\mathfrak{T}_{\mathfrak{G}}$ is $\frac{\lambda_2+3(d-2)}{2(2d-3)}$ that cannot be smaller than $\frac{3(d-2)}{2(2d-3)} \approx \frac{3}{4}$. □

As we need the eigenvalue gap to be arbitrarily small, this approach is not useful for us.

Chapter 6

Matrix Partition

We consider an extension of CSP which is called Matrix Partition problems (MP). We mainly study its generalisations by manipulating the containments of the problem input and by considering problems over arbitrary relational signatures. We show how these generalisations are related to each other in terms of having a dichotomy. Being motivated by Atserias' characterization of FO-definable fragment of CSP by having a finitary duality in [Ats08], we study how different input extensions influence an MP problem to be expressed by finitely many minimal obstructions. Apart from that, we provide an MP problem which is P-time solvable with a trivial clone of polymorphisms. Finally, we provide a logic that is an extension of MMSNP and a fragment of MonadicSNP that contains MP and that becomes a candidate to be P-time equivalent to MP.

6.1 Preliminaries

We deal mostly with labeled complete relational structures, *i.e.*, each relation of arity k is V^k, and tuples are labeled by the elements of a partially ordered set, defined for example in [Sta00].

Definition 20 ($(*,\tau)$-structures)**.** Let $(P_*, \preceq_*)$ be a partially ordered set (poset), and τ be a finite relational signature. A $(*,\tau)$-*structure* is a tuple $\mathfrak{G} := (G; \mathsf{R}_1^{\mathfrak{G}}, \ldots, \mathsf{R}_n^{\mathfrak{G}})$ with G a finite set, and for each $i \in [n]$, $\mathsf{R}_i^{\mathfrak{G}} \colon G^{k_i} \to P_*$ is interpreted as a mapping to the elements of the poset.

We will always denote a $(*,\tau)$-structure by a boldface capital letter, e.g. $\mathfrak{A}$, and its domain by the same letter in plain font, e.g. A. It is worth mentioning that the notion of $(*,\tau)$-structure is different from the one in *universal algebra*, where in the latter case the functional symbol R_i is interpreted in $\mathfrak{G}$ as a function from $G^{k_i} \to G$.

For a $(*,\tau)$-structure $\mathfrak{G}$ and $X \subseteq G$, the substructure of $\mathfrak{G}$ induced by X is the $(*,\tau)$-structure $\mathfrak{G}'$ with domain $G' = X$ and, for $\mathsf{R} \in \tau$ of arity k and $\mathbf{t} \in X^k$, $\mathsf{R}^{\mathfrak{G}'}(\mathbf{t}) = \mathsf{R}^{\mathfrak{G}}(\mathbf{t})$; and we denote by $\mathfrak{G} \setminus X$ the substructure of $\mathfrak{G}$ induced by $G \setminus X$.

We now extend the notion of homomorphism between relational structures to $(*,\tau)$-structures, the difference being the ability to map a tuple to a "greater" one.

Definition 21 (homomorphism for $(*,\tau)$-structures)**.** For two $(*,\tau)$-structures $\mathfrak{G}$ and $\mathfrak{H}$, a mapping $\mathfrak{h} \colon G \to H$ is called a *homomorphism from* $\mathfrak{G}$ *to* $\mathfrak{H}$ if, for each $\mathsf{R} \in \tau$ of arity k, and $\mathbf{t} \in G^k$, $\mathsf{R}^{\mathfrak{G}}(\mathbf{t}) \preceq_* \mathsf{R}^{\mathfrak{H}}\big(\mathfrak{h}(\mathbf{t})\big)$.

As usual, we will write $\mathfrak{h}\colon \mathfrak{G} \to \mathfrak{H}$ to mean that $\mathfrak{h}\colon G \to H$ is a homomorphism from $\mathfrak{G}$ to $\mathfrak{H}$. We say that $\mathfrak{h} : \mathfrak{G} \to \mathfrak{H}$ is *surjective* (resp. *injective*) if $\mathfrak{h} : G \to H$ is surjective (resp. injective).

We can now explain how the notion of homomorphism between $(*, \tau)$-structures subsumes the usual ones. Before, let us recall the partial orders we consider in this thesis.

- $(P_{01}, \preceq_{01})$, where $P_{01} = \{0, 1\}$ and $\preceq_{01}$ is the empty order with 0 and 1 incomparable.
- $(P_{\text{CSP}}, \preceq_{\text{CSP}})$, where $P_{\text{CSP}} = \{0, 1\}$ and $\preceq_{\text{CSP}}$ is a total order with $0 \preceq_{\text{CSP}} 1$.
- $(P_\star, \preceq_\star)$, where $P_\star = \{0, 1, \star\}$ and $\preceq_\star$ is the poset with $0 \preceq_\star \star$ and $1 \preceq_\star \star$, and 0 incomparable with 1.
- $(P_\varnothing, \preceq_\varnothing)$ where $P_\varnothing = \{\varnothing, 0, 1, \star\}$ and $\preceq_\varnothing$ is the poset with $\varnothing \preceq_\varnothing 0 \preceq_\varnothing \star$ and $\varnothing \preceq_\varnothing 1 \preceq_\varnothing \star$, and 0 incomparable with 1.

Remark. If the signature τ is clear from the context, then we will just write $*$-*structure* instead of $(*, \tau)$-structure, for $* \in \{01, \star, \varnothing\}$. Also, if $\tau = \{\mathsf{E}(\cdot, \cdot)\}$, then we will write $*$-*graph* instead. Finally, we will talk about *relational* τ-*structures* and *directed graphs*, instead of (CSP, τ)-structures and CSP-graphs. Furthermore, for any tuple (arc) $\mathbf{t} \in A^k$ of a $*$-structure ($*$-graph) $\mathfrak{A}$ corresponding to a symbol R in τ that is clear from the context, we will call $\mathbf{t}$ a *p-tuple (p-arc)* if $\mathsf{R}^{\mathfrak{A}}(\mathbf{t}) = p$ for some element p of the poset $(P_*, \preceq_*)$.

It is not hard to check that (CSP, τ)-structures are exactly associated with the usual notion of relational τ-structures, and homomorphisms between (CSP, τ)-structures to usual homomorphisms. Notice that homomorphisms between $(01, \tau)$-structures are exactly full homomorphisms on relational structures.

Proposition 6.1.1. *Let $(P_*, \preceq_*)$ and $(P_{*'}, \preceq_{*'})$ be two posets, with $(P_*, \preceq_*)$ a sub-poset of $(P_{*'}, \preceq_{*'})$. Then every $(*, \tau)$-structure is also a $(*', \tau)$-structure, for any τ.*

Particularly, for any τ, every $(01, \tau)$-structure is a $(\star, \tau)$-structure, and every $(\star, \tau)$-structure is a $(\varnothing, \tau)$-structure. For $* \in \{01, \star, \varnothing\}$, we denote by $\mathbf{Struct}_*[\tau]$ the set of all $(*, \tau)$-structures. We use the notation $\mathbf{Struct}_*$ because one can use $(*, \tau)$-structures as objects and homomorphisms as arrows to make a category. From the proposition above, and the definitions of $(P_{01}, \preceq_{01}), (P_\star, \preceq_\star)$ and $(P_\varnothing, \preceq_\varnothing)$, we have the following inclusion:

$$\mathbf{Struct}_{01}[\tau] \subset \mathbf{Struct}_\star[\tau] \subset \mathbf{Struct}_\varnothing[\tau].$$

We can now define the homomorphism problems, that we restrict for conciseness to the four posets: $(P_{01}, \preceq_{01}), (P_\star, \preceq_\star), (P_\varnothing, \preceq_\varnothing), (P_{\text{CSP}}, \preceq_{\text{CSP}})$.

Definition 22. Let τ be a finite signature and $* \in \{01, \star\}$. For a $\star$-structure $\mathfrak{H}$, the problem $\text{MP}^\tau_*(\mathfrak{H})$ denotes the set of all $*$-structures $\mathfrak{G}$ such that there exists a homomorphism $\mathfrak{h}\colon \mathfrak{G} \to \mathfrak{H}$. If $\mathfrak{H}$ is a CSP-structure, then we write $\text{CSP}^\tau(\mathfrak{H})$ as the set of all CSP-structures $\mathfrak{G}$ such that there exists a homomorphism $\mathfrak{h}\colon \mathfrak{G} \to \mathfrak{H}$. We always omit subscript 01 in $\text{MP}^\tau_{01}(\mathfrak{H})$.

The set of all $\varnothing$-structures $\mathfrak{G}$ such that there is a homomorphism $\mathfrak{h}\colon \mathfrak{G} \to \mathfrak{H}$, with $\mathfrak{H}$ a $\varnothing$-structure, is denoted by $\text{MP}^\tau_\varnothing(\mathfrak{H})$.

By MP^τ, $\mathrm{MP}^\tau_\star$, $\mathrm{MP}^\tau_\varnothing$ and CSP^τ we denote, respectively, the families of problems $\mathrm{MP}^\tau(\mathfrak{H})$, $\mathrm{MP}^\tau_\star(\mathfrak{H})$, $\mathrm{MP}^\tau_\varnothing(\mathfrak{H})$ and $\mathrm{CSP}^\tau(\mathfrak{H})$, for all $\star$-structures $\mathfrak{H}$. For $\mathrm{CSP}^\tau(\mathfrak{H})$ we of course demand that $\mathfrak{H}$ is a relational τ-structure, and for $\mathrm{MP}^\tau_\varnothing(\mathfrak{H})$, we consider $\mathfrak{H}$ to be a $(\varnothing, \tau)$-structure. If $\tau = \{\mathsf{E}(\cdot,\cdot)\}$ – the directed graph signature, then we will omit the τ-superscript and will just write MP, $\mathrm{MP}_\star$, $\mathrm{MP}_\varnothing$ and CSP.

Remark. All along the thesis, whenever we consider a problem $\mathrm{MP}^\tau_*(\mathfrak{H})$, for $* \in \{01, \star, \varnothing\}$, we consider that there is no $x \in H$ such that for all $\mathsf{R} \in \tau$, $\mathsf{R}(x, \ldots, x) = \star$. Otherwise, the problem is trivial as then $\mathrm{MP}^\tau_*(\mathfrak{H})$ equals $\mathbf{Struct}^\tau_*$.

Notice that as every 01-structure is also a $\star$-structure, and every $\star$-structure is also a $\varnothing$-structure, our problems are well defined.

Example 6.1.1. Let $\mathfrak{H}$ be as on Figure 6.1 on page 131. We can define three problems: $\mathrm{MP}(\mathfrak{H}), \mathrm{MP}_\star(\mathfrak{H})$, and $\mathrm{MP}_\varnothing(\mathfrak{H})$. The first one takes an input directed graph and checks if all the loopless vertices induce an independent set, if all the vertices with loops induce a clique, and if there is an arc from any vertex with a loop to any loopless one. The second one takes an input $\star$-graph and checks if all vertices with 0-loops induce an independent set, if all the vertices with 1-loops induce a clique, and if there is a 1-arc from any vertex with a 1-loop to any with a 0-loop. The third problem checks if there is a partition of the domain of an input $\varnothing$-graph $\mathfrak{G}$ in two parts G_0, G_1 such that

- for all x, y in G_0, we have $\mathsf{E}^{\mathfrak{G}}(x, y) \in \{\varnothing, 0\}$, or equivalently $\mathsf{E}^{\mathfrak{G}}(x, y) \preceq_\varnothing 0$;
- for all x, y in G_1, we have $\mathsf{E}^{\mathfrak{G}}(x, y) \preceq_\varnothing 1$, and
- for all x in G_1, y in G_0, we have $\mathsf{E}^{\mathfrak{G}}(x, y) \preceq_\varnothing 1$.

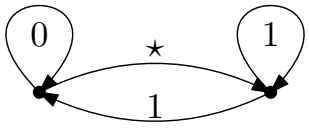


Figure 6.1: An example of a $\star$-graph $\mathfrak{H}$.

△

Nuance in the definition

The original definition of Matrix Partition problem is as follows. Further, we explain why this definition is inconvenient for some questions that we consider.

Definition 23 (Matrix Partitions [FHX07])**.** Let M be a an $n \times n$-matrix with entries on $\{0, 1, \star\}$. A loopless graph $\mathfrak{G}$ admits an *M-partition* if there is a function $\mathfrak{m} : G \to [n]$ such that for all distinct $x, y \in G$, $\mathsf{E}^{\mathfrak{G}}(x, y) \preceq_\star \mathsf{M}[\mathfrak{m}(x), \mathfrak{m}(y)]$.

There are some differences between the definition of M-partition in [FHX07] and the definition of $\mathrm{MP}(\mathfrak{H})$. Unlike Feder and Hell, we consider all possible graphs in the input, not only the loopless ones. Thus, we consider every two vertices that are not necessarily distinct, however, Feder and Hell consider every two distinct ones.

We can put 01**-graphs and $\star$-graphs into the same category.** For convenience, we want to consider directed graphs and $\star$-graphs as objects of the same category $\mathbf{Struct}_\star$. In this case, directed graphs become 01-graphs, according to the definition of MP. Morphisms (arrows) of any category must be transitive. Our definition permits to put digraphs as 01-graphs into the same category as trigraphs, where a morphism between two 01-graphs represents a full homomorphism between the corresponding digraphs. Full homomorphisms are transitive. But they are not transitive in the case of Feder and Hell, we justify it in Example 6.1.2.

Example 6.1.2. Consider three 01-graphs: $\mathfrak{K}_{2,2}$, $\mathfrak{K}_2$, and $\mathfrak{L}$. They are displayed on Figure 6.2. Each edge is 1 and each non-edge is 0 within the corresponding partition matrices. $\mathfrak{K}_{2,2}$ admits $\mathfrak{K}_2$-partition, where the 2 parts of $\mathfrak{K}_{2,2}$ are exactly the partition classes. And

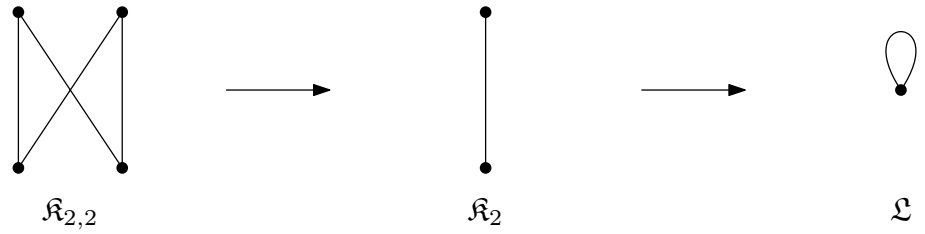


Figure 6.2: The graphs from Example 6.1.2.

$\mathfrak{K}_2$ admits $\mathfrak{L}$-partition, as it is a clique. But $\mathfrak{K}_{1,2}$ does not admit $\mathfrak{L}$-partition as it is not a clique. $\triangle$

We need a less expressive logic for our definition. Feder and Hell's definition requires more expressivity from logical sentences that describe their problems, as they have to use inequalities. When, at the same time, our definition does not have this constraint. In Example 6.1.3, we construct MonadicSNP sentences that describe both problem variations about the same $\star$-graph.

Example 6.1.3. Let $\mathfrak{H}$ be a $\star$-graph with the 2-element domain: $H = \{a, b\}$, and with E interpreted as follows: $\mathsf{E}^{\mathfrak{H}}(a,a) = \mathsf{E}^{\mathfrak{H}}(a,b) = 0$, $\mathsf{E}^{\mathfrak{H}}(b,b) = 1$, and $\mathsf{E}^{\mathfrak{H}}(b,a) = \star$. Then the corresponding Matrix Partition problem of Feder and Hell can be formulated by the sentence:

$$\exists \mathsf{M}_1, \mathsf{M}_2 \, \forall x, y \, \begin{pmatrix} \neg\big(\neg \mathsf{M}_1(x) \wedge \neg \mathsf{M}_2(x)\big) \\ \wedge \neg\big(\mathsf{M}_1(x) \wedge \mathsf{M}_2(x)\big) \end{pmatrix} \wedge \begin{pmatrix} \neg\big(\mathsf{M}_1(x) \wedge \mathsf{M}_1(y) \wedge x \neq y \wedge \mathsf{E}(x,y)\big) \\ \wedge \neg\big(\mathsf{M}_1(x) \wedge \mathsf{M}_2(y) \wedge x \neq y \wedge \mathsf{E}(x,y)\big) \\ \wedge \neg\big(\mathsf{M}_2(x) \wedge \mathsf{M}_2(y) \wedge x \neq y \wedge \neg\mathsf{E}(x,y)\big) \end{pmatrix}.$$

And the problem $\mathrm{MP}(\mathfrak{H})$ can be formulated by a similar sentence but without inequalities:

$$\exists \mathsf{M}_1, \mathsf{M}_2 \, \forall x, y \, \begin{pmatrix} \neg\big(\neg \mathsf{M}_1(x) \wedge \neg \mathsf{M}_2(x)\big) \\ \wedge \neg\big(\mathsf{M}_1(x) \wedge \mathsf{M}_2(x)\big) \end{pmatrix} \wedge \begin{pmatrix} \neg\big(\mathsf{M}_1(x) \wedge \mathsf{M}_1(y) \wedge \mathsf{E}(x,y)\big) \\ \wedge \neg\big(\mathsf{M}_1(x) \wedge \mathsf{M}_2(y) \wedge \mathsf{E}(x,y)\big) \\ \wedge \neg\big(\mathsf{M}_2(x) \wedge \mathsf{M}_2(y) \wedge \neg\mathsf{E}(x,y)\big) \end{pmatrix}.$$

$\triangle$

It is more natural to generalise according to our definition. It is natural when a relational structure has a relational tuple of the form $\mathsf{R}(x,\dots,x)$. It is not clear why we should omit all structures with such tuples. Apart from that, we show, see Section 6.3.2, that without loss of generality we can consider only classes MP^{σ}, where σ consists of a unique relation symbol. The techniques used in the proof cannot be applied to show a similar result for problems where input structures contain no tuples of the form $\mathsf{R}(x,\dots,x)$.

Remark. It is not known if these two definitions are give P-time equivalent classes of problems. It can be shown that any problem $\mathrm{MP}(\mathfrak{H})$ can be reduced to a list matrix partition problem with the matrix M, where M is a $|H|\times|H|$-matrix such that, for any i,j in $[|H|]$, $\mathsf{M}[i,j]=\mathsf{E}^{\mathfrak{H}}(h_i,h_j)$. This implies that if a list matrix partition problem is P-time solvable, then the corresponding MP problem is also P-time solvable.

6.2 Matrix Partition and CSP

Let $\tau=\{\mathsf{R}_1,\dots,\mathsf{R}_n\}$ be a signature, the arity of each R_i denoted by k_i. We prove in this section that there is a signature τ_{CSP} such that any problem in $\mathrm{MP}^{\tau}_{\varnothing}$ is P-time equivalent to a problem in $\mathrm{CSP}^{\tau_{\mathrm{CSP}}}$ and vice versa, that is, $\mathrm{MP}^{\tau}_{\varnothing}\equiv_p\mathrm{CSP}^{\tau_{\mathrm{CSP}}}$.

The signature τ_{CSP} is defined by repeating each symbol of τ two times, one for 0-tuples and one for 1-tuples, $\star$-tuples will be considered as 0- and 1-tuples at the same time. Let $\tau_{\mathrm{CSP}}=\{\mathsf{R}_{1,0},\mathsf{R}_{1,1},\dots,\mathsf{R}_{n,0},\mathsf{R}_{n,1}\}$, for $i\in[n]$, $\mathsf{R}_{i,0},\mathsf{R}_{i,1}$ both have arity k_i.

Every $\varnothing$-structure $\mathfrak{A}_{\varnothing}$ is associated with a relational τ_{CSP}-structure $\mathfrak{A}_{\mathrm{CSP}}$ with the same domain A and the symbols $\mathsf{R}_{i,0},\mathsf{R}_{i,1}$ of τ_{CSP} are interpreted as follows:

$$\forall\mathsf{R}_i\in\tau,\mathbf{t}\in A^{k_i},j\in\{0,1\}\colon \mathsf{R}_{i,j}^{\mathfrak{A}_{\mathrm{CSP}}}(\mathbf{t})=1\Leftrightarrow j\preceq_{\varnothing}\mathsf{R}_i^{\mathfrak{A}_{\varnothing}}(\mathbf{t}). \tag{6.1}$$

Observation 2. $\mathfrak{A}_{\mathrm{CSP}}$ is constructible in P-time in the size of $\mathfrak{A}_{\varnothing}$.

Observation 3. For any $(\varnothing,\tau)$-structure $\mathfrak{A}_{\varnothing}$, there exists a unique relational τ_{CSP}-structure $\mathfrak{A}_{\mathrm{CSP}}$, and for any relational τ_{CSP}-structure, there exists a unique $(\varnothing,\tau)$-structure $\mathfrak{A}_{\varnothing}$ such that eq. (6.1) is satisfied. That is there is a one-to-one correspondence between $(\varnothing,\tau)$-structures and relational τ_{CSP}-structures.

Theorem 6.2.1. *$\mathrm{MP}^{\tau}_{\varnothing}$ and $\mathrm{CSP}^{\tau_{CSP}}$ are P-time equivalent.*

Proof. Let $\mathfrak{A}_{\varnothing}$ be a $\varnothing$-structure. We first prove that $\mathfrak{B}_{\varnothing}\in\mathrm{MP}^{\tau}_{\varnothing}(\mathfrak{A}_{\varnothing})$ if and only if $\mathfrak{B}_{\mathrm{CSP}}\in\mathrm{CSP}^{\tau_{\mathrm{CSP}}}(\mathfrak{A}_{\mathrm{CSP}})$.

Assume that $\mathfrak{B}_{\varnothing}\in\mathrm{MP}^{\tau}_{\varnothing}(\mathfrak{A}_{\varnothing})$ and let $\mathsf{h}:B\to A$ be a homomorphism from $\mathfrak{B}_{\varnothing}$ to $\mathfrak{A}_{\varnothing}$. We will show that the same map h is a homomorphism from $\mathfrak{B}_{\mathrm{CSP}}$ to $\mathfrak{A}_{\mathrm{CSP}}$. For any tuple $\mathbf{t}$ and its image $\mathsf{h}(\mathbf{t})$, we know that for any $\mathsf{R}_{i,j}\in\tau_{\mathrm{CSP}}$:

$$\mathsf{R}_{i,j}^{\mathfrak{B}_{\mathrm{CSP}}}(\mathbf{t})=1\Leftrightarrow j\preceq_{\varnothing}\mathsf{R}_i^{\mathfrak{B}_{\varnothing}}(\mathbf{t})\Rightarrow j\preceq_{\varnothing}\mathsf{R}_i^{\mathfrak{A}_{\varnothing}}\big(\mathsf{h}(\mathbf{t})\big)\Leftrightarrow\mathsf{R}_{i,j}^{\mathfrak{A}_{\mathrm{CSP}}}\big(\mathsf{h}(\mathbf{t})\big)=1.$$

Now, backwards, assume that $\mathfrak{B}_{\mathrm{CSP}}\in\mathrm{CSP}^{\tau_{\mathrm{CSP}}}(\mathfrak{A}_{\mathrm{CSP}})$, and let $\mathsf{h}:B\to A$ a homomorphism from $\mathfrak{B}_{\mathrm{CSP}}$ to $\mathfrak{A}_{\mathrm{CSP}}$. Similarly as in the first part, for a tuple $\mathbf{t}$, we know that for all $\mathsf{R}_i\in\tau,j\in\{0,1\}$:

$$j\preceq_{\varnothing}\mathsf{R}_i^{\mathfrak{B}_{\varnothing}}(\mathbf{t})\Leftrightarrow\mathsf{R}_{i,j}^{\mathfrak{B}_{\mathrm{CSP}}}(\mathbf{t})=1\Rightarrow\mathsf{R}_{i,j}^{\mathfrak{A}_{\mathrm{CSP}}}\big(\mathsf{h}(\mathbf{t})\big)=1\Leftrightarrow j\preceq_{\varnothing}\mathsf{R}_i^{\mathfrak{A}_{\varnothing}}\big(\mathsf{h}(\mathbf{t})\big).$$

This implies that h is a homomorphism from $\mathfrak{B}_{\varnothing}$ to $\mathfrak{A}_{\varnothing}$. □

For $* \in \{01, \star, \varnothing\}$, the notion of homomorphism between $(*, \tau)$-structures admits a *core* notion. It generalises the notion of a core for trigraphs ($\star$-graphs) given in [HN07]. For $* \in \{01, \star, \varnothing\}$, a $(*, \tau)$-structure $\mathfrak{C}$ is called a *core* if any homomorphism $\mathfrak{h} \colon \mathfrak{C} \to \mathfrak{C}$ is an isomorphism, where isomorphism between $(*, \tau)$-structures is the same as usual.

Proposition 6.2.2. *Let $* \in \{01, \star, \varnothing\}$. Then for any $(*, \tau)$-structure $\mathfrak{A}_*$, there exists a unique, up to isomorphism, $(*, \tau)$-structure $\mathfrak{C}_*$ such that it is a core and $\mathfrak{A}_* \leftrightarrows \mathfrak{C}_*$, and $\mathfrak{C}_*$ embeds into $\mathfrak{A}_*$.*

Proof. We know that $\mathfrak{A}_*$ is also a $\varnothing$-structure by Proposition 6.1.1. Then, consider the relational τ_{CSP}-structure $\mathfrak{A}_{\text{CSP}}$ provided by Theorem 6.2.1. It is well-known, see [HN04], that $\mathfrak{A}_{\text{CSP}}$ has the core $\mathfrak{C}_{\text{CSP}}$ embedded into $\mathfrak{A}_{\text{CSP}}$. Let $\mathfrak{C}_*$ be the corresponding $\varnothing$-structure by Theorem 6.2.1, it must also be homomorphically equivalent to $\mathfrak{A}_*$ and be embedded into it. As $\mathfrak{C}_*$ embeds into $\mathfrak{A}_*$, it is also a $*$-structure. Let $\mathfrak{e} : C \to C$ be a non-injective endomorphism. Then the same map $\mathfrak{e}$ will be a non-injective endomorphism of the core $\mathfrak{C}_{\text{CSP}}$ which is impossible. Let $\mathfrak{C}'_*$ be another core of $\mathfrak{A}_*$, that is not isomorphic to $\mathfrak{C}_*$. But then $\mathfrak{C}'_{\text{CSP}}$ must be the core of $\mathfrak{A}_{\text{CSP}}$ and $\mathfrak{C}_{\text{CSP}} \ncong \mathfrak{C}'_{\text{CSP}}$ which is impossible as $\mathfrak{C}_{\text{CSP}}$ is a core. □

If, for a class of problems $\mathcal{C}$, $\mathcal{C} \subseteq_p \text{CSP}$, then $\mathcal{C}$ has a dichotomy. We manage to embed the class $\text{MP}_\varnothing$ into CSP because of the smallest element $\varnothing$. Now we show that this happens every time when the poset $(P_*, \preceq_*)$ is a lattice.

A poset $(P, \preceq)$ is called a *lattice* if, for any two elements x, y in P there are two elements:

- $x \vee y$ – the least upper bound, that is, for any element z, we have

$$x \preceq z \text{ and } y \preceq z \text{ implies } x \vee y \preceq z;$$

- $x \wedge y$ – the greatest lower bound, that is, for any element z, we have

$$z \preceq x \text{ and } z \preceq y \text{ implies } z \preceq x \wedge y.$$

In the rest of this section, we prove the following theorem.

Theorem 6.2.3. *Let τ be a finite relational signature. Suppose that $(P_*, \preceq_*)$ is a lattice. Then, for any $(*, \tau)$-structure $\mathfrak{A}_*$ there is a τ_{CSP}-structure $\mathfrak{A}_{CSP}$ such that $MP^\tau_*(\mathfrak{A}_*)$ is P-time equivalent to $CSP^{\tau_{CSP}}(\mathfrak{A}_{CSP})$, where τ_{CSP} is a finite relational signature.*

We first explain how to construct $\mathfrak{A}_{\text{CSP}}$ from $\mathfrak{A}_*$.

Construction 17. The signature τ_{CSP} is constructed as follows:

$$\tau_{\text{CSP}} := \big\{ \mathsf{R}_l \mid \mathsf{R} \in \tau, l \in P_* \setminus \{\min\} \big\},$$

where min denotes the minimal element of the lattice. Each R_l has the same arity as R has.

The domain of $\mathfrak{A}_{\text{CSP}}$ is the same as the domain of $\mathfrak{A}_*$. For any k-ary R in τ, and for any k-tuple $\mathbf{a}$ of the elements of $\mathfrak{A}_*$, for any l in $P_* \setminus \{\min\}$, $\mathbf{a}$ belongs to $\mathsf{R}_l^{\mathfrak{A}_{\text{CSP}}}$ if and only if $\mathsf{R}^{\mathfrak{A}_*} \succeq_* l$. This is the end of Construction 17.

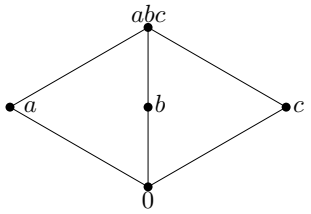


Figure 6.3: The Hasse diagram of $(P_*, \preceq_*)$.

Example 6.2.1. Let $(P_*, \preceq_*)$ be as on Figure 6.3.

Let τ be the graph signature: $\tau = \{\mathsf{E}(\cdot,\cdot)\}$. Then τ_{CSP} contains 4 relational symbols: $\mathsf{E}_a, \mathsf{E}_b, \mathsf{E}_c$, and E_{abc}. For a $*$-graph $\mathfrak{A}_*$, the structure $\mathfrak{A}_{\mathrm{CSP}}$ is constructed as follows.

- For any $(x, y) \in A^2$ such that $\mathsf{E}^{\mathfrak{A}_*}(x, y) = 0$, it is contained in no τ_{CSP}-relation of $\mathfrak{A}_{\mathrm{CSP}}$.
- For any $(x, y) \in A^2$ such that $\mathsf{E}^{\mathfrak{A}_*}(x, y) = a$, it is contained in $\mathsf{E}_a^{\mathfrak{A}_{\mathrm{CSP}}}$. The cases $\mathsf{E}^{\mathfrak{A}_*}(x, y) = b$ and $\mathsf{E}^{\mathfrak{A}_*}(x, y) = c$ are treated similarly.
- For any $(x, y) \in A^2$ such that $\mathsf{E}^{\mathfrak{A}_*}(x, y) = abc$, it is contained in all four relations $\mathsf{E}_a^{\mathfrak{A}_{\mathrm{CSP}}}, \mathsf{E}_b^{\mathfrak{A}_{\mathrm{CSP}}}, \mathsf{E}_c^{\mathfrak{A}_{\mathrm{CSP}}}, \mathsf{E}_{abc}^{\mathfrak{A}_{\mathrm{CSP}}}$.

△

To prove Theorem 6.2.3, we need the two following propositions.

Proposition 6.2.4. *Let $\mathfrak{A}_*, \mathfrak{B}_*$ be two $(*, \tau)$-structures, and $\mathfrak{A}_{CSP}, \mathfrak{B}_{CSP}$ be the relational structures that are obtained by Construction 17. Then, $\mathfrak{B}_* \to \mathfrak{A}_*$ if and only if $\mathfrak{B}_{CSP} \to \mathfrak{A}_{CSP}$.*

Proof. Let $\mathsf{h}\colon B \to A$ be a mapping between the domains. Pick some R-tuple $\mathbf{b}$, it is mapped to $\mathsf{h}(\mathbf{b})$. Suppose that $\mathsf{R}^{\mathfrak{B}_*}(\mathbf{b}) = l$. By construction, we have $\mathbf{b} \in \mathsf{R}_{l'}^{\mathfrak{B}_{\mathrm{CSP}}}$, for any $l' \preceq_* l$. Also, by construction, $\mathsf{R}^{\mathfrak{A}_*}\big(\mathsf{h}(\mathbf{b})\big) \succeq_* l$ if and only if $\mathsf{h}(\mathbf{b})$ is contained in $\mathsf{R}_{l'}^{\mathfrak{A}_{\mathrm{CSP}}}$, for any $l' \preceq_* l$. This implies that h is a homomorphism between $\mathfrak{B}_*$ and $\mathfrak{A}_*$ if and only if it is a homomorphism between $\mathfrak{B}_{\mathrm{CSP}}$ and $\mathfrak{A}_{\mathrm{CSP}}$. □

Proposition 6.2.5. *Let $\mathfrak{G}$ be a finite τ_{CSP}-structure. Suppose that, for any R in τ of arity k and for any k-ary tuple $\mathbf{x}$ of $\mathfrak{G}$, there is l in P_* such that $\mathbf{x}$ belongs to $\mathsf{R}_m^{\mathfrak{G}}$, for any $m \preceq_* l$. Then there is a $(*, \tau)$-structure $\mathfrak{B}_*$ such that $\mathfrak{G} = \mathfrak{B}_{CSP}$, where $\mathfrak{B}_{CSP}$ is obtained from $\mathfrak{B}_*$ by Construction 17.*

Proof. Let $\mathfrak{B}_*$ have the same domain as $\mathfrak{G}$ has. For any $\mathsf{R} \in \tau$ of arity k, and for any k-ary tuple $\mathbf{x}$, we put $\mathsf{R}^{\mathfrak{B}_*}(\mathbf{x}) = l$, where l is the maximal element such that $\mathbf{x}$ is contained in $\mathsf{R}_l^{\mathfrak{G}}$. Let $\mathfrak{B}_{\mathrm{CSP}}$ be obtained from $\mathfrak{B}_*$ by Construction 17. By construction, $\mathfrak{B}_{\mathrm{CSP}}$ and $\mathfrak{G}$ is the same structure. □

Proof of Theorem 6.2.3. For a $(*, \tau)$-structure $\mathfrak{A}_*$, we construct in P-time in size of $\mathfrak{A}_*$, the relational structure $\mathfrak{A}_{\mathrm{CSP}}$, by Construction 17. By Proposition 6.2.4 on page 135, we have $\mathrm{MP}_*^{\tau}(\mathfrak{A}_*) \leq_p \mathrm{CSP}^{\tau_{\mathrm{CSP}}}(\mathfrak{A}_{\mathrm{CSP}})$.

Now, consider an arbitrary relational τ_{CSP}-structure $\mathfrak{G}$, we will find a $(*, \tau)$-structure $\mathfrak{B}_*$ such that $\mathfrak{G} \to \mathfrak{A}_{\mathrm{CSP}}$ if and only if $\mathfrak{B}_* \to \mathfrak{A}_*$. For any R in τ of arity k, consider a k-tuple $\mathbf{x}$ of $\mathfrak{G}$. Let $X_{\mathbf{x}}$ denote the following subset of P_*:

$$X_{\mathbf{x}} := \big\{ l \in P_* \mid \mathbf{x} \in \mathsf{R}_l^{\mathfrak{G}} \big\}.$$

If there is a homomorphism $\mathfrak{g}\colon \mathfrak{G} \to \mathfrak{A}_{\mathrm{CSP}}$, then the tuple $\mathfrak{g}(\mathbf{x})$ must belong to $\mathsf{R}_m^{\mathfrak{A}_{\mathrm{CSP}}}$, for any m which is smaller than the least upper bound $\bigvee X_{\mathbf{x}}$ of the elements of $X_{\mathbf{x}}$. For any such m, we add $\mathbf{x}$ to $\mathsf{R}_m^{\mathfrak{G}}$. Observe that, after doing this, the result can be mapped to $\mathfrak{A}_{\mathrm{CSP}}$ if and only if the original structure$\mathfrak{G}$ can be mapped to $\mathfrak{A}_{\mathrm{CSP}}$, by the construction of $\mathfrak{A}_{\mathrm{CSP}}$. For any other k-tuple, we do a similar operation. After that, for any k-tuple $\mathbf{y}$, $X_{\mathbf{y}}$ contains all the elements l' such that $l' \preceq_* \bigvee X_{\mathbf{y}}$. For any other $\mathsf{R}' \in \tau$ of arity k', we do a similar operation.

The structure that we obtain from $\mathfrak{G}$ satisfies the condition from Proposition 6.2.5 on page 135, this means that there is a $(*, \tau)$-structure $\mathfrak{B}_*$ such that $\mathfrak{G}$ and $\mathfrak{B}_{\mathrm{CSP}}$ is the same structure. By Proposition 6.2.4, $\mathfrak{B}_* \to \mathfrak{A}_*$ if and only if $\mathfrak{B}_{\mathrm{CSP}} \to \mathfrak{A}_{\mathrm{CSP}}$. □

6.3 Generalised Matrix Partition

6.3.1 Input extension

In this section we will prove the following theorem.

Theorem 6.3.1. *For any finite signature τ, MP^τ and $MP^\tau_\star$ are P-time equivalent.*

In order to prove the P-time equivalence, we will show that for any $\star$-structure $\mathfrak{H}$, $\mathrm{MP}^\tau(\mathfrak{H}) \equiv_p \mathrm{MP}^\tau_\star(\mathfrak{H})$. We first do the proof for $\star$-graphs, and then explain how to modify the construction for any τ. Hell and Nešetřil proved in [HN07] that for any $\star$-graph $\mathfrak{G}$ there is a 01-graph $\mathfrak{G}_{01}$ such that $\mathfrak{G} \in \mathrm{MP}_\star(\mathfrak{H}) \Leftrightarrow \mathfrak{G}_{01} \in \mathrm{MP}(\mathfrak{H})$ using probabilistic arguments. We prove the equivalence by giving a deterministic algorithm running in P-time.

In order to prove that for any $\star$-graph $\mathfrak{G}$, there is a 01-graph $\mathfrak{G}_{01}$ such that $\mathfrak{G} \in \mathrm{MP}_\star(\mathfrak{H})$ if and only if $\mathfrak{G}_{01} \in \mathrm{MP}(\mathfrak{H})$, we will use the notion of Hadamard matrices.

An $n \times n$-matrix H_n, which entries are in $\{1, -1\}$, is called a *Hadamard matrix* if

$$\mathsf{H}_n \cdot \mathsf{H}_n^T = n \cdot \mathsf{I}_n,$$

where I_n is the identity matrix of size n, and H^T is the transpose of H.

Hadamard matrices exist for any power of 2.

Lemma 6.3.2 ([Wal23])**.** *For every positive integer $n > 1$, one can construct in time $2^{poly(n)}$ a $2^n \times 2^n$-Hadamard matrix.*

If H_n is an $n \times n$-Hadamard matrix, that we assume its rows and columns indexed by $[n]$, then for any two sets $A, B \subseteq [n]$, we denote by $\mathsf{H}_n[A, B]$ the submatrix of H_n, whose rows are indexed by A and columns are indexed by B. If all the entries of $\mathsf{H}_n[A, B]$ are equal, then we call $\mathsf{H}_n[A, B]$ a *monochromatic submatrix*. We will need the following to prove that if $\mathfrak{G}_{01} \in \mathrm{MP}(\mathfrak{H})$, then $\mathfrak{G} \in \mathrm{MP}_\star(\mathfrak{H})$.

Lemma 6.3.3 ([Alo86, PRS88])**.** *Let H_n be an $n \times n$-Hadamard matrix, whose rows and columns are indexed by $[n]$. Then, for any two disjoint sets $A, B \subseteq [n]$, with $|A| = |B| > \sqrt{n}$, the submatrix $\mathsf{H}_n[A, B]$ of H_n is not monochromatic.*

Remark. We use Hadamard matrices because the distribution of 1s is close to the uniform one. It is not the only class that can be used in this case. Every graph with a good expanding property, see [Alo86], would also be sufficient to prove Theorem 6.3.1. Hadamard matrices were the first ones that we found that had this property of being close to uniform, so we decided to use them in this thesis.

Proof of Theorem 6.3.1. Let us first explain the case of $\star$-graphs. Let $\mathfrak{H}$ be a $\star$-graph with $m = |H|$. We will show the P-time equivalence between the problems $\mathrm{MP}(\mathfrak{H})$ and $\mathrm{MP}_\star(\mathfrak{H})$.

Every 01-graph is also a $\star$-graph, so $\mathrm{MP}_{01}(\mathfrak{H})$ trivially reduces to $\mathrm{MP}_\star(\mathfrak{H})$. For the opposite direction, let us construct for every $\star$-graph $\mathfrak{G}$, a 01-graph $\mathfrak{G}_{01}$ such that $\mathfrak{G} \in \mathrm{MP}_\star(\mathfrak{H})$ if and only $\mathfrak{G}_{01} \in \mathrm{MP}_{01}(\mathfrak{H})$.

Let k be the smallest positive integer such that $2^k > 4m^2 + 1$, and let H_{2^k} be the Hadamard matrix ensured by Lemma 6.3.2. Let the domain of $\mathfrak{G}_{01}$ be the disjoint union $\bigsqcup_{g \in G} V_g$, where for all $g \in G$, $|V_g| = 2^k$. Let us enumerate the set V_g as $\{v_{g,1}, \ldots, v_{g,2^k}\}$, for each $g \in G$. Now, for each $v_{g_1,i} \in V_{g_1}$ and $v_{g_2,j} \in V_{g_2}$, $1 \leq i, j \leq 2^k$,

$$\mathsf{E}^{\mathfrak{G}_{01}}(v_{g_1,i}, v_{g_2,j}) = \begin{cases} \mathsf{E}^{\mathfrak{G}}(g_1, g_2) & \text{if } \mathsf{E}^{\mathfrak{G}}(g_1, g_2) \neq \star, \\ (\mathsf{H}_{2^k}[i,j] + 1)/2 & \text{otherwise.} \end{cases}$$

Observe that in the case $\mathsf{E}^{\mathfrak{G}}(g_1, g_2) = \star$, if $\mathsf{H}_{2^k}[i,j] = 1$, then $\mathsf{E}^{\mathfrak{G}_{01}}(v_{g_1,i}, v_{g_2,j}) = 1$, otherwise it is equal to 0.

By construction, there exists a surjective homomorphism $\pi \colon \mathfrak{G}_{01} \to \mathfrak{G}$ such that for all $g \in G$, $\pi(V_g) = g$. If there exists a homomorphism $\mathfrak{h} \colon \mathfrak{G} \to \mathfrak{H}$, then, by transitivity, $\mathfrak{h} \circ \pi : G_{01} \to H$ will be a homomorphism. Suppose that there exists a homomorphism $\mathfrak{h}_{01} \colon \mathfrak{G}_{01} \to \mathfrak{H}$. By pigeonhole principle, for every V_g, there is a set of size at least $\frac{|V_g|}{m}$ elements of V_g that are mapped by $\mathfrak{h}_{01}$ to the same element of $\mathfrak{H}$, and let us call it $A_g \subseteq V_g$, for each $g \in G$. Let us define a map $\mathfrak{h} \colon G \to H$ with $\mathfrak{h}(g) = \mathfrak{h}_{01}(A_g)$. Now, for every two elements $g_1, g_2 \in G$, if $\mathsf{E}^{\mathfrak{G}}(g_1, g_2) \in \{0, 1\}$, then for all $a_1 \in A_{g_1}, a_2 \in A_{g_2}$, $\mathsf{E}^{\mathfrak{G}_{01}}(a_1, a_2) = \mathsf{E}^{\mathfrak{G}}(g_1, g_2)$, so $\mathsf{E}^{\mathfrak{G}}(g_1, g_2) \preceq_\star \mathsf{E}^{\mathfrak{H}}\big(\mathfrak{h}(g_1), \mathfrak{h}(g_2)\big)$. If $\mathsf{E}^{\mathfrak{G}}(g_1, g_2) = \star$, then $\mathsf{H}_{2^k}[A_{g_1}, A_{g_2}]$ is of size at least $\frac{4m^2+1}{m} \times \frac{4m^2+1}{m}$, where A_g is identified with the set $\{i \in [2^k] \mid v_{g,i} \in A_g\}$. One checks easily that there are subsets B_1 of A_{g_1}, and B_2 of A_{g_2}, that do not intersect and both of size at least $\frac{|V_g|}{2m}$. Observe that

$$\frac{|V_g|}{2m} \geq \frac{4m^2 + 1}{2m} \geq \sqrt{4m^2 + 1}.$$

By Lemma 6.3.3, the submatrix $\mathsf{H}_{2^k}[B_1, B_2]$ is not monochromatic. This means that $\mathsf{E}^{\mathfrak{H}}\big(\mathfrak{h}(g_1), \mathfrak{h}(g_2)\big) = \star$ and $\mathfrak{h}$ is a homomorphism from $\mathfrak{G}$ to $\mathfrak{H}$, and we are done.

Let us now prove the general case. Let $\tau = \{\mathsf{R}_1, \ldots, \mathsf{R}_p\}$ a signature, with k_i the arity of R_i, for $i \in [p]$. We recall that $\mathrm{MP}^\tau(\mathfrak{H})$ trivially reduces to $\mathrm{MP}^\tau_\star(\mathfrak{H})$ as it is the same problem, but with restricted inputs.

For the other direction, we use the same technique as in the proof for the binary case, we construct a 01-structure $\mathfrak{G}_{01}$. For a given input $\mathfrak{G}$, and for any element $g \in G$, we introduce a set $V_g = \{v_{g,1}, \ldots, v_{g,2^k}\}$ of size 2^k such that $2^k \geq 4|H|^2 + 1$ and k is the smallest such positive integer. Let also H_{2^k} be the Hadamard matrix guaranteed by Lemma 6.3.2. Now, the domain G_{01} of $\mathfrak{G}_{01}$ is the disjoint union $\bigsqcup_{g \in G} V_g$. For each $\mathsf{R}_i \in \tau$ and for each tuple $(v_{g_1,i_1}, v_{g_2,i_2}, \ldots, v_{g_{k_i},i_{k_i}})$,

$$\mathsf{R}_i^{\mathfrak{G}_{01}}(v_{g_1,i_1}, \ldots, v_{g_{k_i},i_{k_i}}) = \begin{cases} \mathsf{R}_i^{\mathfrak{G}}(g_1, \ldots, g_{k_i}) & \text{if } \mathsf{R}_i^{\mathfrak{G}}(g_1, \ldots, g_{k_i}) \neq \star, \\ (\mathsf{H}_{2^k}[i_1, i_2] + 1)/2 & \text{otherwise.} \end{cases}$$

Suppose now that there exists $\mathfrak{h}_{01} \colon \mathfrak{G}_{01} \to \mathfrak{H}$. Then, by pigeonhole principle, in each set V_g, there is a set of size at least $\frac{|V_g|}{|H|}$ elements that are mapped to the same element

of H, denoted by A_g. Then, the sets A_{g_1} and A_{g_2} define a submatrix of H_{2^k} of size at least $2\sqrt{2^k}$ and thus it is not monochromatic by the same argument as in the proof for the binary case. □

6.3.2 Signature simplification

Recall that a *primitive-positive formula* $\varphi(x_1, \ldots, x_n)$ is a first-order formula (FO^τ) of the form

$$\exists x_{n+1}, \ldots, x_m.(\psi_1 \wedge \cdots \wedge \psi_l)$$

where each ψ_i is either $x_s = x_j$, true, or $\mathsf{R}(x_{i_1}, \ldots, x_{i_k}) = 1$.

Let $\tau = \{\mathsf{R}_1, \ldots, \mathsf{R}_n\}, \tau' = \{\mathsf{S}_1, \ldots, \mathsf{S}_m\}$ be two signatures, and $\mathfrak{A}, \mathfrak{A}'$ be relational τ- and τ'-structures over the same domain A. We say that $\mathfrak{A}$ *pp-defines* $\mathfrak{A}'$ if for every k-ary relation $\mathsf{S}_j^{\mathfrak{A}'}$ of $\mathfrak{A}'$ there exists a primitive-positive formula $\varphi_j \in \mathsf{FO}^\tau$ with k free variables such that for all $(a_1, \ldots, a_k) \in A^k$, $\mathsf{S}_j^{\mathfrak{A}'}(a_1, \ldots, a_n) = 1 \Leftrightarrow \mathfrak{A}' \models \varphi_j(a_1/x_1, \ldots, a_n/x_n)$.

Theorem 6.3.4. *[BK17] Suppose that a relational τ-structure $\mathfrak{A}$ pp-defines a relational τ'-structure $\mathfrak{A}'$. Then the problem $CSP^{\tau'}(\mathfrak{A}')$ reduces in P-time to $CSP^\tau(\mathfrak{A})$.*

6.3.3 From directed graphs to many relations

Let $\tau = \{\mathsf{R}_1, \ldots, \mathsf{R}_n\}$ be a finite signature with arities $k_1, \ldots, k_n$, and such that $k_1 \geq 2$. We show that the existence of a dichotomy for the class of problems $\mathrm{MP}_\star^\tau$ implies the existence of a dichotomy for the class of $\star$-graphs $\mathrm{MP}_\star$. Let $\gamma = \{\mathsf{E}(\cdot,\cdot)\}$ be the directed graph signature and let $\gamma_{\mathrm{CSP}} = \{\mathsf{E}_0(\cdot,\cdot), \mathsf{E}_1(\cdot,\cdot)\}$ be obtained from γ by the construction from Section 6.2.

Theorem 6.3.5. *For every $\star$-graph $\mathfrak{H}_\star$ there exists a $(\star, \tau)$-structure $\mathfrak{A}_\star$ such that the problems $MP_\star(\mathfrak{H}_\star)$ and $MP_\star^\tau(\mathfrak{A}_\star)$ are P-time equivalent.*

Proof. Let us recall from the Section 6.2 that there is a one-to-one correspondence between a $(\varnothing, \tau)$-structure $\mathfrak{A}_\varnothing$ and a relational τ_{CSP}-structure $\mathfrak{A}_{\mathrm{CSP}}$ such that for any two $(\varnothing, \tau)$-structures $\mathfrak{A}_\varnothing, \mathfrak{B}_\varnothing$:

$$\mathfrak{B}_\varnothing \to \mathfrak{A}_\varnothing \Leftrightarrow \mathfrak{B}_{\mathrm{CSP}} \to \mathfrak{A}_{\mathrm{CSP}}.$$

Now, let us consider a $\star$-graph $\mathfrak{H}_\star$ with its corresponding relational γ_{CSP}-structure $\mathfrak{H}_{\mathrm{CSP}}$. We construct the τ_{CSP}-structure $\mathfrak{A}_{\mathrm{CSP}}$ by the following pp-definition:

$$\forall j \in \{0,1\}\ \mathsf{R}_{1,j}^{\mathfrak{A}_{\mathrm{CSP}}}(x_1, \ldots, x_{k_1}) = 1 \Leftrightarrow \mathsf{E}_j^{\mathfrak{H}_{\mathrm{CSP}}}(x_1, x_2) = 1; \tag{6.2}$$

$$\forall i > 1, j \in \{0,1\}\ \mathsf{R}_{i,j}^{\mathfrak{A}_{\mathrm{CSP}}}(x_1, \ldots, x_{k_i}) = 1 \Leftrightarrow \mathsf{true}. \tag{6.3}$$

Observe that the relational γ_{CSP}-structure $\mathfrak{H}_{\mathrm{CSP}}$ is also pp-definable from the relational τ_{CSP}-structure $\mathfrak{A}_{\mathrm{CSP}}$:

$$\mathsf{E}^{\mathfrak{H}_{\mathrm{CSP}}}(x_1, x_2) = 1 \Leftrightarrow \exists x_3, \ldots, x_{k_1}.\mathsf{R}_1^{\mathfrak{A}_{\mathrm{CSP}}}(x_1, \ldots, x_{k_1}). \tag{6.4}$$

Now consider any $\star$-graph $\mathfrak{G}_\star$. Every $\star$-graph is also a $\varnothing$-graph, so there is a relational γ_{CSP}-structure $\mathfrak{G}_{\mathrm{CSP}}$ such that $\mathfrak{G}_\star \to \mathfrak{H}_\star$ if and only if $\mathfrak{G}_{\mathrm{CSP}} \to \mathfrak{H}_{\mathrm{CSP}}$. By the pp-definability in eq. (6.4) and Theorem 6.3.4, we construct a relational τ_{CSP}-structure $\mathfrak{B}_{\mathrm{CSP}}$

such that $\mathfrak{G}_{\mathrm{CSP}} \to \mathfrak{H}_{\mathrm{CSP}}$ if and only if $\mathfrak{B}_{\mathrm{CSP}} \to \mathfrak{A}_{\mathrm{CSP}}$. From $\mathfrak{B}_{\mathrm{CSP}}$ we obtain a $(\star, \tau)$-structure $\mathfrak{B}_\star$ such that $\mathfrak{B}_{\mathrm{CSP}} \to \mathfrak{A}_{\mathrm{CSP}}$ if and only if $\mathfrak{B}_\star \to \mathfrak{A}_\star$. Observe that, because $\mathfrak{G}_\star$ is a $\star$-graph, in $\mathfrak{G}_{\mathrm{CSP}}$ for any $(x, y) \in G^2$, we have either $\mathsf{E}_0^{\mathfrak{G}_{\mathrm{CSP}}}(x, y) = 1$ or $\mathsf{E}_1^{\mathfrak{G}_{\mathrm{CSP}}}(x, y) = 1$; thus in $\mathfrak{B}_{\mathrm{CSP}}$ any relation other than R_1 is interpreted trivially and for each tuple $\mathbf{x} \in B^{k_1}$ either $\mathsf{R}_{1,0}^{\mathfrak{B}_{\mathrm{CSP}}}(\mathbf{x}) = 1$ or $\mathsf{R}_{1,1}^{\mathfrak{B}_{\mathrm{CSP}}}(\mathbf{x}) = 1$. So, $\mathfrak{B}_\star$ is indeed a $(\star, \tau)$-structure, that finishes the reduction from $\mathrm{MP}_\star(\mathfrak{H}_\star)$ to $\mathrm{MP}_\star^\tau(\mathfrak{A}_\star)$.

For the other direction, consider any $(\star, \tau)$-structure $\mathfrak{B}_\star$. Similarly, we construct a relational τ_{CSP}-structure $\mathfrak{B}_{\mathrm{CSP}}$, and by the pp-definition in eqs. (6.2) and (6.3), we can compute a relational γ_{CSP}-structure $\mathfrak{G}_{\mathrm{CSP}}$ such that $\mathfrak{G}_{\mathrm{CSP}} \to \mathfrak{H}_{\mathrm{CSP}}$ if and only if $\mathfrak{B}_\star \to \mathfrak{A}_\star$, and then a $\star$-graph $\mathfrak{G}_\star$ such that $\mathfrak{B}_\star \to \mathfrak{A}_\star$ if and only if $\mathfrak{G}_\star \to \mathfrak{H}_\star$. With similar arguments as in the other direction, we can prove that $\mathfrak{G}_\star$ is indeed a $\star$-graph. We have then shown that $\mathrm{MP}_\star(\mathfrak{G}_\star) \equiv_p \mathrm{MP}_\star^\tau(\mathfrak{A}_\star)$ □

One notices that the proof of Theorem 6.3.5 is still correct if we replace γ by any relation R of arity $\ell \geq 2$, we require in this case that R_1 has arity at least ℓ.

6.3.4 From many relations to one

Suppose that $\tau = \{\mathsf{R}_1, \ldots, \mathsf{R}_p\}$, R_i has arity k_i, let $k = \max_i k_i$. We show that for any such τ there exists $\tilde{\tau} = \{\mathsf{R}\}$ with R of arity $k + p - 1$ such that for any $(\star, \tau)$-structure $\mathfrak{A}$ there exists a $(\star, \tilde{\tau})$-structure $\tilde{\mathfrak{A}}$ such that $\mathrm{MP}_\star^\tau(\mathfrak{A})$ and $\mathrm{MP}_\star^{\tilde{\tau}}(\tilde{\mathfrak{A}})$ are P-time equivalent. This means that $\mathrm{MP}_\star^\tau \subseteq_p \mathrm{MP}_\star^{\tilde{\tau}}$.

Now we will describe how the $\tilde{\tau}$-structure $\tilde{\mathfrak{A}}$ is constructed. If A is the domain of $\mathfrak{A}$, then the domain $\tilde{A}$ of $\tilde{\mathfrak{A}}$ is $\tilde{A} = A \sqcup \{c_A\}$, with c_A a new element. The relation $\mathsf{R}^{\tilde{\mathfrak{A}}}$ is defined as follows:

- Let $\mathcal{A}_1 = \{\tilde{\mathbf{t}} = (\underbrace{c_A, \ldots, c_A}_{i-1}, \mathbf{t}, \underbrace{c_A, \ldots, c_A}_{k+p-k_i-i}) \mid \mathsf{R}_i \in \tau, \mathbf{t} \in A^{k_i}\}$, and

$$\text{for all, } \tilde{\mathbf{t}} \in \mathcal{A}_1, \ \mathsf{R}^{\tilde{\mathfrak{A}}}(\tilde{\mathbf{t}}) = \mathsf{R}_i^{\mathfrak{A}}(\mathbf{t}); \tag{6.5}$$

- Let $\mathcal{A}_2 = \{(c_A, \ldots, c_A)\}$, then for all $\tilde{\mathbf{t}} \in \mathcal{A}_2 \colon \mathsf{R}^{\tilde{\mathfrak{A}}}(\tilde{\mathbf{t}}) = 1$;

- Let $\mathcal{A}_3 = \tilde{A}^{k+p-1} \setminus (\mathcal{A}_1 \sqcup \mathcal{A}_2)$, then for all $\tilde{\mathbf{t}} \in \mathcal{A}_3 \colon \mathsf{R}^{\tilde{\mathfrak{A}}}(\tilde{\mathbf{t}}) = 0$.

Now we will prove one direction of the P-time equivalence. The size of $\tilde{\mathfrak{A}}$ is polynomial in $|A|$, so the construction takes P-time, and below we show that $\mathfrak{B} \to \mathfrak{A} \Leftrightarrow \tilde{\mathfrak{B}} \to \tilde{\mathfrak{A}}$.

Lemma 6.3.6. *$MP_\star^\tau(\mathfrak{A})$ reduces in polynomial time to $MP_\star^{\tilde{\tau}}(\tilde{\mathfrak{A}})$.*

Proof. Let $\mathfrak{B}$ be an input instance of the problem $\mathrm{MP}_\star^\tau(\mathfrak{A})$. Assume that there is $\mathsf{h} \colon \mathfrak{B} \to \mathfrak{A}$ – a homomorphism. We will show that $\tilde{\mathsf{h}} \colon \tilde{B} \to \tilde{A}$ such that $\tilde{\mathsf{h}}(c_B) = c_A$ and for all $x \in B \setminus \{c_B\} \colon \tilde{\mathsf{h}}(x) = \mathsf{h}(x)$, is a homomorphism.

Recall that $\tilde{B}^{k+p-1} = \mathcal{B}_1 \sqcup \mathcal{B}_2 \sqcup \mathcal{B}_3$. Consider $\tilde{\mathbf{t}} = (c_B, \ldots, c_B, \mathbf{t}, c_B, \ldots, c_B) \in \mathcal{B}_1$, where $\mathbf{t} = (b_1, \ldots, b_{k_i}) \in B^{k_i}$ for $\mathsf{R}_i \in \tau$. Then $\tilde{\mathsf{h}}(\tilde{\mathbf{t}}) = (c_A, \ldots, c_A, \mathsf{h}(\mathbf{t}), c_A, \ldots, c_A) \in \mathcal{A}_1$. As h is a homomorphism, we have that by eq. (6.5):

$$\mathsf{R}^{\tilde{\mathfrak{B}}}(\tilde{\mathbf{t}}) = \mathsf{R}_i^{\mathfrak{B}}(\mathbf{t}) \preceq_\star \mathsf{R}_i^{\mathfrak{A}}\big(\mathsf{h}(\mathbf{t})\big) = \mathsf{R}^{\tilde{\mathfrak{A}}}\big(\tilde{\mathsf{h}}(\tilde{\mathbf{t}})\big).$$

For $\tilde{\mathbf{t}} \in \mathcal{B}_2$, we have that $\tilde{\mathsf{h}}(\tilde{\mathbf{t}}) = (c_A, \dots, c_A)$, so $\mathsf{R}^{\tilde{\mathfrak{A}}}\big(\tilde{\mathsf{h}}(\tilde{\mathbf{t}})\big) = \mathsf{R}^{\tilde{\mathfrak{B}}}(\tilde{\mathbf{t}}) = 1$. Let us consider a tuple $\tilde{\mathbf{t}} = (x_1, \dots, x_{k+p-1}) \in \mathcal{B}_3$. We know that $\tilde{\mathsf{h}}(x) = c_A$ if and only if $x = c_B$, thus $\tilde{\mathsf{h}}(\tilde{\mathbf{t}}) \in \mathcal{A}_3$. Then $\mathsf{R}^{\tilde{\mathfrak{A}}}\big(\tilde{\mathsf{h}}(\tilde{\mathbf{t}})\big) = \mathsf{R}^{\tilde{\mathfrak{B}}}(\tilde{\mathbf{t}}) = 0$. We have shown that $\tilde{\mathsf{h}}$ is a homomorphism.

Assume that there is $\tilde{\mathsf{h}}\colon \tilde{\mathfrak{B}} \to \tilde{\mathfrak{A}}$ – a homomorphism. We know that $x = c_B$ if and only if $\mathsf{R}^{\tilde{\mathfrak{B}}}(x, \dots, x) = 1$, and otherwise $\mathsf{R}^{\tilde{\mathfrak{B}}}(x, \dots, x) = 0$. We also know the same thing for $\tilde{\mathfrak{A}}$. Thus, $x = c_B$ if and only if $\tilde{\mathsf{h}}(x) = c_A$. This allows us to correctly construct $\mathsf{h}\colon \mathfrak{B} \to \mathfrak{A}$, where for all $x \in B$, $\mathsf{h}(x) = \tilde{\mathsf{h}}(x)$.

For any $\mathsf{R}_i \in \tau$ and $\mathbf{t} \in B^{k_i}$, $\mathbf{t}$ is associated with $\tilde{\mathbf{t}} = (c_B, \dots, c_B, \mathbf{t}, c_B \dots, c_B) \in \mathcal{B}_1$ and its image $\mathsf{h}(\mathbf{t}) \in A^{k_i}$ is associated with $\tilde{\mathsf{h}}(\tilde{\mathbf{t}}) = (c_A, \dots, c_A, \mathsf{h}(\mathbf{t}), c_A, \dots, c_A) \in \mathcal{A}_1$. We know that by the construction of $\tilde{\mathfrak{A}}$ and $\tilde{\mathfrak{B}}$, and by eq. (6.5):

$$\mathsf{R}_i^{\mathfrak{B}}(\mathbf{t}) = \mathsf{R}^{\tilde{\mathfrak{B}}}(\tilde{\mathbf{t}}) \preceq_\star \mathsf{R}^{\tilde{\mathfrak{A}}}\big(\tilde{\mathsf{h}}(\tilde{\mathbf{t}})\big) = \mathsf{R}_i^{\mathfrak{A}}\big(\mathsf{h}(\mathbf{t})\big).$$

So, h is a homomorphism and $\mathrm{MP}^{\tau}_{\star}(\mathfrak{A})$ reduces to $\mathrm{MP}^{\tilde{\tau}}_{\star}(\tilde{\mathfrak{A}})$. □

Now we have to find in polynomial time for any input $(\star, \tilde{\tau})$-structure $\tilde{\mathfrak{G}}$ of $\mathrm{MP}^{\tilde{\tau}}_{\star}(\mathfrak{A})$ a $(\star, \tau)$-structure $\mathfrak{B}$ such that

$$\tilde{\mathfrak{G}} \to \tilde{\mathfrak{A}} \Leftrightarrow \mathfrak{B} \to \mathfrak{A}.$$

Lemma 6.3.7. *$MP^{\tilde{\tau}}_{\star}(\tilde{\mathfrak{A}})$ reduces in polynomial time to $MP^{\tau}_{\star}(\mathfrak{A})$.*

Proof. Let $\tilde{\mathfrak{G}}$ be an input instance of $\mathrm{MP}^{\tilde{\tau}}_{\star}(\tilde{\mathfrak{A}})$. Firstly, for any element $x \in \tilde{G}$, we check whether $\mathsf{R}^{\tilde{\mathfrak{G}}}(x, \dots, x) = \star$. If such an x exists, then we cannot map $\tilde{\mathfrak{G}}$ to $\tilde{\mathfrak{A}}$ as for all $y \in \tilde{A}$ we have that $\mathsf{R}^{\tilde{\mathfrak{A}}}(y, \dots, y) \in \{0, 1\}$. This can be checked in time linear in $|\tilde{G}|$. In this case, we output some fixed NO input instance of $\mathrm{MP}^{\tau}_{\star}(\mathfrak{A})$, e.g., some $\mathfrak{B}$ where there is $b \in B$ and $\mathsf{R}_i^{\mathfrak{B}}(b, \dots, b) = \star$ for all $\mathsf{R}_i \in \tau$.

Now we can assume that, for all $x \in \tilde{G}$, $\mathsf{R}^{\tilde{\mathfrak{G}}}(x, \dots, x) \in \{0, 1\}$. We divide the elements of $\tilde{G}$ into two classes: $\tilde{G} = C_1 \sqcup C_0$ by the following rule:

$$\text{for all } x \in \tilde{G},\ x \in C_i \Leftrightarrow \mathsf{R}^{\tilde{\mathfrak{G}}}(x, \dots, x) = i. \tag{6.6}$$

Observe that if there exists a homomorphism $\mathsf{h}\colon \tilde{\mathfrak{G}} \to \tilde{\mathfrak{A}}$, then for all $x \in \tilde{G}\colon \mathsf{h}(x) = c_A \Leftrightarrow x \in C_1$. We are going to construct a $\tilde{\tau}$-structure $\tilde{\mathfrak{B}}$ with the following properties:

1. $\tilde{\mathfrak{G}} \to \tilde{\mathfrak{B}}$;
2. $\tilde{\mathfrak{G}} \to \tilde{\mathfrak{A}} \Leftrightarrow \tilde{\mathfrak{B}} \to \tilde{\mathfrak{A}}$;
3. Either we can check in P-time that $\tilde{\mathfrak{B}} \not\to \tilde{\mathfrak{A}}$ or there exists a τ-structure $\mathfrak{B}$ such that $\tilde{\mathfrak{B}}$ can be obtained from $\mathfrak{B}$ by the construction described above in this section.

The domain $\tilde{B} := C_0 \sqcup \{c_B\}$. The element c_B should be considered as the result of identifying all in C_1 into a single element, namely c_B.

Let us consider a tuple $\tilde{\mathbf{t}} = (b_1, \dots, b_{k+p-1}) \in \tilde{B}^{k+p-1}$. Denote by $\mathcal{I}_{\tilde{\mathbf{t}}} \subseteq [k+p-1]$ the set of indices such that $b_i = c_B$ in $\tilde{\mathbf{t}}$. Denote by $\mathcal{C}_{\tilde{\mathbf{t}}} \subseteq G^{k+p-1}$ the class of all tuples $(x_1, \dots, x_{k+p-1}) \in \tilde{G}^{k+p-1}$ such that

$$\forall i \in [k+p-1]\colon\ (i \in \mathcal{I}_{\tilde{\mathbf{t}}} \Rightarrow x_i \in C_1)\ \wedge (i \notin \mathcal{I}_{\tilde{\mathbf{t}}} \Rightarrow b_i = x_i)\,.$$

The interpretation $\mathsf{R}^{\tilde{\mathfrak{B}}}$ is defined as follows, here $\bigvee$ denotes the least upper bound of two element with respect to the ordering $\preceq_\star$:

$$\mathsf{R}^{\tilde{\mathfrak{B}}}(\tilde{\mathbf{t}}) = \bigvee_{(x_1,\dots,x_{k+p-1})\in\mathcal{C}_{\tilde{\mathbf{t}}}} \mathsf{R}^{\tilde{\mathfrak{G}}}(x_1,\dots,x_{k+p-1}). \tag{6.7}$$

Observe that we can construct $\tilde{\mathfrak{B}}$ in time polynomial in the size of the input $\tilde{\mathfrak{G}}$.

Let us check the property 1, that $\tilde{\mathfrak{G}} \to \tilde{\mathfrak{B}}$. Let us consider a map $\pi\colon \tilde{G} \to \tilde{B}$ s.t.

- if $x \in C_1$, then $\pi(x) = c_B$;
- if $x \in C_0$, then $\pi(x) = x$.

Consider a tuple $\tilde{\mathbf{x}} = (x_1,\dots,x_{k+p-1}) \in \tilde{G}^{k+p-1}$ and $\pi(\tilde{\mathbf{x}}) = (b_1,\dots,b_{k+p-1}) \in \tilde{B}^{k+p-1}$ where

- $b_i = c_B$, if $x_i \in C_1$;
- $b_i = x_i$, otherwise.

As $\tilde{\mathbf{x}} \in \mathcal{C}_{\pi(\tilde{\mathbf{x}})}$, by eq. (6.7) we have $\mathsf{R}^{\tilde{\mathfrak{G}}}(\tilde{\mathbf{x}}) \preceq_\star \mathsf{R}^{\tilde{\mathfrak{B}}}\big(\pi(\tilde{\mathbf{x}})\big)$. This proves that π is a homomorphism.

Let us check the property 2, that $\tilde{\mathfrak{G}} \to \tilde{\mathfrak{A}} \Leftrightarrow \tilde{\mathfrak{B}} \to \tilde{\mathfrak{A}}$. As $\tilde{\mathfrak{G}} \to \tilde{\mathfrak{B}}$, we need to show only one direction, *i.e.*, $\Rightarrow$. Assume that there is $\mathsf{h}_G\colon \tilde{\mathfrak{G}} \to \tilde{\mathfrak{A}}$ – a homomorphism. Observe that, for all x, $x \in C_1 \Leftrightarrow \mathsf{h}_G(x) = c_A$. We define a map h_B as follows:

- if $x = c_B$, then $\mathsf{h}_B(x) = c_A$;
- if $x \neq c_B$, then $\mathsf{h}_B(x) = \mathsf{h}_G(x)$.

Consider a tuple $\tilde{\mathbf{t}} = (b_1,\dots,b_{k+p-1}) \in \tilde{B}^{k+p-1}$. Observe that $\mathsf{h}_B(\tilde{\mathbf{t}}) = \mathsf{h}_G(\mathcal{C}_{\tilde{\mathbf{t}}})$ that is any tuple from $\mathcal{C}_{\tilde{\mathbf{t}}}$ is mapped to $\mathsf{h}_B(\tilde{\mathbf{t}})$ by h_G. We know that

$$\mathsf{R}^{\tilde{\mathfrak{A}}}\big(\mathsf{h}_B(\tilde{\mathbf{t}})\big) \succeq_\star \mathsf{R}^{\tilde{\mathfrak{G}}}(x_1,\dots,x_{k+p-1})$$

for all $(x_1,\dots,x_{k+p-1}) \in \mathcal{C}_{\tilde{\mathbf{t}}}$. Thus,

$$\mathsf{R}^{\tilde{\mathfrak{A}}}\big(\mathsf{h}_B(\tilde{\mathbf{t}})\big) \succeq_\star \bigvee_{(x_1,\dots,x_{k+p-1})\in\mathcal{C}_{\tilde{\mathbf{t}}}} \mathsf{R}^{\tilde{\mathfrak{G}}}(x_1,\dots,x_{k+p-1}) = \mathsf{R}^{\tilde{\mathfrak{B}}}(\tilde{\mathbf{t}}),$$

where $\bigvee$ denote the least upper bound of the elements. This shows that h_B is a homomorphism.

Finally, we need to check the property 3 to finish the proof. Recall that we split all the tuples $(b_1,\dots,b_{k+p-1}) \in \tilde{B}^{k+p-1}$ into three classes: $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$. Observe that for any homomorphism $\mathsf{h} : \tilde{\mathfrak{B}} \to \tilde{\mathfrak{A}}$, we have that for any $x \in \tilde{B}$, $(x = c_B \Leftrightarrow \mathsf{h}(x) = c_A)$; then, for any $j \in [3]$, $\mathsf{h}(\mathcal{B}_j) \subseteq \mathcal{A}_j$. At first, we look at the tuple $\tilde{\mathbf{t}} = (c_B,\dots,c_B) \in \mathcal{B}_2$. By eqs. (6.6) and (6.7), we know that $\mathsf{R}^{\tilde{\mathfrak{B}}}(\tilde{\mathbf{t}}) \succeq_\star 1$. If $\mathsf{R}^{\tilde{\mathfrak{B}}}(\tilde{\mathbf{t}}) = \star$, then we output some fixed NO input instance of $\mathrm{MP}^\tau_\star(\mathfrak{A})$ for $\tilde{\mathfrak{G}}$. If $\mathsf{R}^{\tilde{\mathfrak{B}}}(\tilde{\mathbf{t}}) = 1$, then we continue.

Now, we look at every tuple $\tilde{\mathbf{t}} \in \mathcal{B}_3$ and check whether $\mathsf{R}^{\tilde{\mathfrak{B}}}(\tilde{\mathbf{t}}) = 0$. If there exists $\tilde{\mathbf{t}} \in \mathcal{B}_3$ such that $\mathsf{R}^{\tilde{\mathfrak{B}}}(\tilde{\mathbf{t}}) \neq 0$, then we output some fixed NO input instance of $\mathrm{MP}^\tau_\star(\mathfrak{A})$ for $\tilde{\mathfrak{G}}$. If, for all tuples of $\mathcal{B}_3$, we have that $\mathsf{R}^{\tilde{\mathfrak{B}}}(\tilde{\mathbf{t}}) = 0$, then we continue. We can do all these checks in time polynomial in $|\tilde{G}|$.

Now we can assume that $\mathsf{R}^{\tilde{\mathfrak{B}}}(\tilde{\mathbf{t}}_2) = 1$ and $\mathsf{R}^{\tilde{\mathfrak{B}}}(\tilde{\mathbf{t}}_3) = 0$, for all $\tilde{\mathbf{t}}_2 \in \mathcal{B}_2, \tilde{\mathbf{t}}_3 \in \mathcal{B}_3$. We are ready to construct the $(\star,\tau)$-structure $\mathfrak{B}$:

- the domain B of $\mathfrak{B}$ is $\tilde{B} \setminus \{c_B\}$;
- for any relation $\mathsf{R}_i \in \tau$ and any tuple $\mathbf{t} = (b_1, \ldots, b_{k_i}) \in B^{k_i}$ it is interpreted as follows:

$$\mathsf{R}_i^{\mathfrak{B}}(\mathbf{t}) = \mathsf{R}^{\tilde{\mathfrak{B}}}(\underbrace{c_B, \ldots, c_B}_{i-1}, \mathbf{t}, \underbrace{c_B, \ldots, c_B}_{k+p-k_i-i}). \tag{6.8}$$

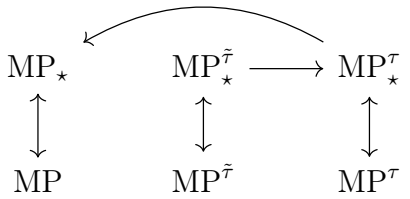
If we apply the $\tilde{(\cdot)}$-transformation to this $(\star, \tau)$-structure $\mathfrak{B}$, then we will get $\tilde{\mathfrak{B}}$, because for all tuples of $\mathcal{B}_2, \mathcal{B}_3$: $\mathsf{R}^{\tilde{\mathfrak{B}}}$ always has values 1 and 0 correspondingly, and for all tuples of $\mathcal{B}_1$ there is a one-to-one correspondence with the tuples of all $\mathsf{R}_i \in \tau$, the equivalence of their values is provided by eqs. (6.5) and (6.8). By Lemma 6.3.6, $\mathfrak{B} \to \mathfrak{A}$ if and only if $\tilde{\mathfrak{B}} \to \tilde{\mathfrak{A}}$. We have shown that, for any $(\star, \tilde{\tau})$-structure $\tilde{\mathfrak{G}}$, we can find in time polynomial in $|\tilde{G}|$ a $(\star, \tau)$-structure $\mathfrak{B}$ such that $\tilde{\mathfrak{G}} \to \tilde{\mathfrak{A}} \Leftrightarrow \tilde{\mathfrak{B}} \to \tilde{\mathfrak{A}}$. Thus $\mathrm{MP}^{\tilde{\tau}}_{\star}(\tilde{\mathfrak{A}})$ reduces in polynomial time to $\mathrm{MP}^{\tau}_{\star}(\mathfrak{A})$. □

Lemma 6.3.6 and Lemma 6.3.7 provide the following statement.

Theorem 6.3.8. *If the class of problems $MP^{\tilde{\tau}}_{\star}$ has a dichotomy, then the class $MP^{\tau}_{\star}$ has a dichotomy.*

Observe that in order to prove the other direction, for every $(\star, \tilde{\tau})$-structure $\mathfrak{A}$, we have to find a $(\star, \tau)$-structure $\hat{\mathfrak{A}}$ such that $\mathrm{MP}^{\tau}_{\star}(\hat{\mathfrak{A}})$ and $\mathrm{MP}^{\tilde{\tau}}_{\star}(\mathfrak{A})$ are P-time equivalent. We show in the next section the difficulties to obtain such a reduction.

The dichotomy implications between the considered classes are displayed on the diagram below. Each arrow shows an implication of the existence of a dichotomy, *i.e.*, if the class at the tail has a dichotomy, then the class at the head has it. The vertical ones are shown in Section 6.3.1, and the horizontal ones are shown in Section 6.3.2. One can see now that the existence of a dichotomy for $\mathrm{MP}^{\tilde{\tau}}$ implies such existence for all other classes considered there.



6.3.5 From one relation to directed graphs

We do not prove that for any $(\star, \tilde{\tau})$-structure $\mathfrak{H}$, with $\tilde{\tau}$ consisting of one single symbol, there exists a $\star$-graph $\mathfrak{H}_2$ such that $\mathrm{MP}^{\tilde{\tau}}_{\star}(\mathfrak{H}) \equiv_p \mathrm{MP}_{\star}(\mathfrak{H}_2)$. However, we will discuss some necessary conditions for the existence of such a reduction. And also we will discuss why approaches that are similar to the one used in [BDJN15, FV98] cannot be applied to the homomorphism problems considered in this thesis, in particular, Matrix Partition Problems.

For the simplicity of the notations, we will consider the reduction from $\tilde{\tau} = \{\mathsf{R}(\cdot, \cdot, \cdot)\}$ to $\star$-graphs. In the very beginning, we are going to show that if there exists such a correspondence between $(\star, \tilde{\tau})$-structures and $\star$-graphs, then the size of the domain of the constructed $\star$-graph must be significantly greater than the one of the corresponding ternary $\star$-structure.

Let $\mathcal{G}_2^n$ be the class of all $\star$-graphs on n elements that are cores and pairwise not homomorphically equivalent. Recall that a core $\mathfrak{G} \in \mathcal{G}_2^n$ cannot have an element x with $\mathsf{E}^{\mathfrak{G}}(x,x) = \star$. Let $\mathcal{G}_3^n$ be the class of $(\star, \tilde{\tau})$-structures with the same property.

Lemma 6.3.9. *For every positive integer n, $|\mathcal{G}_3^n| \geq |\mathcal{G}_2^n| \cdot 3^{n(n-1)^2-1}$.*

Proof. We suppose that all $\mathfrak{G} \in \mathcal{G}_2^n$ have the same domain $\{a, x_1, \ldots, x_{n-1}\}$, we fix one element a, and linearly order the elements from $\{x_1, \ldots, x_{n-1}\}$ with $x_i < x_j$ if $i < j$. Let $\mathfrak{G} \in \mathcal{G}_2^n$. We will construct a family of $(\star, \tilde{\tau})$-structures $\mathcal{G}_{\mathfrak{G}}$ of size $3^{n(n-1)^2-1}$ such that every two $(\star, \tilde{\tau})$-structures from there will not be homomorphically equivalent. We will construct such a class for every $\star$-graph in $\mathcal{G}_2^n$, and then show that any two structures from different classes will not be homomorphically equivalent as well.

Any $\mathfrak{G}_3 \in \mathcal{G}_{\mathfrak{G}}$ must satisfy the following properties:

- the domain is the same as the one of $\mathfrak{G}$: $G_3 = G = \{a, x_1, \ldots, x_{n-1}\}$;
- for all $x, y \in G_3$: $\mathsf{R}^{\mathfrak{G}_3}(a, x, y) = \mathsf{E}^{\mathfrak{G}}(x, y)$ – we define all the relations of $\mathfrak{G}$ using only the triples that have a on the first coordinate;
- for all $x \in G_3 \setminus \{a\}$: $\mathsf{R}^{\mathfrak{G}_3}(x, x, x) = 1 - \mathsf{E}^{\mathfrak{G}}(a, a)$ – all the elements other than a have the value on the loop, that is different from the loop $\mathsf{R}^{\mathfrak{G}_3}(a, a, a) = \mathsf{E}^{\mathfrak{G}}(a, a)$; the *loop property*
- for all $i, j \in [n]$, $\mathsf{R}^{\mathfrak{G}_3}(x_i, x_j, a) = \star$ if $i < j$ and $\mathsf{R}^{\mathfrak{G}_3}(x_i, x_j, a) = 0$ if $i \geq j$; the *linear ordering property*
- fix one $(x_{i_1}, x_{i_2}, x_{i_3}) \in (G_3 \setminus \{a\})^3$, such that the number i_1, i_2, i_3 are not all equal, and set $\mathsf{R}^{\mathfrak{G}_3}(x_{i_1}, x_{i_2}, x_{i_3}) = \star$.

The values are restricted for the n^2 tuples that are associated with the arcs of $\mathfrak{G}$, for the $(n-1)^2$ tuples of the linear ordering, and for the $(n-1)$ loops, and for the triple $(x_{i_1}, x_{i_2}, x_{i_3})$. For any other triple $(x, y, z) \in G^3$, there is no restriction on the value of $\mathsf{R}^{\mathfrak{G}_3}$ among $\{0, 1, \star\}$. Thus,

$$|\mathcal{G}_{\mathfrak{G}}| = 3^{n^3-n^2-(n-1)^2-(n-1)-1} = 3^{n(n-1)^2-1}.$$

Let us consider $\mathfrak{A}, \mathfrak{B} \in \mathcal{G}_{\mathfrak{G}}$, suppose that there is a homomorphism $\mathsf{h}\colon \mathfrak{A} \to \mathfrak{B}$, then $\mathsf{h}(a) = a$ and for all $x \neq a$, $\mathsf{h}(x) \neq a$ by the loop property of $\mathfrak{G}_3$. Also, by the linear ordering property, we have that for all $x \in A$, $\mathsf{h}(x) = x$. But these two structures differ on at least one tuple, this is a contradiction.

Let us consider $\mathfrak{A}_1 \in \mathcal{G}_{\mathfrak{G}_1}, \mathfrak{A}_2 \in \mathcal{G}_{\mathfrak{G}_2}$ – structures from different classes of two $\star$-graphs $\mathfrak{G}_1, \mathfrak{G}_2$ that are not hom-equivalent. Suppose that there is no homomorphism from $\mathfrak{G}_1$ to $\mathfrak{G}_2$ and there is a homomorphism $\mathsf{h}\colon \mathfrak{A}_1 \to \mathfrak{A}_2$. If $\mathsf{E}^{\mathfrak{G}_1}(a, a) \neq \mathsf{E}^{\mathfrak{G}_2}(a, a)$, then by the loop property, for all $x \in A_1 \setminus \{a\}$, $\mathsf{h}(x) = a$, this is a contradiction as $\mathsf{E}^{\mathfrak{G}_2}(a, a) \neq \star$ on one hand and $\star = \mathsf{R}^{\mathfrak{A}_1}(x_{i_1}, x_{i_2}, x_{i_3}) \preceq_\star \mathsf{R}^{\mathfrak{A}_2}(a, a, a) = \mathsf{E}^{\mathfrak{G}_2}(a, a)$ on the other hand. Thus, we assume that $\mathsf{E}^{\mathfrak{G}_1}(a, a) = \mathsf{E}^{\mathfrak{G}_2}(a, a)$ and that $\mathsf{h}(a) = a$ and, for all $x \neq a$, $\mathsf{h}(x) \neq a$ (again by the loop property), but, by the linear ordering property, we must have that $\mathsf{h}(x) = x$. The homomorphism h implies that the identity mapping on the set $G = \{a, x_1, \ldots, x_{n-1}\}$ is a homomorphism from $\mathfrak{G}_1$ to $\mathfrak{G}_2$ that is a contradiction as $\mathfrak{G}_1$ and $\mathfrak{G}_2$ are pairwise not homomorphically equivalent.

This proves that we are able to construct at least $|\mathcal{G}_2^n| \cdot 3^{n(n-1)^2-1}$ $(\star, \tilde{\tau})$-structures such that any two of them are not homomorphically equivalent. □

This lemma ensures that when we make a correspondence between ternary and binary structures, in the general case we need to add a lot of elements to the binary one.

Corollary 6.3.10. *Let $n, m \in \mathbb{N}$. If $|\mathcal{G}_3^n| < |\mathcal{G}_2^m|$, then $m > \sqrt{n(n-1)^2 - 1}$.*

Proof. The number of all possible ways to assign one of three values to each of the m^2 pairs equals 3^{m^2}. Then, by Lemma 6.3.9:

$$3^{m^2} \geq |\mathcal{G}_2^m| > |\mathcal{G}_3^n| \geq 3^{n(n-1)^2-1}|\mathcal{G}_2^n| \geq 3^{n(n-1)^2-1} \Rightarrow m^2 > n(n-1)^2 - 1.$$

□

We will argue that all the approaches similar to the one used in [BDJN15] do not work for the case of $\mathrm{MP}_\star$ (equivalently MP, see Section 6.3.1). Such an approach can be described by these steps: for any $(\star, \tilde{\tau})$-structure $\mathfrak{H}_3$ the corresponding $\star$-graph $\mathfrak{H}_2$ is constructed as follows:

1. take the same domain $H_2 = H_3$ and
2. substitute every tuple $\mathsf{R}^{\mathfrak{H}_3}(x_1, x_2, x_3)$ by a $\star$-graph $\mathfrak{T}^v_{x_1x_2x_3}$ for $v \in \{0, 1, \star\}$ that contains only these 3 elements x_1, x_2, x_3 among those of H_3. Letter $\mathfrak{T}$ stands for "tuple" and the superscript v is for "value". It is required that for two different tuples $\mathbf{t_1}$ and $\mathbf{t_2}$, the domains of $\mathfrak{T}^{v_1}_{\mathbf{t_1}}$ and of $\mathfrak{T}^{v_1}_{\mathbf{t_2}}$ intersect only on H_3.

So, the domain of $\mathfrak{H}_2$ is the union of the domain H_3 of the $(\star, \tilde{\tau})$-structure $\mathfrak{H}_3$ and the domains of all the $\star$-graphs $\mathfrak{T}^v$ that represent the tuples of $\mathfrak{H}_3$:

$$H_2 = H_3 \cup \bigcup_{(x_1,x_2,x_3)\in H_3^3,\ \mathsf{R}^{\mathfrak{H}_3}(x_1,x_2,x_3)=v} T^v_{x_1x_2x_3}.$$

This union is not disjoint because each $\mathfrak{T}^v$ contains elements of H_3.

In [BDJN15] every such $\mathfrak{T}^v$ was a balanced directed graph obtained from the star with three leaves by subdividing each arc p times, for some p, the leaves being the elements of H_3. So, during the reduction from CSP on directed graphs to $\mathrm{CSP}^{\tilde{\tau}}$, it was clear which elements of the input directed graph are associated with the elements of the domain of the $\tilde{\tau}$-structure from which this directed graph is reduced. This constructive approach can be generalised by the following list of conditions applied to $\mathfrak{H}_2$. This means that any construction of $\mathfrak{H}_2$ that satisfies them, is called similar to the construction given in [BDJN15]. All the results in this section are related only to such similar constructions.

1. For each $\star$-graph $\mathfrak{T}^v_{x_1x_2x_3}$ that represents a tuple $\mathsf{R}^{\mathfrak{H}_3}(x_1, x_2, x_3) = v$, the problem $\mathrm{MP}_\star(\mathfrak{T}^v_{x_1x_2x_3})$ is solvable in P-time and $v_x = \mathsf{R}^{\mathfrak{H}_3}(x_1, x_2, x_3) \preceq_\star \mathsf{R}^{\mathfrak{H}_3}(y_1, y_2, y_3) = v_y$ if and only if $\mathfrak{T}^{v_x}_{x_1x_2x_3} \to \mathfrak{T}^{v_y}_{y_1y_2y_3}$.
2. Let $\mathfrak{H}_2, \mathfrak{H}'_2$ be two $\star$-graphs obtained from $(\star, \tilde{\tau})$-structures $\mathfrak{H}_3, \mathfrak{H}'_3$ by this approach. Then, for any homomorphism $\mathsf{h}\colon \mathfrak{H}_2 \to \mathfrak{H}'_2$, it is true that for all $x \in H_2$, $x \in H_3 \subseteq H_2 \Leftrightarrow \mathsf{h}(x) \in H'_3 \subseteq H'_2$.
3. For each $\star$-graph $\mathfrak{A}$ that is an input instance of $\mathrm{MP}_\star(\mathfrak{H}_2)$, one can decide in time polynomial in $|A|$ which elements of $\mathfrak{A}$ can only map to the elements of H_3. That is, we can decide, for every $x \in A$, if any $\mathsf{h}\colon \mathfrak{A} \to \mathfrak{H}_2$, implies that $\mathsf{h}(x) \in H_3 \subset H_2$. Also, for every $x \in A$, either any homomorphism from $\mathfrak{A} \to \mathfrak{H}_2$ maps x to H_3, or any homomorphism from $\mathfrak{A} \to \mathfrak{H}_2$ maps x to $H_2 \setminus H_3$.

4. For two elements $w, w' \in H_2$ such that $w, w' \notin H_3$ and w, w' do not belong to the same T^v_{xyz}, then $\mathsf{E}^{\mathfrak{H}_2}(w, w') = 0$.

5. Let $\mathfrak{A}$ be a $\star$-graph. Suppose that for some $v \neq v'$ there is $\mathsf{h}\colon \mathfrak{A} \to \mathfrak{T}^v_{xyz}$ and $\mathfrak{A} \not\to \mathfrak{T}^{v'}_{xyz}$. Suppose that, for every $a_0, a_n \in A$ such that $\mathsf{h}(a_0), \mathsf{h}(a_n) \in \{x, y, z\}$, there exist $a_1, \dots, a_{n-1} \in A$ such that:

   - for every $1 \leq i < n$, $\mathsf{h}(a_i) \notin \{x, y, z\}$,
   - for every $0 \leq i < n$, $\mathsf{E}^{\mathfrak{A}}(a_i, a_{i+1}) \neq 0$ or $\mathsf{E}^{\mathfrak{A}}(a_{i+1}, a_i) \neq 0$.

   Then, for any other $\mathsf{h}'\colon \mathfrak{A} \to \mathfrak{T}^v_{xyz}$ and for all $a \in A$ such that $\mathsf{h}(a) \in \{x, y, z\}$, we have that $\mathsf{h}(a) = \mathsf{h}'(a)$.

In particular, the reduction from CSP^τ to CSP on directed graphs in [BDJN15] satisfies the first four conditions. The fifth one cannot be applied to CSP because there are no three different types of $\mathfrak{T}^v$ in that case. Any polynomial time reduction satisfying these five conditions, cannot prove the P-equivalence with $\star$-graphs, unless $\mathrm{CSP} \equiv_p \mathrm{MP}$. We assume that $\mathrm{CSP} \equiv_p \mathrm{MP}_\varnothing$ by Section 6.2.

Proposition 6.3.11. *Let a $\star$-graph $\mathfrak{H}_2$ be constructed from some $(\star, \tilde{\tau})$-structure $\mathfrak{H}_3$ and satisfy all the five conditions above. Then $\mathrm{MP}^{\tilde{\tau}}_\star(\mathfrak{H}_3)$ reduces in P-time to $\mathrm{MP}_\star(\mathfrak{H}_2)$, and $\mathrm{MP}_\star(\mathfrak{H}_2)$ reduces in P-time to $\mathrm{MP}^{\tilde{\tau}}_\varnothing(\mathfrak{H}_3)$.*

Proof. Consider $(\star, \tilde{\tau})$-structures $\mathfrak{G}_3, \mathfrak{H}_3$ and the corresponding $\star$-graphs $\mathfrak{G}_2$ and $\mathfrak{H}_2$ that satisfy the conditions 1–5. If there is $\mathsf{h}\colon \mathfrak{G}_3 \to \mathfrak{H}_3$, then, by the conditions 1 and 2, there is $\mathsf{h}_2\colon \mathfrak{G}_2 \to \mathfrak{H}_2$. If there is $\mathsf{h}_2\colon \mathfrak{G}_2 \to \mathfrak{H}_2$, then, by the condition 2, one can consider the restriction h of this map on the set G_3, and the codomain of this map will be the set H_3. By the condition 1, h is a homomorphism between $\mathfrak{G}_3$ and $\mathfrak{H}_3$.

Now, consider any $\star$-graph $\mathfrak{A}$ from the input of $\mathrm{MP}_\star(\mathfrak{H}_2)$. By the condition 3, we can mark in P-time all the elements that can map only to the elements of H_3, denote the set containing them by A_3. Then on the set $A \setminus A_3$ we define the following equivalence relation $eq(\cdot, \cdot)$: for two elements $a_0, a_n \in A \setminus A_3$, we say that $eq(a_0, a_n)$ if there exists a sequence of elements $a_0, a_1, \dots, a_{n-1}, a_n \in A \setminus A_3$ such that for any $0 \leq i < n$, either $\mathsf{E}^{\mathfrak{A}}(a_i, a_{i+1}) \neq 0$ or $\mathsf{E}^{\mathfrak{A}}(a_{i+1}, a_i) \neq 0$. For every eq-equivalence class A_a (containing an element a), consider an induced $\star$-subgraph $\mathfrak{A}_a$ on the subset consisting of A_a itself together with those $b \in A_3$ such that there exists $c \in A_a$ such that either $\mathsf{E}^{\mathfrak{A}}(b, c) \neq 0$ or $\mathsf{E}^{\mathfrak{A}}(b, c) \neq 0$. Below we will show that the image of every $\mathfrak{A}_a$ can only be contained in some $\mathfrak{T}^v_{xyz}$.

Claim 6.3.12. *If there is $\mathsf{h}\colon \mathfrak{A}_a \to \mathfrak{H}_2$, then $\mathsf{h}(A_a) \subseteq T^v_{xyz}$ for some $\mathfrak{T}^v_{xyz}$.*

Proof of Claim 6.3.12. For any two elements a_0, a_n of the eq-equivalence class A_a, there exists a sequence $a_1, \dots, a_{n-1}$ of elements of A_a such that, for any $0 \leq i \leq n-1$, one of $\mathsf{E}^{\mathfrak{A}}(a_i, a_{i+1})$ and $\mathsf{E}^{\mathfrak{A}}(a_{i+1}, a_i)$ is not 0. As for all $a \in A \setminus A_3$ and for all $\mathsf{h}'\colon \mathfrak{A} \to \mathfrak{H}_2$, $\mathsf{h}'(a) \notin H_3$, $\mathsf{h}(a_0), \dots, \mathsf{h}(a_n) \in H_2 \setminus H_3$ (by condition (3)). Then, by the condition 4, that is in $\mathfrak{H}_2$ any two elements w and w' belonging to different $\mathfrak{T}^v, \mathfrak{T}^{v'}$, $\mathsf{E}^{\mathfrak{H}_2}(w, w') = 0$, we have that all $\mathsf{h}(a_0), \dots, \mathsf{h}(a_n)$ are in the same $\mathfrak{T}^v$. □

By the condition 1, we find in P-time for every $\mathfrak{A}_a$ the list of values $v \in \{0, 1, \star\}$ such that $\mathfrak{A}_a$ maps to $\mathfrak{T}^v_{xyz}$. If $\mathfrak{A}_a \not\to \mathfrak{T}^v_{xyz}$ for any v, then there is no way that $\mathfrak{A}$ can be mapped to $\mathfrak{H}_2$ and we reject this instance. Among all v such that $\mathfrak{A}_a \to \mathfrak{T}^v$, we label $\mathfrak{A}_a$ with the

smallest possible such v with respect to $\preceq_\star$. If $\mathfrak{A}_a$ maps to $\mathfrak{T}^v_{xyz}$ for any possible v, then we say that $\mathfrak{A}_a$ is $\varnothing$-labeled. Introduce a new equivalence relation $map(\cdot,\cdot)$ on the set A_3, we say that $map(a_1,a_2)$ if there exists $\mathfrak{A}_a \ni a_1,a_2$ and there is $\mathtt{h}\colon \mathfrak{A}_a \to \mathfrak{T}^v$ such that $\mathtt{h}(a_1)=\mathtt{h}(a_2)$. By the condition 5, for any $a_1,a_2\in A_3$ such that $map(a_1,a_2)$: there is $\mathtt{h}\colon \mathfrak{A}\to\mathfrak{H}_2 \Rightarrow \mathtt{h}(a_1)=\mathtt{h}(a_2)$. Let us construct a new $\star$-graph $\mathfrak{A}_2$ based on $\mathfrak{A}$. Take the domain $A_2=A_3/map$ and, for any $(a_1,a_2,a_3)\in(A_2)^3$, add a gadget $\mathfrak{T}^v_{a_1a_2a_3}$ following the rules below. Consider $\mathfrak{A}_a$ labeled with $v\neq\varnothing$ such that for any element $x\in H_3$ (or y or z) of $\mathfrak{T}^v_{xyz}$ there exists an A_3-element a_x of $\mathfrak{A}_a$ such that a_x is mapped to x. In this case we substitute $\mathfrak{A}_a$ by $\mathfrak{T}^v_{a_xa_ya_z}$ for $a_x,a_y,a_z\in A_2$. Consider those $\mathfrak{A}_a$ labeled with $v\neq\varnothing$ where there exists an element $x\in H_3$ (or y or z) of $\mathfrak{T}^v_{xyz}$ so that no element a_x of $\mathfrak{A}_a$ maps to x. For such a case we add to A_2 a new element $a_{\mathfrak{A}_a,x}$ and substitute $\mathfrak{A}_a$ by $\mathfrak{T}^v$ for the corresponding 3 elements of A_2. All the eq-equivalence classes A_a labeled with $\varnothing$ are not substituted with anything in $\mathfrak{A}_2$. The $\star$-graph $\mathfrak{A}_2$ is associated with a $(\varnothing,\tilde{\tau})$-structure $\mathfrak{A}_3$ as follows: each triple a_1,a_2,a_3 of A_3 is either contained in $\mathfrak{T}^v_{a_1a_2a_3}$ or not. If yes, then we set $\mathtt{R}^{\mathfrak{A}_3}(a_1,a_2,a_3)=v$, if not, then $\mathtt{R}^{\mathfrak{A}_3}(a_1,a_2,a_3)=\varnothing$. It is routine to check now that $\mathfrak{A}\to\mathfrak{H}_2$ if and only if $\mathfrak{A}_3\to\mathfrak{H}_3$. □

6.4 Minimal Obstructions

We prove that the inclusion-minimal obstructions considered in [FHX07] coincide with finitary duality in $\mathbf{Struct}_{01}$. We also show that being characterised by a finite set of inclusion-minimal obstructions in $\mathbf{Struct}_{01}$ is equivalent to be characterised by a finite set of inclusion-minimal obstructions in $\mathbf{Struct}_\star$. The main results of this section are summarised in the following.

Theorem 6.4.1. *Let $\mathfrak{H}$ be a $\star$-structure. Then, the following are equivalent.*

1. *MP($\mathfrak{H}$) has finitary duality.*
2. *$Obs^{\subset}_{01}(\mathfrak{H})$ is finite.*
3. *$MP_\star(\mathfrak{H})$ has finitary duality.*
4. *$Obs^{\subset}_{\star}(\mathfrak{H})$ is finite.*

Throughout this section, let $\tau=\{\mathtt{R}_1,\dots,\mathtt{R}_p\}$ be a fixed signature. We recall that, for $*$ in $\{01,\star,\varnothing\}$, $\mathbf{Struct}_*$ is the set of all $(*,\tau)$-structures.

Let us now recall the definitions of obstructions from [FHX07] and of finitary duality, that we extend to all structures.

Definition 24 ([FHX07])**.** Let $*$ be in $\{01,\star,\varnothing\}$ and let $\mathfrak{H}$ be a $\star$-structure. A $*$-structure $\mathfrak{G}$ is called an *inclusion-minimal obstruction* for $\mathrm{MP}_*(\mathfrak{H})$ if $\mathfrak{G}\not\to\mathfrak{H}$ and for all v in G, $\mathfrak{G}\setminus\{v\}\to\mathfrak{H}$. The set of all obstructions for the problem $\mathrm{MP}_*(\mathfrak{H})$ is denoted by $\mathrm{Obs}^{\subset}_*(\mathfrak{H})$.

Definition 25. Let $*$ be in $\{01,\star,\varnothing\}$ and let $\mathfrak{H}$ be a $\star$-structure. We say that a set $\mathcal{F}$ of $*$-structures is a *duality set* for the problem $\mathrm{MP}_*(\mathfrak{H})$ if

$$\mathfrak{G}\notin \mathrm{MP}_*(\mathfrak{H}) \iff \mathfrak{F}\to\mathfrak{G}, \qquad \text{for some } \mathfrak{F}\in\mathcal{F}.$$

If, moreover, the set $\mathcal{F}$ is finite, we say that $\mathrm{MP}_*(\mathfrak{H})$ has *finitary duality*.

We prove that the inclusion-minimal obstruction set is finite if and only if there is a finite duality set. We also show how their finiteness depends on the category from which the input structures are taken. We summarise on the diagram below all the dependencies that we have proved. Every arrow of this diagram means the implication.

$$\begin{array}{ccccc} |\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})| < \infty & \longleftrightarrow & |\mathrm{Obs}^{\subset}_{\star}(\mathfrak{H})| < \infty & \longleftarrow & |\mathrm{Obs}^{\subset}_{\varnothing}(\mathfrak{H})| < \infty \\ \updownarrow & & \updownarrow & & \updownarrow \\ \mathrm{MP}(\mathfrak{H}) \text{ has f.d.} & \longleftrightarrow & \mathrm{MP}_{\star}(\mathfrak{H}) \text{ has f.d.} & \longleftarrow & \mathrm{MP}_{\varnothing}(\mathfrak{H}) \text{ has f.d.} \end{array}$$

Let us first prove that on $\mathbf{Struct}_{01}$ the inclusion-minimal obstruction set is also a duality set.

Proposition 6.4.2. *$Obs^{\subset}_{01}(\mathfrak{H})$ is a duality set for MP($\mathfrak{H}$). Moreover, among all duality sets, $Obs^{\subset}_{01}(\mathfrak{H})$ is the minimal one by inclusion.*

Proof. Let $\mathfrak{A} \not\to \mathfrak{H}$. We start iteratively removing arbitrary elements from $\mathfrak{A}$ until the substructure induced on the elements of the rest is an inclusion-minimal obstruction, *i.e.*, if we remove any element, then the resulting structure will map to $\mathfrak{H}$. Such a substructure belongs to $\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})$ and maps to $\mathfrak{A}$, thus, $\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})$ is a duality set.

Let $\mathcal{F}$ be a duality set for $\mathrm{MP}(\mathfrak{H})$ such that $|\mathcal{F}| \leq |\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})|$. We can assume without loss of generality that $\mathcal{F} \subseteq \mathrm{Obs}^{\subset}_{01}(\mathfrak{H})$: any $\mathfrak{F}$ in $\mathcal{F}$ has an induced substructure that belongs to $\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})$, so we can substitute $\mathfrak{F}$ with this substructure.

Let $\mathfrak{G}$ be in $\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})$. $\mathfrak{G}$ is a core, otherwise it contains a proper induced substructure that does not map to $\mathfrak{H}$, a contradiction. Assume now that there exists $\mathfrak{G}_1$ in $\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})$ such that there is a homomorphism $\mathfrak{h}\colon \mathfrak{G}_1 \to \mathfrak{G}$, and that $\mathfrak{G} \not\to \mathfrak{G}_1$. Let $G' = \mathfrak{h}(G_1)$ and let $\mathfrak{G}'$ be the substructure of $\mathfrak{G}$ induced by G'. If $\mathfrak{G}'$ is a proper induced substructure of $\mathfrak{G}$, then by the assumption of inclusion-minimality, and by transitivity of homomorphism: $\mathfrak{G}_1 \to \mathfrak{H}$ – a contradiction. Thus, $\mathfrak{h}(G_1) = G$, but since $\mathfrak{h}$ is a full homomorphism, $\mathfrak{G}$ is either a proper induced substructure of $\mathfrak{G}_1$, which would contradict our assumption that $\mathfrak{G}_1$ is a core, or is isomorphic to $\mathfrak{G}_1$, which would contradict the assumption that $\mathfrak{G} \not\to \mathfrak{G}_1$. We can then conclude that $\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})$ contains no proper subsets that are duality sets for $\mathrm{MP}(\mathfrak{H})$, which means that this duality set is a subset of any other duality set. □

Corollary 6.4.3. *$|Obs^{\subset}_{01}(\mathfrak{H})| < \infty$ if and only if MP($\mathfrak{H}$) has finitary duality.*

The result of Proposition 6.4.2 does not hold on $\mathbf{Struct}_{\star}$.

Proposition 6.4.4. *For any 01-structure $\mathfrak{H}$, $Obs^{\subset}_{\star}(\mathfrak{H})$ is not a duality set of minimal size.*

Proof. Pick some vertex x in the domain H of $\mathfrak{H}$. Consider a $\star$-graph $\mathfrak{G} = (\{u, v\}, \mathsf{E}^{\mathfrak{G}})$ with $\mathsf{E}^{\mathfrak{G}}(u, u) = \mathsf{E}^{\mathfrak{G}}(v, v) = \mathsf{E}^{\mathfrak{H}}(x, x)$ and $\mathsf{E}^{\mathfrak{G}}(u, v) = \mathsf{E}^{\mathfrak{G}}(v, u) = \star$. Also consider a $\star$-graph $\mathfrak{G}'$ obtained from $\mathfrak{G}$ by setting $\mathsf{E}^{\mathfrak{G}'}(v, u) = 0$, and keeping the rest as in $\mathfrak{G}$, see Figure 6.4.

Both $\mathfrak{G}$ and $\mathfrak{G}'$ belong to $\mathrm{Obs}^{\subset}_{\star}(\mathfrak{H})$ as $\mathfrak{H}$ has an element x such that $\mathsf{E}^{\mathfrak{H}}(x, x) = \mathsf{E}^{\mathfrak{G}}(u, u) = \mathsf{E}^{\mathfrak{G}}(v, v)$ and as $\mathfrak{G} \not\to \mathfrak{H}$ and similarly $\mathfrak{G}' \not\to \mathfrak{H}$ because they both have a $\star$-arc and $\mathfrak{H}$ is a 01-graph. Also, $\mathfrak{G}' \to \mathfrak{G}$ and $\mathfrak{G} \not\to \mathfrak{G}'$, so $\mathfrak{G}$ can be removed from $\mathrm{Obs}^{\subset}_{\star}(\mathfrak{H})$ if it is a duality set. For and arbitrary signature τ the proof will be similar. □

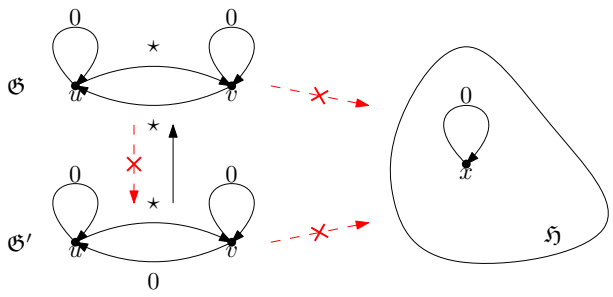


Figure 6.4: The $\star$-graphs $\mathfrak{G}$ and $\mathfrak{G}'$ from the input of $\mathrm{MP}_\star(\mathfrak{H})$. Without loss of generality, we assume that $\mathsf{E}^{\mathfrak{H}}(x,x)=0$, the case when it is equal to 1 is equivalent.

6.4.1 01- and $\star$-obstruction sets agree on being finite

We prove Theorem 6.4.1 in this section. From now on, we fix a $\star$-structure $\mathfrak{H}$. We have showed, see Corollary 6.4.3, that Theorem 6.4.1(1.) and Theorem 6.4.1(2.) are equivalent. We prove the other equivalences with the following propositions. Corollary 6.4.6 of the following Proposition 6.4.5 proves that Theorem 6.4.1(3.) is equivalent to Theorem 6.4.1 (1.).

Proposition 6.4.5. *A family of* 01*-structures* $\mathcal{F}$ *is a duality set for* $MP_\star(\mathfrak{H})$ *if and only if* $\mathcal{F}$ *is a duality set for* $MP(\mathfrak{H})$.

Proof. Let $\mathcal{F}$ be a duality set for $\mathrm{MP}_\star(\mathfrak{H})$. Any 01-structure $\mathfrak{G}$ is also a $\star$-structure. So, if $\mathfrak{G} \not\to \mathfrak{H}$, then $\mathfrak{F} \to \mathfrak{G}$ for some $\mathfrak{F}$ in $\mathcal{F}$. This means that $\mathcal{F}$ is a duality set for $\mathrm{MP}(\mathfrak{H})$.

Let $\mathcal{F}$ be a finite duality set for $\mathrm{MP}(\mathfrak{H})$. Let $\mathfrak{G}$ be a $\star$-structure that does not map to $\mathfrak{H}$. By Theorem 6.3.1, for any $\mathfrak{G}$ in $\mathbf{Struct}_\star \setminus \mathbf{Struct}_{01}$ there exists $\mathfrak{G}_{01}$ in $\mathbf{Struct}_{01}$ such that

- there is a surjective homomorphism $\pi_{\mathfrak{G}} \colon \mathfrak{G}_{01} \twoheadrightarrow \mathfrak{G}$;
- $\mathfrak{G} \not\to \mathfrak{G}_{01}$;
- $\mathfrak{G}$ is in $\mathrm{MP}_\star(\mathfrak{H})$ if and only if $\mathfrak{G}_{01}$ is in $\mathrm{MP}(\mathfrak{H})$.

As $\mathfrak{G}_{01}$ is in $\mathbf{Struct}_{01}$ and as it also does not map to $\mathfrak{H}$, there exists $\mathfrak{F}$ in $\mathcal{F}$ such that $\mathfrak{F} \to \mathfrak{G}_{01}$. By transitivity, $\mathfrak{F} \to \mathfrak{G}$, this means that $\mathcal{F}$ is also a duality set for $\mathrm{MP}_\star(\mathfrak{H})$. □

Corollary 6.4.6. $MP_\star(\mathfrak{H})$ *has finitary duality if and only if* $MP(\mathfrak{H})$ *has finitary duality.*

Proof. Let $\mathcal{F}$ be a finite duality set for $\mathrm{MP}_\star(\mathfrak{H})$. Every 01-structure $\mathfrak{G}$ is also a $\star$-structure. So, $\mathfrak{G} \not\to \mathfrak{H}$ implies $\mathfrak{F} \to \mathfrak{G}$ for some $\mathfrak{F}$ in $\mathcal{F}$. Such a structure $\mathfrak{F}$ must also be a 01-structure, thus, the subfamily $\mathcal{F}_{01} \subseteq \mathcal{F}$ consisting of 01-structures is a finite duality set for $\mathrm{MP}(\mathfrak{H})$. □

Remark. Proposition 6.4.2 provides that $\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})$ is the minimal by inclusion duality set for $\mathrm{MP}(\mathfrak{H})$. Proposition 6.4.5 and the proof of Corollary 6.4.6 allow us to conclude that $\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})$ is also the minimal by inclusion duality set for $\mathrm{MP}_\star(\mathfrak{H})$. So, without loss of generality, we can think of $\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})$ when we consider a duality set for $\mathrm{MP}_\star(\mathfrak{H})$.

The following statement proves the equivalence between Theorem 6.4.1(2.) and Theorem 6.4.1(4.).

Proposition 6.4.7. *$Obs^{\subset}_{\star}(\mathfrak{H})$ is finite if and only if $Obs^{\subset}_{01}(\mathfrak{H})$ is finite.*

Proof. Let us first prove the right implication. As any 01-structure is also a $\star$-structure, we can conclude that $\mathrm{Obs}^{\subset}_{01}(\mathfrak{H}) \subseteq \mathrm{Obs}^{\subset}_{\star}(\mathfrak{H})$.

Let us now turn our attention to the left implication. Let us consider the class $\overline{\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})}$, that is obtained from $\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})$ by taking all $\star$-structures $\mathfrak{A}$ such that there exists a surjective homomorphism from $\mathfrak{B}$ to $\mathfrak{A}$ for some $\mathfrak{B}$ in $\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})$. Observe that $|\overline{\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})}|$ is finite. We know by Theorem 6.3.1 that, for any $\mathfrak{G}$ in $\mathrm{Obs}^{\subset}_{\star}(\mathfrak{H})$ there exists a 01-structure $\mathfrak{G}_{01}$ such that:

- there is a surjective homomorphism $\pi_{\mathfrak{G}} \colon \mathfrak{G}_{01} \twoheadrightarrow \mathfrak{G}$;
- $\mathfrak{G} \not\to \mathfrak{G}_{01}$;
- $\mathfrak{G}$ is in $\mathrm{MP}_{\star}(\mathfrak{H})$ if and only if $\mathfrak{G}_{01}$ is in $\mathrm{MP}(\mathfrak{H})$.

As $\mathfrak{G} \notin \mathrm{MP}_{\star}(\mathfrak{H})$, we can conclude that $\mathfrak{G}_{01} \notin \mathrm{MP}(\mathfrak{H})$. And as $\mathfrak{G}_{01}$ is a 01-structure, there exists $\mathfrak{G}'_{01}$ in $\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})$ such that $\mathfrak{G}'_{01}$ is an induced substructure of $\mathfrak{G}_{01}$, and thus, by transitivity, $\mathfrak{G}'_{01} \to \mathfrak{G}$. By inclusion-minimality of $\mathfrak{G}$ (recall that $\mathfrak{G}$ is in $\mathrm{Obs}^{\subset}_{\star}(\mathfrak{H})$), this homomorphism is surjective, *i.e.*, $\mathfrak{G}$ belongs to $\overline{\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})}$. We have thus proved that $\mathrm{Obs}^{\subset}_{\star}(\mathfrak{H}) \subseteq \overline{\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})}$, *i.e.*, is finite. □

6.4.2 Looking at obstructions in $\mathrm{Struct}_{\varnothing}$

The goal now is to prove the remaining arrows on the diagram from page 147. As in the previous section, let $\mathfrak{H}$ be a fixed $\star$-structure.

Proposition 6.4.8. *If $MP_{\varnothing}(\mathfrak{H})$ has finitary duality, then $MP_{\star}(\mathfrak{H})$ has finitary duality.*

Proof. Let $\mathcal{F}_{\varnothing}$ be the finite duality set for $\mathrm{MP}_{\varnothing}(\mathfrak{H})$. Let $\mathcal{F}_{\star}$ be a duality set for $\mathrm{MP}_{\star}(\mathfrak{H})$. Recall that without loss of generality we can assume $\mathcal{F}_{\star} = \mathrm{Obs}^{\subset}_{01}(\mathfrak{H})$. For any $\mathfrak{G}$ in $\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})$, we have $\mathfrak{G} \not\to \mathfrak{H}$, so we have $\mathfrak{G}_{\varnothing} \to \mathfrak{G}$, for some $\mathfrak{G}_{\varnothing}$ in $\mathcal{F}_{\varnothing}$. This homomorphism must be surjective because any proper induced substructure of $\mathfrak{G}$ maps to $\mathfrak{H}$. As $\mathcal{F}_{\varnothing}$ is finite, there is some constant c such that, for any $\mathfrak{G}_{\varnothing}$ in $\mathcal{F}_{\varnothing}$, $|G_{\varnothing}| < c$. But then we also have $|G| < c$, for any $\mathfrak{G}$ in $\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})$. □

We now prove a similar statement for inclusion-wise minimal obstructions.

Proposition 6.4.9. *If $Obs^{\subset}_{\varnothing}(\mathfrak{H})$ is finite, then $Obs^{\subset}_{\star}(\mathfrak{H})$ is finite.*

Proof. Any $\star$-structure is also a $\varnothing$-structure. Then, $\mathrm{Obs}^{\subset}_{\star}(\mathfrak{H}) \subseteq \mathrm{Obs}^{\subset}_{\varnothing}(\mathfrak{H})$. □

We are now going to prove that $\mathrm{Obs}^{\subset}_{\varnothing}(\mathfrak{H})$ is finite if and only if $\mathrm{MP}_{\varnothing}(\mathfrak{H})$ has finitary duality.

Proposition 6.4.10. *If $MP_{\varnothing}(\mathfrak{H})$ has finitary duality, then $Obs^{\subset}_{\varnothing}(\mathfrak{H})$ is finite.*

Proof. Let $\mathcal{F}_{\varnothing}$ be a finite duality set for $\mathrm{MP}_{\varnothing}(\mathfrak{H})$. Consider any $\mathfrak{G}$ in $\mathrm{Obs}^{\subset}_{\varnothing}(\mathfrak{H})$. Suppose that it is not in $\mathcal{F}_{\varnothing}$; then there exists $\mathfrak{T}$ in $\mathcal{F}_{\varnothing}$ such that $\mathfrak{T} \to \mathfrak{G}$. Moreover, we know that $\mathfrak{T}$ always maps surjectively to $\mathfrak{G}$, because otherwise the substructure of $\mathfrak{G}$ induced by the image of $\mathfrak{T}$ would not map to $\mathfrak{H}$, contradicting that $\mathfrak{G}$ is an inclusion-minimal obstruction. The set of $\varnothing$-structures $\mathfrak{G}$ such that $\mathfrak{T}$ surjectively maps to $\mathfrak{G}$ is finite because $|G| \leq |T|$. Thus, for every $\varnothing$-structure in $\mathrm{Obs}^{\subset}_{\varnothing}(\mathfrak{H})$ there exists $\mathfrak{T}$ in $\mathrm{Obs}^{\rightarrow}_{\varnothing}(\mathfrak{H})$ such that $\mathfrak{T}$ surjectively maps to $\mathfrak{G}$, and we can then conclude that $\mathrm{Obs}^{\subset}_{\varnothing}(\mathfrak{H})$ is finite. □

We state the following which finishes the proof of the diagram from page 147.

Proposition 6.4.11. *If $Obs^{\subset}_{\varnothing}(\mathfrak{H})$ is finite, then $MP_{\varnothing}(\mathfrak{H})$ has finitary duality.*

Proof. It is sufficient to show that $\mathrm{Obs}^{\subset}_{\varnothing}(\mathfrak{H})$ is a duality set. Suppose that $\mathfrak{G} \not\to \mathfrak{H}$, for some $\mathfrak{G}$ in $\mathbf{Struct}_{\varnothing}$. Then we can remove elements from $\mathfrak{G}$ until it is no longer possible. The resulting structure belongs to $\mathrm{Obs}^{\subset}_{\varnothing}(\mathfrak{H})$ and is embedded into $\mathfrak{G}$ as it is an induced substructure. This means that $\mathrm{Obs}^{\subset}_{\varnothing}(\mathfrak{H})$ is a duality set. □

It is possible to show that there exists $\mathfrak{H}$ such that $\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})$ is finite and $\mathrm{Obs}^{\subset}_{\varnothing}(\mathfrak{H})$ is infinite. Hence, there are no arrows on the diagram from page 147 from the right column ($\mathbf{Struct}_{\varnothing}$) to the other two.

Proposition 6.4.12. *Let $\mathfrak{H} = \mathfrak{K}_2$ be a 01-graph, the clique on 2 vertices. Then $Obs^{\subset}_{01}(\mathfrak{H})$ is finite and $Obs^{\subset}_{\varnothing}(\mathfrak{H})$ is infinite.*

Proof. Feder and Hell proved in [FH08] that once $\mathfrak{H}$ is a 01-graph, the inclusion-minimal obstructions for $\mathrm{MP}(\mathfrak{H})$ have bounded size. Thus, $\mathrm{Obs}^{\subset}_{01}(\mathfrak{H})$ is finite.

Let us show that $\mathrm{Obs}^{\subset}_{\varnothing}(\mathfrak{H})$ is infinite. Consider a $\varnothing$-graph $\mathfrak{C}_n$ on the domain $v_1, \ldots, v_n$ with $\mathsf{E}^{\mathfrak{C}_n}(v_i, v_{i+1}) = 1$ for all i in $[n-1]$ and with $\mathsf{E}^{\mathfrak{C}_n}(v_n, v_1) = 1$, and with all other arcs equal to $\varnothing$. The problem $\mathfrak{C}_n \to \mathfrak{H}$ is equivalent to the 2-coloring of a directed cycle that is a directed graph, for which we know that odd cycles are all inclusion-minimal obstructions. Similarly, deleting any vertex from $\mathfrak{C}_n$ creates a $\varnothing$-graph that maps to $\mathfrak{H}$. Thus, the set $\mathcal{C} = \{\mathfrak{C}_n \mid n \text{ is odd}\}$ is an infinite set of inclusion-minimal obstructions for $\mathrm{MP}_{\varnothing}(\mathfrak{H})$. □

6.5 Matrix Partition and its relation with logic

Feder and Vardi show in [FV98] that MMSNP and CSP denote the same class of problems, up to P-time equivalence. This means that this class can be studied from both logic and homomorphism perspectives. We discuss logics that potentially could be P-time equivalent to the class MP.

Let τ be the input relational signature, and $\sigma = \{\mathsf{M}_1, \ldots, \mathsf{M}_s\}$ be the existential signature consisting of unary relation symbols. Recall that the logic MonadicSNP from Chapter 2 consists of the sentences of the following form:

$$\exists \mathsf{M}_1, \ldots, \mathsf{M}_s \, \forall \mathbf{x} \bigwedge_{i=1}^{m} \neg\big(\alpha_i \wedge \beta_i\big),$$

where, for any i in $[m]$, α_i is a conjunction of τ-atoms or negated τ-atoms, and β_i is a conjunction of σ-atoms or negated σ-atoms.

Clearly, MP is a subclass of MonadicSNP, that is, every MP problem can be defined by a MonadicSNP sentence.

Proposition 6.5.1. *For any $s \times s$ matrix M consisting of $0, 1, *$, there is a sentence Φ_M in MonadicSNP such that $SAT(\Phi_M)$ is P-time equivalent to $MP(M)$.*

Proof. Let $\tau = \{\mathsf{E}(\cdot,\cdot)\}$ be the input signature of Φ_{M} and $\sigma = \{\mathsf{M}_1, \dots, \mathsf{M}_s\}$ be the existential signature. Then, the sentence Φ_{M} is of the following form:

$$\exists \mathsf{M}_1, \dots, \mathsf{M}_s \; \forall x \Big(\neg\big(\neg\mathsf{M}_1(x) \wedge \dots \wedge \neg\mathsf{M}_s(x)\big) \wedge \bigwedge_{i,j\in[s], i\neq j} \neg\big(\mathsf{M}_i(x) \wedge \mathsf{M}_j(x)\big) \Big) \forall x, y \bigwedge_{i,j\in[s]} \neg\phi_{ij}.$$

Here, the first negated conjunct requires that any element must belong to at least one σ-relation, and the next s negated conjuncts restrict an element to be in two different σ-relations. Then, every ϕ_{ij} depends on the value of the corresponding matrix element m_{ij} of M:

- if $m_{ij} = 0$, then $\phi_{ij} := \big(\mathsf{M}_i(x) \wedge \mathsf{M}_j(y) \wedge \mathsf{E}(x,y)\big)$;
- if $m_{ij} = 1$, then $\phi_{ij} := \big(\mathsf{M}_i(x) \wedge \neg\mathsf{M}_j(y) \wedge \mathsf{E}(x,y)\big)$;
- if $m_{ij} = *$, then $\phi_{ij} := \bot$, and, consequently, $\neg\phi_{ij} = \top$. In this case, we just can omit ϕ_{ij}, as $*$ gives no restriction.

By definition of MP, a directed graph satisfies Φ_{M} if and only if it admits M-partition. □

By Theorem 2.0.1 in Chapter 2, MonadicSNP does not have a dichotomy. But this does not mean that MP does not have it, as we do not know if MonadicSNP can be embedded into MP. In fact, being written in the MonadicSNP form, MP problems define a small fragment of MonadicSNP, which can be described by the following property satisfied by Φ_{M}, for any M.

Observation 4. Any negated conjunct of Φ_{M} contains at most one τ-atom or negated τ-atom. Which implies that, within any negated conjunct, all its τ-atoms have the same polarity: either they are all non-negated or all negated.

It is easy to show that the set of MonadicSNP sentences satisfying the property from Observation 4 coincides with the class MP. For any such sentence Φ, we can modify its existential signature σ in order to make the σ-relations form a partition of the input structure elements. This can be done similarly to the normal form transformation in Section 4.2. Where we, firstly, ensure that, for any negated conjunct $\neg\phi$ of Φ and for any variable x in ϕ, and for any σ-relation M in σ, this negated conjunct $\neg\phi$ contains either the σ-atom $\mathsf{M}(x)$ or its negated version $\neg\mathsf{M}(x)$. And then we replace σ by a new signature $\mathbf{2}^\sigma$, where every $\mathbf{2}^\sigma$-relation is associated with a subset of σ. Finally, we demand that these new $2^{|\sigma|}$ relation symbols form a partition. The resulting sentence still satisfies the property. Now, it is straightforward to construct an equivalent MP problem.

So, it makes sense to consider a more expressive logic, where, within any negated conjunct, its τ-atoms have the same polarity.

Definition 26. Let τ be the input signature and let $\sigma = \{\mathsf{M}_1, \dots, \mathsf{M}_s\}$ be the existential signature consisting of unary relation symbols. We say that a τ-sentence Φ belongs to the class MMSNP_* if it can be written in the following form:

$$\exists \mathsf{M}_1, \dots, \mathsf{M}_s \; \forall \mathbf{x} \bigwedge_{i=1}^{m} \neg\big(\alpha_i \wedge \beta_i\big),$$

where, for any i in $[m]$, either α_i is a conjunction only of non-negated τ-atoms or it is a conjunction only of negated τ-atoms; and β_i is a conjunction of σ-atoms or negated σ-atoms.

Clearly, MMSNP_* contains MP. Regarding this class, we can formulate the following problem.

Problem 2. Is it true that for any sentence Φ in MMSNP_* there is a problem in MP which is P-time equivalent to SAT(Φ)?

The logic MMSNP_* is strictly more expressive than MMSNP.

Proposition 6.5.2. *There is a sentence Φ_* in MMSNP$_*$ such that, for any Φ in MMSNP, Φ_* is not logically equivalent to Φ.*

Proof. Let Φ_* be a sentence on directed graphs that accepts only complete graphs with loops: $\forall x, y\ \mathsf{E}(x, y)$. Theorem 3 in [FV03] states that the class of structures satisfying a MonadicSNP sentence Ψ is closed under inverse homomorphisms if and only if Ψ is logically equivalent to a MMSNP sentence. As SAT(Φ_*) is not closed under inverse homomorphisms, it cannot be expressed by an MMSNP sentence. □

6.6 Matrix Partition and polymorphisms

Let $\mathfrak{A}$ be a $(\star, \tau)$-structure with the domain A. A mapping $\mathsf{f}\colon \underbrace{A \times \cdots \times A}_{n} \to A$ is called a *polymorphism* if, for all R in τ of arity k, for any n tuples $\mathbf{t}_1, \dots, \mathbf{t}_n$ in A^k, we have

$$\bigwedge_{i=1}^{n} \mathsf{R}^{\mathfrak{A}}(\mathbf{t}_i) \preceq_{\varnothing} \mathsf{R}^{\mathfrak{A}}\big(\mathsf{f}(\mathbf{t}_1, \dots, \mathbf{t}_n)\big),$$

where $\bigwedge$ denotes the greatest lower bound of the elements and $\mathsf{f}(\mathbf{t}_1, \dots, \mathbf{t}_n)$ denotes $\big(\mathsf{f}(t_1^1, \dots, t_n^1), \dots, \mathsf{f}(t_1^k, \dots, t_n^k)\big)$. The set of all polymorphisms of $\mathfrak{A}$ is denoted by Pol($\mathfrak{A}$).

A polymorphism $\mathsf{s}\colon A^4 \to A$ is called *Siggers* if it satisfies the following identity:

$$\forall x, y, z \in A\ \mathsf{s}(x, y, z, x) = \mathsf{s}(y, x, y, z). \tag{6.9}$$

The following is proved in [Zhu20] for Constraint Satisfaction Problems (CSP) on finite relational structures.

Theorem 6.6.1. *Let $\mathfrak{A}$ be a finite relational structure over some finite signature. Then either*

- *there is a Siggers polymorphism $\mathsf{s}\colon \mathfrak{A}^4 \to \mathfrak{A}$ and CSP($\mathfrak{A}$) is in P, or*
- *there is no Siggers polymorphism and CSP($\mathfrak{A}$) is NP-complete.*

We can reduce any Matrix Partition problem in P-time to some CSP such that the Matrix Partition problem is equivalent to this CSP being restricted on complete input instances. In particular, for any $\star$-graph $\mathfrak{H}$, we can construct a relational structure $\mathfrak{H}_{\mathrm{CSP}}$ over a signature $\{\mathsf{E}_0(\cdot,\cdot), \mathsf{E}_1(\cdot,\cdot)\}$. Here, by *complete input* we mean those relational structures $\mathfrak{A}$ such that for all x, y in A, (x, y) belongs exactly to one of $\mathsf{E}_0^{\mathfrak{A}}$ and $\mathsf{E}_1^{\mathfrak{A}}$.

Our goal is to show that Matrix Partition problems cannot be characterised by the conditions from Theorem 6.6.1. To do that, we provide a $\star$-graph $\mathfrak{P}$ such that MP($\mathfrak{P}$) is P-time solvable but the corresponding structure $\mathfrak{P}_{\mathrm{CSP}}$ does not have a Siggers polymorphism and thus CSP($\mathfrak{P}_{\mathrm{CSP}}$) is NP-complete.

We provide a $\star$-graph such that the corresponding Matrix Partition problem is solvable in P-time and that it has no Siggers polymorphism. Consider a $\star$-graph $\mathfrak{P}$ as on Figure 6.5. We have chosen it to be of the form of a directed path because it is well-known that list matrix partition problems are P-time solvable on them (we prove it formally, though). So the objective is to prove that such structure has no nontrivial polymorphisms. By *trivial* we mean a projection.

Figure 6.5: A $\star$-path $\mathfrak{P}$ of length 4.

Theorem 6.6.2. *$\mathfrak{P}$ does not have Siggers polymorphism and MP($\mathfrak{P}$) is P-time solvable.*

We prove Theorem 6.6.2 at the end of this section after proving all the lemmas.

Lemma 6.6.3 (Theorem 2 in [FHSS11])**.** *Let $\mathfrak{H}$ be a $\star$-graph. Let $\mathfrak{H}^-$ be a directed graph obtained from $\mathfrak{H}$ by adding an arc for every pair (x, y) in H^2 such that $\mathsf{E}^{\mathfrak{H}}(x, y) \in \{1, \star\}$. Then, if $\mathfrak{H}^-$ is an oriented tree, then the list matrix partition problem of $\mathfrak{H}$ is P-time equivalent to the list homomorphism problem of $\mathfrak{H}^-$.*

Lemma 6.6.4. *If $\mathsf{s}\colon A^4 \to P$ is a polymorphism, then it is* conservative*. That is, for any x_1, x_2, x_3, x_4 in P, we have $\mathsf{s}(x_1, x_2, x_3, x_4) \in \{x_1, x_2, x_3, x_4\}$.*

Proof. As (x, y) is the only pair such that $\mathsf{E}^{\mathfrak{P}}(x, y) = \star$, we must have $\mathsf{s}(x, x, x, x) = x$ and $\mathsf{s}(y, y, y, y) = y$. As x is the only element of P that is adjacent to w by a 1-arc, we must have $\mathsf{s}(w, w, w, w) = w$. As y is the only element of P that is adjacent to z by a 1-arc, we must have $\mathsf{s}(z, z, z, z) = z$.

Suppose that $\mathsf{s}(w, w, x, x) \in \{y, z\}$. If it equals y, then $\mathsf{s}(x, x, y, y) = z$, but then, for any value of $\mathsf{s}(y, y, z, z)$, s violates the polymorphism property. If $\mathsf{s}(w, w, x, x) = z$, then, for any value of $\mathsf{s}(x, x, y, y)$, s violates the polymorphism property. The proof, for other tuples from $\{w, x\}^4 \cup \{x, y\}^4 \cup \{y, z\}^4$, is similar.

Suppose that $\mathsf{s}(w, w, y, y) \in \{x, z\}$. If it equals x, then it contradicts to $\mathsf{s}(w, w, w, w) = w$. If it equals z, then, for any value of $\mathsf{s}(x, x, z, z)$, s violates the polymorphism property. The proof, for other tuples from $\{w, y\}^4 \cup \{x, z\}^4$ is similar.

Suppose that $\mathsf{s}(w, w, z, z) \in \{x, y\}$. If it equals x, then it contradicts to $\mathsf{s}(w, w, w, w) = w$. If it equals y, then it contradicts to $\mathsf{s}(z, z, z, z) = z$. The proof for other tuples from $\{w, z\}^4$ is similar.

Suppose that $\mathsf{s}(w, w, x, y) = z$; then, for any value of $\mathsf{s}(x, x, y, z)$, s violates the polymorphism property. The proof for other tuples from $\{w, x, y\}^4 \cup \{x, y, z\}^4$ is similar.

Suppose that $\mathsf{s}(w, w, x, z) = y$; then it contradicts $\mathsf{s}(z, z, z, z) = z$. The proof for other tuples from $\{w, x, z\}^4 \cup \{w, y, z\}^4$ is similar. □

Order w, x, y, z as $w \prec x \prec y \prec z$. Take two tuples $\mathbf{t}, \mathbf{t}'$ in $\{w, x\}^4 \cup \{x, y\}^4 \cup \{y, z\}^4$. We say that $\mathbf{t} \preceq \mathbf{t}'$ if, for any i in $[4]$, $t_i \preceq t'_i$, where $\preceq$ means $\prec$ or $=$. This ordering is displayed on Figure 6.6 for tuples from $\{x, y\}$.

Lemma 6.6.5. *Let s be a polymorphism of $\mathfrak{P}$. Then, for any $\mathbf{t}, \mathbf{t}'$ in $\{w, x\}^4 \cup \{x, y\}^4 \cup \{y, z\}^4$, if $\mathbf{t} \preceq \mathbf{t}'$, then $\mathsf{s}(\mathbf{t}) \preceq \mathsf{s}(\mathbf{t}')$.*

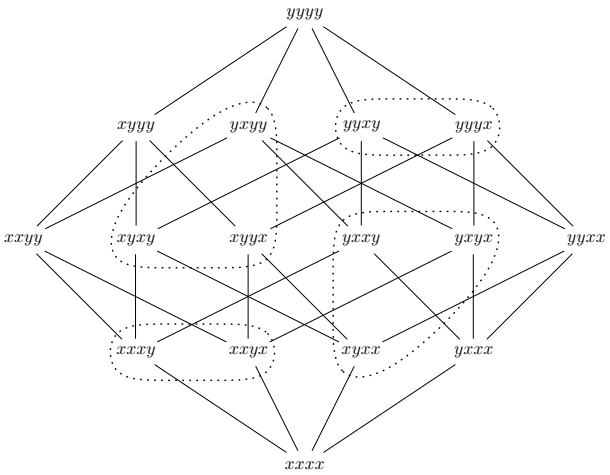


Figure 6.6: The Hasse diagram of the $\preceq$-ordering of $\{x, y\}^4$. Dotted closed curves highlight those tuples that are mapped to the same element by a Siggers polymorphism.

Proof. If $\mathbf{t}$ and $\mathbf{t}'$ consist of elements of different two-element sets, then, by Lemma 6.6.4, we are done.

Suppose that $\mathbf{t}, \mathbf{t}'$ belong to $\{w, x\}^4$. By Lemma 6.6.4, $\mathsf{s}(\mathbf{t}), \mathsf{s}(\mathbf{t}')$ belong to $\{w, x\}$. Suppose that $\mathbf{t} \preceq \mathbf{t}'$ and $\mathsf{s}(\mathbf{t}) \not\preceq \mathsf{s}(\mathbf{t}')$, that is, $\mathsf{s}(\mathbf{t}) = x$ and $\mathsf{s}(\mathbf{t}') = w$. On one hand, $\mathsf{E}^{\mathfrak{P}}(t'_i, t_i) = 0$, for any i in [4]. But $\mathsf{E}^{\mathfrak{P}}\big(\mathsf{s}(\mathbf{t}', \mathbf{t})\big) = \mathsf{E}^{\mathfrak{P}}(w, x) = 1$ which contradicts s being a polymorphism. The proof for other tuples from $\{w, x\}^4 \cup \{y, z\}^4$ is similar.

Suppose that, for $\mathbf{t}, \mathbf{t}'$ in $\{x, y\}^4$, we have $\mathbf{t} \preceq \mathbf{t}'$. Let $\mathbf{u}$ and $\mathbf{u}'$ be obtained from $\mathbf{t}$ and $\mathbf{t}'$ by replacing each occurence of x with w and each occurence of y with x. Then we have $\mathbf{u} \preceq \mathbf{u}'$, and, as they both belong to $\{w, x\}^4$, we have $\mathsf{s}(\mathbf{u}) \preceq \mathsf{s}(\mathbf{u}')$. Suppose that $\mathsf{s}(\mathbf{t}) \not\preceq \mathsf{s}(\mathbf{t}')$; then, by Lemma 6.6.4, $\mathsf{s}(\mathbf{t}) = y$ and $\mathsf{s}(\mathbf{t}') = x$. Then, $\mathsf{s}(\mathbf{u}') = w$ and, consequently, $\mathsf{s}(\mathbf{u}) = w$. As $\bigwedge_{i=1}^{4} \mathsf{E}^{\mathfrak{P}}(u_i, t_i) \not\preceq_{\varnothing} \mathsf{E}^{\mathfrak{P}}\big(\mathsf{s}(\mathbf{u}), \mathsf{s}(\mathbf{t})\big)$, s is not a polymorphism. □

Lemma 6.6.6. *Let s be a Siggers polymorphism of $\mathfrak{P}$. Let $\mathbf{t}$ be a 4-tuple from $\{w, x\}^4 \cup \{x, y\}^4 \cup \{y, z\}^4$, where 3 out of 4 elements are equal to a, for some a in P. Then, $\mathsf{s}(\mathbf{t}) = a$.*

Proof. Suppose that $\mathbf{t}$ is in $\{x, y\}^4$. We are going to show that, if 3 out of 4 elements of $\mathbf{t}$ are equal to x, then $\mathsf{s}(\mathbf{t}) \neq y$, so it can only be mapped to x. Observe that it is sufficient to show that (x, y, x, x) does not map to y. Because, by Siggers property, $\mathsf{s}(x, y, x, x) = \mathsf{s}(y, x, y, x)$ and, by Lemma 6.6.5, $\mathsf{s}(x, x, y, x) \preceq \mathsf{s}(y, x, y, x)$ and $\mathsf{s}(y, x, x, x) \preceq \mathsf{s}(y, x, y, x)$, and, by Siggers property, $\mathsf{s}(x, x, y, x) = \mathsf{s}(x, x, x, y)$. See Figure 6.6 on page 154.

Suppose that $\mathsf{s}(x, y, x, x) = y$. Then, $\mathsf{s}(y, z, y, y) = z$, as $\mathsf{E}^{\mathfrak{P}}(x, y) = \mathsf{E}^{\mathfrak{P}}(y, z) = 1$ and as (y, z) is the only 1-arc that goes from y. By Siggers property, $\mathsf{s}(z, y, y, z) = \mathsf{s}(y, z, y, y) = z$. Therefore, by Lemma 6.6.5, $\mathsf{s}(z, y, y, z) \preceq \mathsf{s}(z, y, z, z)$, so $\mathsf{s}(z, y, z, z) = z$. But, $\mathsf{E}^{\mathfrak{P}}(y, z) = 1$ and $\mathsf{E}^{\mathfrak{P}}(x, z) = \mathsf{E}^{\mathfrak{P}}(y, y) = 0$, that contradicts the assumption that s is a polymorphism. We conclude that $\mathsf{s}(x, y, x, x) = x$, and all other tuples from $\{x, y\}^4$, where 3 out of 4 elements are equal to x, must be mapped to x. Similarly, all tuples of $\{x, y\}^4$, where 3 out of 4 elements equal y, must be mapped to y. The proof for tuples from $\{w, x\}^4 \cup \{y, z\}^4$ is similar. □

For a set A, a ternary function $\mathsf{f} \colon A^3 \to A$ is called *majority* if, for any x, y in A,

$$\mathsf{f}(x, x, y) = \mathsf{f}(x, y, x) = \mathsf{f}(y, x, x) = x.$$

As an example, let A be equal to $[n]$, and consider the ternary *median* function $\mathfrak{m}\colon A^3 \to A$ that is defined as follows. Let x, y, z be three elements from A. Suppose that $\{x, y, z\} = \{a, b, c\}$, where $a \leq b \leq c$. Then $\mathfrak{m}(x, y, z)$ is defined to be b.

Proof of Theorem 6.6.2. Consider two tuples $\mathbf{t} = (w, w, x, y)$ and $\mathbf{t}' = (x, y, w, w)$. Suppose that $\mathbf{s}$ is a Siggers polymorphism of $\mathfrak{P}$. By Lemma 6.6.4, $\mathbf{s}(\mathbf{t}), \mathbf{s}(\mathbf{t}')$ belongs to $\{w, x, y\}$. Suppose that $\mathbf{s}(\mathbf{t}) = x$; then, by Lemma 6.6.6, $\mathbf{s}(w, w, x, w) = w$, this is a contradiction to $\mathbf{s}$ being a polymorphism. Suppose that $\mathbf{s}(\mathbf{t}) = y$; then, by Lemma 6.6.6, $\mathbf{s}(z, z, z, y) = z$, this is a contradiction to $\mathbf{s}$ being a polymorphism. Similarly, having $\mathbf{s}(x, w, w, w) = w$ and $\mathbf{s}(z, y, z, z)$ prevents $\mathbf{s}(\mathbf{t}')$ from belonging to $\{x, y\}$. Thus, both these tuples must be mapped to w. If $\mathbf{s}(w, w, x, y) = w$, then $\mathbf{s}(y, y, x, x) = y$ and, consequently, $\mathbf{s}(x, x, w, w) = x$. On the other hand, we have $\mathbf{s}(x, y, w, w) = w$, this contradicts to $\mathbf{s}$ being a polymorphism. So, any map $\mathbf{s}$ with Siggers property cannot be a polymorphism of $\mathfrak{P}$.

For any $\star$-graph $\mathfrak{H}$, the MP problem $\mathrm{MP}(\mathfrak{H})$ can be reduced to the list matrix partition problem associated with $\mathfrak{H}$. For an input directed graph $\mathfrak{G}$ of $\mathrm{MP}(\mathfrak{H})$, do the following. Denote by $\mathfrak{G}'$ the graph that is obtained from $\mathfrak{G}$ by removing all the loops. For x in G, if $\mathsf{E}^{\mathfrak{G}}(x, x) = 0$, then the list L_x consists of all h in H such that $\mathsf{E}^{\mathfrak{H}}(h, h) = 0$; if $\mathsf{E}^{\mathfrak{G}}(x, x) = 1$, then L_x consists of all h in H such that $\mathsf{E}^{\mathfrak{H}}(h, h) = 1$. Then, we reduce $\mathfrak{G}$ to $\mathfrak{G}'$ with the family of lists $\{L_x \mid x \in G\}$.

Let $\mathfrak{P}^-$ denote the directed path of length 4. $\mathfrak{P}^-$ has the median polymorphism $\mathfrak{m}$ with respect to the ordering $w < x < y < z$ on the set $P = \{w, x, y, z\}$. By Bulatov's characterisation for list-homomorphism problems from [Bul03], having a majority polymorphism implies that the corresponding list-homomorphism problem is P-time solvable.

By Lemma 6.6.3, the list-matrix partition problem associated with $\mathfrak{P}$ can also be solved in P-time, as the list-homomorphism problem is P-time solvable for $\mathfrak{P}^-$. As $\mathrm{MP}(\mathfrak{P})$ can be reduced to a P-time solvable problem, it is also P-time solvable. □

Chapter 7

Conclusion

In this part we briefly summarize the main results of this thesis and discuss possible perspectives for further work. We explain our ideas quite informally.

Although the main objective of this thesis is to study the dichotomy question for extensions of MMSNP and CSP, we manage to solve it only for MMSNP with guarded inequalities, which happens to be P-time equivalent to MMSNP, and, thus, has a dichotomy. For each of the other classes that we have considered, this question still remains unanswered.

Regarding MMSNP_2, we believe that one should study algebraic properties of ω-categorical CSP templates that are associated with MMSNP_2 sentences or, more broadly, with regular infinite MMSNP sentences. Maybe, there is a chance to apply the same methods as for MMSNP, see [BMM18]. Apart from that, we also see some potential in constructing a right expander structure that can finish the reduction. It requires deep understanding of the construction properties that we want our expander to have. Results that we obtain in this thesis about infinite MMSNP and expander structures for MMSNP_2, could be useful for proving a dichotomy for this class.

Speaking of Matrix Partition problems (MP), we do not think that it is possible to provide a PvsNP-complete characterization that uses polymorphisms, as in CSP. Because we provide a $\star$-graph with a tractable MP problem that has no Siggers polymorphism. This means that it has only trivial polymorphisms. The class MP is contained in MMSNP_* that is a fragment of MonadicSNP. A dichotomy for MMSNP_* would imply a dichotomy for MP as well. We do not think that the complexity of MP problems can be characterized by having some forbidden induced substructures, similarly as it is done for undirected graphs in [HN90]. Because MP extends CSP on digraphs, and CSP on digraphs is not characterized in this manner.

It makes sense to pose questions about "grey zone" classes that are similar to those that have already been studied on CSP. We have considered one such question on both MMSNP_2 and MP: is it possible to simplify the signature so that we can always assume only problems on directed graphs? In both cases, we only manage to show that we can consider, without loss of generality, signatures with a unique relation symbol. Though, it seems much harder to reduce the arity of this relation symbol. It is hard to be done for MP because this class does not have a notion for "nothing", which is similar to the "non-edge" notion in CSP. However, for MMSNP_2 it seems to be more doable as it is not clear what the obstacle is. It remains just to understand well the properties that we need from the directed graphs. These properties must be, in some sense, dual to the "balanced" property that is used for CSP. Because both MMSNP and MMSNP_2 classes

are formulated dually with respect to CSP.

The question of containment of two MMSNP problems is also well-studied, see [Mad10]. It happens that $\mathrm{SAT}(\Phi_1)$ is contained in $\mathrm{SAT}(\Phi_2)$ if and only if we can replace existential relations σ_1 of Φ_1 with existential relations σ_2 of Φ_2 such that for any negated conjunct of Φ_2 there exists a more restrictive negated conjunct of the recoloured Φ_1. The map from σ_1 to σ_2 is called a *recolouring*, it is used in the proof of the dichotomy theorem for MMSNP in [BMM18]. For MMSNP_2 logic, it is easy to show that containment implies the existence of a recolouring. However, the other directions seems to be more complicated, the main obstacle is the existence of duplicated tuples.

For a complexity class, it is important to understand its relation with the first-order logic FO. This is done by characterizing the problems that are FO-definable. For example, see [Ats08], for a finite relational structure $\mathfrak{A}$, $\mathrm{CSP}(\mathfrak{A})$ is FO-definable if and only if $\mathrm{CSP}(\mathfrak{A})$ has a *finitary duality*, *i.e.*, there is a finite family $\mathcal{D}$ of structures such that, for any input structure $\mathfrak{B}$, we have $\mathfrak{B} \to \mathfrak{A}$ if and only if $\forall \mathfrak{D} \in \mathcal{D}\ \mathfrak{D} \not\to \mathfrak{B}$. This question can be asked for MP problems. This is the reason why we studied minimal obstructions for these problems. This question is unlikely to be solved for MP by a similar approach as it is done for CSP, because MP does not have a notion for "nothing". That is, by the same reason as it is hard to reduce the arity.

Appendix A

Maximum cut on interval graphs of interval count two is NP-complete

By Ladner's theorem, P $\neq$ NP implies the existence of NP-intermediate problems. However, we do not know about any naturally formulated problem that happens to be NP-intermediate. There are many NP problems that are not known to be NP-complete or P-time solvable, *e.g.* Graph Isomorphism and Integer Factoring. Some problems that are NP-complete for general case become P-time solvable if the input is restricted. And, for some cases of the input restriction, the questions about the problem complexity are still open. In this appendix, we solve one of such questions. Although it is not related to Feder and Vardi's logic MMSNP, this result concerns the P-NP-complete dichotomy question. This is why we decide to add it to the thesis.

A.1 Introduction

For a graph $G = \big(V(G), E(G)\big)$, a *cut* is a partition of $V(G)$ into two disjoint subsets. Any cut determines a *cut set* which is the set of all edges that have one endpoint in one part and the other endpoint in the other part. The *size* of a cut is the cardinality of its cut set. The *maximum cut problem*, denoted by MaxCut, asks for a cut of maximum size in G.

MaxCut is a fundamental and well-known NP-complete problem [GJ90]. The weighted version of the problem is one of Karp's original 21 NP-complete problems [Kar72]. MaxCut remains NP-hard even for cubic graphs [BK99], split graphs [BJ94], co-bipartite graphs [BJ94], unit disk graphs [DK07], total graphs [Gur99], and interval graphs [ABMR21]. On the positive side, polynomial time algorithms are known for planar graphs [Had75], line graphs [Gur99], graphs not contractible to K_5 [Bar83] and graphs with bounded treewidth [BJ94].

There are two papers that mainly motivate our research. First, a recent proof of NP-completeness of MaxCut on interval graphs provided by Adhikary et al. in [ABMR21]. Second, a more recent result of de Figueiredo et al. in [dFdMdSOS21], where they extend the result of the first paper by proving that MaxCut is NP-complete on graphs of interval count four. Using the technique of the above work, de Figueiredo et al. prove the NP-completeness of MaxCut on permutation graphs as well, which too was open for a long time [dFdMdSOS22]. The bounding of the number of interval lengths brings us closer to the final goal: to characterize MaxCut for unit interval graphs as they are exactly

interval graphs of interval count one. There were attempts to provide a polynomial-time algorithm for unit interval graphs [BKN99, BES17], but they both were later shown to be incorrect [BdFG+04, KMN20]. In this paper we extend the result of the second paper by proving the following theorem which brings us as close as possible to MaxCut on unit interval graphs.

Theorem A.1.1. MaxCut *on interval graphs of interval count two is NP-complete.*

A.2 Preliminaries

For a, b in $\mathbb{R}$, an *interval* between a and b is the set of all $x \in \mathbb{R}$ such that $a \leq x \leq b$. A family $\mathcal{M}$ of intervals $\{I_1, \ldots, I_n\}$ is called an *interval model* of a graph G if one can order the vertices of G: $V(G) = \{g_1, \ldots, g_n\}$ such that $\forall i, j \in [n], g_i g_j \in E(G)$ iff $I_i \cap I_j \neq \varnothing$. Here and elsewhere, we use the notation $[n]$ for the set $\{1, 2, \ldots, n\}$, where $n \in \mathbb{N}$. A graph G is called an *interval* graph if there exists a family of intervals $\mathcal{M}_G$ that is an interval model of G. A graph G is said to have *interval count* c if there exist a set $C \subset \mathbb{R}$ of size c and an interval model $\mathcal{M}$ of G, where the length of any interval of $\mathcal{M}$ belongs to C. We usually think of a MaxCut partition as of a coloring of intervals into two colors: R for red and B for blue. So, when we say that some interval is colored in some color, we mean that it belongs to the corresponding partition class. If an interval X has a color $c \in \{R, B\}$, then we frequently denote it by $\text{Color}(X) = c$. If $\mathcal{B}$ is a family of intervals that all have a color $c \in \{R, B\}$, then we also can write $\text{Color}(\mathcal{B}) = c$.

A.3 Background

We first start with the reduction of Adhikary et al. in [ABMR21]. They reduced MaxCut on cubic graphs to MaxCut on interval graphs. In their paper, each vertex and edge of the original cubic graph was represented by a set of intervals, called *vertex* and *edge gadgets* respectively. The interval model consisted of first all the vertex gadgets, and then all the edge gadgets arranged from left to right. If an edge was incident to a vertex, then the corresponding vertex and edge gadgets were "linked" by a pair of very long intervals whose left ends intersected the corresponding vertex gadgets and right ends intersected the edge gadget. They were called *link intervals*. The number of intervals in any gadget was much greater than the total number of link intervals in the graph. It was shown that, in a MaxCut partition of this interval graph, each vertex gadget or edge gadget could have only two possible partitions. For a vertex gadget, these two partitions were made to correspond to its membership in one of the partitioning sets for MaxCut of the cubic graph. If two adjacent vertices of the cubic graph belonged to different sets, then the corresponding edge gadget would make more cut edges with link intervals than if these vertex gadgets were in the same set. Thus, a maximum sized cut of the cubic graph always corresponded to a maximum sized cut of the constructed interval graph and vice versa, proving the latter to be NP-complete.

In the above reduction, intervals of two different lengths were used to construct the gadgets. However, the length of each link interval depended on the relative positions of its vertex and edge gadgets. So the total number of different lengths of link intervals was linearly dependent on the size of the cubic graph. So, this interval graph seemed to be far

away from unit interval graphs, for which the problem was still open. De Figueiredo et al. in [dFdMdSOS21] made a very important advancement in this regard. They showed that MaxCut was NP-complete even when the total number of lengths used for the intervals was only 4, i.e. when the interval count of the graph was 4. In their paper, an extra gadget, resembling the vertex and edge gadgets, was used as a "joining gadget" between link intervals. Instead of having a link interval running through the entire length between its corresponding vertex and edge gadgets, they used a chain of link intervals joined to each other with the use of joining gadgets. A link chain is a sequence of link intervals with

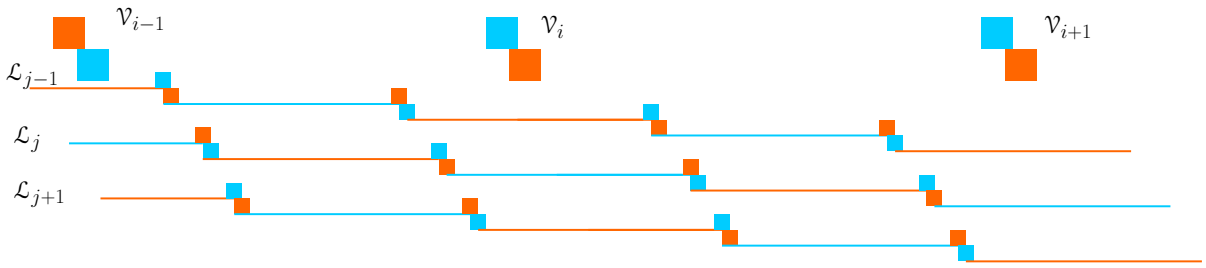


Figure A.1: Link chains between consecutive vertex gadgets.

a join gadget between every two consecutive link intervals. Such a join gadget partially intersects only particular link intervals. The link intervals of all other link chains that intersect the join gadget, intersect it fully. The join gadgets ensure that in a maximum cut partition, the link intervals of every link chain are colored with red and blue alternately. Thus, link intervals of arbitrary lengths in the original reduction can be replaced by link intervals of a single size. However, one problem remained. Between consecutive vertex or edge gadgets, the link chains had their join gadgets in a fixed order. For example, consider vertex gadgets $\mathcal{V}_{i-1}$, $\mathcal{V}_i$ and $\mathcal{V}_{i+1}$, and link chains $\mathcal{L}_{j-1}$, $\mathcal{L}_j$ and $\mathcal{L}_{j+1}$ (Fig. A.1. If in the gap between $\mathcal{V}_{i-1}$ and $\mathcal{V}_i$, the join gadget of $\mathcal{L}_{i-1}$ occurred first, followed by a join gadget each of $\mathcal{L}_j$ and $\mathcal{L}_{j+1}$ respectively, then they would occur in the same order in the gap between $\mathcal{V}_i$ and $\mathcal{V}_{i+1}$. This would pose a problem if the structure of the graph required $\mathcal{L}_{j-1}$ and $\mathcal{L}_{j+1}$ to have a partial intersection with the same edge gadget. In such a case, a second type of link interval with a different length would be used to end link chain $\mathcal{L}_{j+1}$ early and enable it to partially intersect the same edge gadget along with $\mathcal{L}_{j-1}$. Thus, their reduction needed intervals of two lengths for the vertex, edge and join gadgets, and another two lengths as the link intervals, totalling to an interval count of four.

In the present work, we bring the interval count by two due to our modifications in the structure of gadgets and link chains. Instead of using separate lengths of short and long intervals, the vertex, edge and join gadgets are now three consecutive cliques composed of intervals of the same length, with the middle clique being twice as large as the two other cliques. The link chains do not need two separate lengths of intervals anymore. The problem of non-consecutive link chains partially intersecting the same edge gadget is solved by using a *switch gadget*, which again is a sequence of cliques, that changes the relative positions of join gadgets of link chains. This brings the interval count down to two.

A.4 Overview of the reduction

We reduce MaxCut on cubic graphs to MaxCut on interval graphs of interval count two. For every cubic graph G, we construct an interval graph H of interval count 2 such that any MaxCut partition of G corresponds to some MaxCut partition of H. In this section, we provide basic information about the reduction.

Interval sizes We can use intervals of only two different sizes: *short* and *long*. Later, when we give an explicit proof, we assume that short intervals have length 1, and long intervals have length α, for some $\alpha > 1$ that depends on the size of the input cubic graph.

Blocks A *block* is an interval model consisting of short intervals that start from and terminate at the same coordinates. It is usually denoted by a letter $\mathcal{B}$. The size $|\mathcal{B}|$ of a block $\mathcal{B}$ is the number of intervals within it. We can without loss of generality draw blocks as squares. A block of size x is an interval model of the complete graph K_x on x vertices. When an interval model contains several blocks, they can be linearly ordered. The ordering depends on the coordinates of the left ends of the blocks. Every gadget that we use is constructed from blocks.

Block coloring Consider any interval model consisting of blocks $\mathcal{B}_1, \ldots, \mathcal{B}_s$, they are linearly ordered from the leftmost to the rightmost, say that $\mathcal{B}_1$ is the leftmost and $\mathcal{B}_s$ is the rightmost. Then we say that the coloring of the blocks is *alternating* if, for every $i \in [s-1]$, $\mathcal{B}_i$ and $\mathcal{B}_{i+1}$ are entirely in different classes, i.e., $\text{Color}(\mathcal{B}_i) \neq \text{Color}(\mathcal{B}_{i+1})$. The coloring is *almost alternating except for* $(\mathcal{B}_i, c)$ if there is a constant value c such that at most c intervals of the block $\mathcal{B}_i$ have the same color as $\mathcal{B}_{i-1}$ and $\mathcal{B}_{i+1}$ and, for every $j \notin \{i-1, i\}$, we have $\text{Color}(\mathcal{B}_j) \neq \text{Color}(\mathcal{B}_{j+1})$.

Size of the reduction graph Before we formally describe the reduction graph H, we define all the gadgets that are used in the construction of H. It is convenient to prove the necessary properties of a gadget right after it is defined. Sometimes, for a property to hold, it is required that H satisfies certain numerical constraints. We list them now, and, when we explicitly construct H, we make sure that these constraints hold. Suppose that $|V(G)| = n$; then $|E(G)| = \frac{3n}{2}$.

1. The size of every block of H depends either on a parameter x (for vertex, join, and switch gadgets) or on a parameter k (for edge gadgets). It is required that $x, k \in \Omega(n^6)$.

2. For every block of H there are at most $\eta := 3n$ long intervals that intersect it.

3. Every long interval of H intersects at most $\mu := 30n + 16$ blocks.

A.5 3-blocks

A *3-block* is an interval model consisting of three blocks $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$ such that $|\mathcal{B}_1| = |\mathcal{B}_3|$ and $|\mathcal{B}_2| = 2|\mathcal{B}_1|$. Every interval of $\mathcal{B}_1$ intersects every interval of $\mathcal{B}_1 \sqcup \mathcal{B}_2$ and no interval

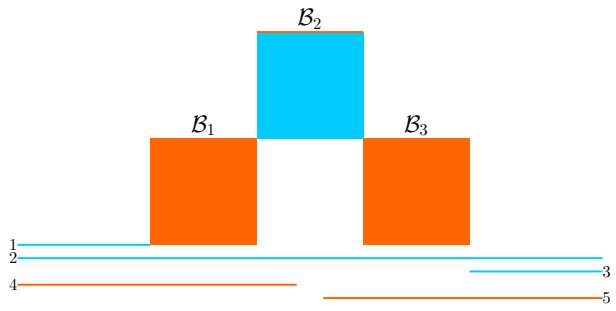


Figure A.2: A 3-block interval model together with five long link intervals intersecting it in any possible way.

of $\mathcal{B}_3$. Every interval of $\mathcal{B}_2$ intersects any other interval of the model. Every interval of $\mathcal{B}_3$ intersects every interval of $\mathcal{B}_2 \sqcup \mathcal{B}_3$ and no interval of $\mathcal{B}_1$.

Usually, we draw a 3-block as it is done on Figure A.2. We say that an interval *terminates at* a block $\mathcal{B}$ if $\mathcal{B}$ is the rightmost block that contains the right end of the interval. Similarly, we say that an interval *starts from* $\mathcal{B}$ if $\mathcal{B}$ is the leftmost block that contains the left end of the interval. For a long interval, we say that it *overlaps* $\mathcal{B}$ if it intersects the block but neither terminates at it nor starts from it. For example, on Figure A.2, there is a short interval that terminates at $\mathcal{B}_1$, another short interval that starts from $\mathcal{B}_3$, and three long intervals that overlap all the three blocks.

3-block is a basic concept of this paper because almost all other gadgets are constructed using 3-blocks. The reason of its usefulness is that, if the block size is large enough, then, under any MaxCut coloring, the coloring of the blocks $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$ of a 3-block is almost alternating except for $(\mathcal{B}_2, c)$; and the value of c depends on the number of long intervals that overlap $\mathcal{B}_2$. This is proved in Lemma A.5.1. Denote the total number of link intervals by η. In the following lemma we show that a 3-block follows an alternating pattern of partition for all maximum sized cuts, except a few intervals in the middle block.

Lemma A.5.1. *Let $\mathcal{B} = (\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3)$ be a 3-block in H of size $(x, 2x, x)$. Then any maximum cut coloring of H satisfies the following conditions:*

1. *all intervals of $\mathcal{B}_1$ and $\mathcal{B}_3$ get the same color.*
2. *all intervals of $\mathcal{B}_2$ get the opposite color except for at most η intervals.*

Proof. Consider a maximum cut coloring of H.

- We denote by $\mathcal{L}_1$ all the long intervals that terminate at $\mathcal{B}_1$.
- We denote by $\mathcal{L}_2$ all the long intervals that overlap $\mathcal{B}_1$, $\mathcal{B}_2$, and $\mathcal{B}_3$.
- Those long intervals that start from $\mathcal{B}_3$ are denoted by $\mathcal{L}_3$.
- Those long intervals that terminate at $\mathcal{B}_2$ are denoted by $\mathcal{L}_4$.
- Finally, $\mathcal{L}_5$ denotes all the long intervals that start from $\mathcal{B}_2$.

Let r_i and b_i be the numbers of intervals that respectively get red and blue colors in $\mathcal{L}_i$, for any $i \in [5]$. Without loss of generality, assume that most intervals of $\mathcal{L}_1$ get the color red, that is, $r_1 > b_1$.

Also let $y_1, 2x - y_2$, and y_3 be the number of intervals of respectively $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$ that are colored blue. Hence $x - y_1, y_2$, and $x - y_3$ are the numbers of intervals of respectively $\mathcal{B}_1, \mathcal{B}_2$, and $\mathcal{B}_3$ that are colored red. Now observe that

1. The number of cut edges formed between the blocks and the long intervals is $r_1y_1 + b_1(x - y_1) + r_2(y_1 + 2x - y_2 + y_3) + b_2(2x - y_1 + y_2 - y_3) + r_3y_3 + b_3(x - y_3) + r_4(y_1 + 2x - y_2) + b_4(x - y_1 + y_2) + r_5(2x - y_2 + y_3) + b_5(y_2 + x - y_3)$.

2. The number of cut edges formed by the short intervals of the same block among themselves is $y_1(x - y_1) + y_2(2x - y_2) + y_3(x - y_3)$.

3. The number of cut edges formed by the short intervals of the different blocks is $y_1y_2 + (x - y_1)(2x - y_2) + y_2y_3 + (2x - y_2)(x - y_3)$.

Denote the quantity $r_i - b_i$ by Δ_i. Therefore the number of cut edges, denoted by C, contributed by the 3-block $\mathcal{B}$ is:

$$\begin{aligned} C =& 4x^2 + x(2r_2 + 2r_4 + 2r_5 + b_1 + 2b_2 + b_3 + b_4 + b_5) - \\ & - xy_1 - xy_3 - y_1^2 - y_2^2 - y_3^2 + 2y_1y_2 + 2y_2y_3 + \\ & + \Delta_1y_1 + \Delta_2(y_1 - y_2 + y_3) + \Delta_3y_3 + \Delta_4(y_1 - y_2) + \Delta_5(y_3 - y_2). \end{aligned}$$

Now let us modify the cut by making $\text{Color}(\mathcal{B}_1) = \text{Color}(\mathcal{B}_3) = R$ and $\text{Color}(\mathcal{B}_2) = B$. This means that we have $y_1 = y_2 = y_3 = 0$. So the number of cut edges, denoted by C_{RBR}, contributed by the 3-block $\mathcal{B}$ in this case is:

$$C_{RBR} = 4x^2 + x(2r_2 + 2r_4 + 2r_5 + b_1 + 2b_2 + b_3 + b_4 + b_5).$$

If we instead have $\text{Color}(\mathcal{B}_1) = \text{Color}(\mathcal{B}_3) = B$ and $\text{Color}(\mathcal{B}_2) = R$ (i.e., $y_1 = y_3 = x, y_2 = 2x$), then the number of cut edges, denoted by C_{BRB}, contributed by the 3-block $\mathcal{B}$ would be:

$$C_{BRB} = 4x^2 + x(r_1 + 2r_2 + r_3 + r_4 + r_5 + 2b_2 + 2b_4 + 2b_5).$$

Without loss of generality, assume that $C_{RBR} \geq C_{BRB}$. This implies that

$$\begin{aligned} & C_{RBR} - C_{BRB} \geq 0 \implies \\ & \Delta_1 + \Delta_3 - \Delta_4 - \Delta_5 \leq 0. \end{aligned}$$

Now observe that $C - C_{RBR}$ equals

$$
\begin{aligned}
&- y_1^2 - y_2^2 - y_3^2 - xy_1 - xy_3 + 2y_1y_2 + 2y_2y_3 \\
&\quad + \Delta_1 y_1 + \Delta_2(y_1 - y_2 + y_3) + \Delta_3 y_3 + \Delta_4(y_1 - y_2) + \Delta_5(y_3 - y_2) = \\
&= - \left[(y_1 - y_2 + y_3) - \frac{1}{2}(\Delta_2 + \Delta_4 + \Delta_5)\right]^2 + \frac{1}{4}(\Delta_2 + \Delta_4 + \Delta_5)^2 + \\
&+ 2y_1y_3 - xy_1 - xy_3 + \Delta_1 y_1 + \Delta_3 y_3 - \Delta_4 y_3 - \Delta_5 y_1 \leq \\
&\leq \frac{1}{4}(\Delta_2 + \Delta_4 + \Delta_5)^2 + 2y_1y_3 - xy_1 - xy_3 + (\Delta_1 - \Delta_5)y_1 + (\Delta_3 - \Delta_4)y_3 = \\
&= \frac{(\Delta_2 + \Delta_4 + \Delta_5)^2}{4} + 2y_1y_3 - xy_1 - xy_3 + (\Delta_1 - \Delta_5)(y_1 - y_3) + (\Delta_1 + \Delta_3 - \Delta_4 - \Delta_5)y_3 \\
&\leq \frac{1}{4}(\Delta_2 + \Delta_4 + \Delta_5)^2 + 2y_1y_3 - xy_1 - xy_3 + (\Delta_1 - \Delta_5)(y_1 - y_3) = \\
&\quad \text{(since we assumed that } \Delta_1 + \Delta_3 - \Delta_4 - \Delta_5 \leq 0) \\
&= \frac{(\Delta_2 + \Delta_4 + \Delta_5)^2}{4} + \frac{(\Delta_1 - \Delta_5)^2}{4} + y_1(y_1 - x) + y_3(y_3 - x) - \left(y_1 - y_3 - \frac{(\Delta_1 - \Delta_5)}{2}\right)^2 \\
&\leq \frac{1}{4}(\Delta_2 + \Delta_4 + \Delta_5)^2 + \frac{1}{4}(\Delta_1 - \Delta_5)^2 + y_1(y_1 - x) + y_3(y_3 - x).
\end{aligned}
$$

Notice that if $y_1 \notin \{0, x\}$, then $(x - y_1)y_1 \geq x - 1$. Similarly, if $y_3 \notin \{0, x\}$, then $(x - y_3)y_3 \geq x - 1$. Since we assumed that either $(y_1 \notin \{0, x\})$ or $(y_3 \notin \{0, x\})$, it implies that $y_1(y_1 - x) + y_3(y_3 - x) \leq 1 - x$. So we have $C - C_{RBR} \leq \frac{1}{4}(\Delta_2 + \Delta_4 + \Delta_5)^2 + \frac{1}{4}(\Delta_1 - \Delta_5)^2 + 1 - x$. But since $x > n^6$ and each $\Delta_i < n^3$, we conclude that $C - C_{RBR} < 0 \implies C < C_{RBR}$. This contradicts the maximality of the first cut. Therefore we conclude that $\big(y_1 \in \{0, x\}\big)$ and $\big(y_3 \in \{0, x\}\big)$, i.e., $\mathcal{B}_1$ and $\mathcal{B}_3$ are monochromatic.

Now we show that $\text{Color}(\mathcal{B}_1) = \text{Color}(\mathcal{B}_3)$. For the sake of contradiction, assume that $\text{Color}(\mathcal{B}_1) \neq \text{Color}(\mathcal{B}_3)$. Thus, $|y_1 - y_3| = x$. Again, we calculate the difference $C - C_{RBR}$, which is at most:

$$
\begin{aligned}
&\frac{1}{4}(\Delta_2 + \Delta_4 + \Delta_5)^2 + \frac{1}{4}(\Delta_1 - \Delta_5)^2 + y_1(y_1 - x) + y_3(y_3 - x) - \left(y_1 - y_3 - \frac{1}{2}(\Delta_1 - \Delta_5)\right)^2 \\
&= \frac{1}{4}(\Delta_2 + \Delta_4 + \Delta_5)^2 + \frac{1}{4}(\Delta_1 - \Delta_5)^2 - \left(y_1 - y_3 - \frac{1}{2}(\Delta_1 - \Delta_5)\right)^2 \\
&\leq \frac{1}{4}(\Delta_2 + \Delta_4 + \Delta_5)^2 - x^2 + x|\Delta_1 - \Delta_5|.
\end{aligned}
$$

Since $x > n^6$ and each $\Delta_i < n^3$, we have that $C - C_{RBR} < 0$, which contradicts the maximality of the first cut. So we conclude that $\text{Color}(\mathcal{B}_1) = \text{Color}(\mathcal{B}_3)$, this proves (1).

Without loss of generality, suppose that $\text{Color}(\mathcal{B}_1) = \text{Color}(\mathcal{B}_3) = R$. So in order to prove (2), we have to show that except at most η intervals, all intervals of $\mathcal{B}_2$ get color blue. The number of cut edges contributed by $\mathcal{B}$ is

$$C = 4x^2 + x(2r_2 + 2r_4 + 2r_5 + b_1 + 2b_2 + b_3 + b_4 + b_5). - y_2(y_2 + \Delta_2 + \Delta_4 + \Delta_5).$$

Now we show that $y_2 \leq b_2 + b_4 + b_5$. For the sake of contradiction assume that $y_2 > b_2 + b_4 + b_5$. Let us modify the cut by coloring all intervals of $\mathcal{B}_2$ blue, i.e. by making $y_2 = 0$. The number of cut edges contributed by $\mathcal{B}$ in the modified cut is:

$$C_{RBR} = 4x^2 + x(2r_2 + 2r_4 + 2r_5 + b_1 + 2b_2 + b_3 + b_4 + b_5).$$

Since we have only modified y_2, the difference in cut sizes of these two cuts is:

$$C - C_{RBR} = -y_2(y_2 + r_2 - b_2 + r_4 - b_4 + r_5 - b_5)$$

Since $y_2 > b_2 + b_4 + b_5$, we conclude that $C - C_{RBR} < 0$, which contradicts C being a maximal cut. Hence $y_2 < b_2 + b_4 + b_5 < \eta$.

□

A.6 Gadgets

After defining a 3-block and showing how it is colored under a MaxCut partition, we can use it to construct different gadgets of the reduction graph H. For each gadget that we introduce, we state in a corresponding lemma how it can be colored under a MaxCut partition of H. For gadgets that are based on 3-blocks, these lemmas follow from Lemma A.5.1 and from the gadget construction. Although there is a gadget whose construction is not based on 3-blocks, the lemma that describes its MaxCut partition within H is proved using methods similar to Lemma A.5.1.

Vertex gadget The interval model of a vertex gadget consists of three 3-blocks and two short intervals: $\mathcal{V} := \mathcal{B}^1 \sqcup \mathcal{B}^2 \sqcup \mathcal{B}^3 \sqcup \{S_{12}, S_{23}\}$. The blocks $\mathcal{B}^i_1, \mathcal{B}^i_2, \mathcal{B}^i_3$ of each 3-block $\mathcal{B}^i$ have sizes $x, 2x, x$ correspondingly, for some $x > 0$. The 3-blocks are connected by short intervals S_{12}, S_{23}, they are called *short link* intervals. There are three long intervals L_1, L_2, L_3 adjacent to a vertex gadget, they start from $\mathcal{B}^1_3, \mathcal{B}^2_3, \mathcal{B}^3_3$, where, for each i in [3], $\mathcal{B}^i_3$ is the rightmost block of $\mathcal{B}^i$.

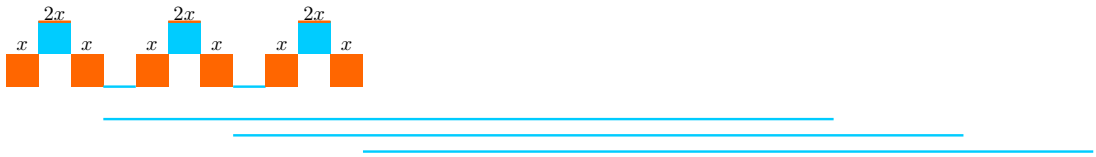


Figure A.3: A vertex gadget interval model.

Under any MaxCut partition, the coloring of each 3-block of a vertex gadget is almost alternating. Moreover, each pair of blocks that are connected by a short link interval must have the same color, and the short link interval must have the opposite color. The colors of L_1, L_2, L_3 are opposite to the colors of blocks from which they start. An optimal coloring of a vertex gadget, under a MaxCut partition, is displayed on Figure A.3. We prove this in Lemma A.6.1.

Lemma A.6.1. *Let $\mathcal{V}$ be a vertex gadget of H (see Figure A.3). Then, in a* MaxCut *partition of H, the two following statements hold:*

1. *for all $i \in [3]$, all the intervals of the blocks $\mathcal{B}^i_1$ and $\mathcal{B}^i_3$ belong to the same class, and*
2. *the intervals of each of the blocks $\mathcal{B}^1_2, \mathcal{B}^2_2$, and $\mathcal{B}^3_2$, except for at most η intervals within each block, belong to the other class, as well as S_{12} and S_{23}.*

Proof. By Lemma A.5.1, for $i \in [3]$, $\mathcal{B}^i$ must be almost alternating except for η intervals in $\mathcal{B}^i_2$.

If two adjacent 3-blocks are colored differently, then the short link interval between them will add x to the MaxCut. If they have the same color, then it will add $2x$, so all the three 3-blocks will have a similar partition: either all RBR or all BRB except for a small fraction in every central block $\mathcal{B}_2^i$. □

Let $v \in V(G)$ be a vertex of a cubic graph G. It corresponds to some vertex gadget. Then the color of v, under a MaxCut partition, is the same as the color assigned to the long intervals starting from the vertex gadget.

Edge gadget An edge gadget consists of two 3-blocks connected by a short link interval: $\mathcal{E} := \mathcal{B}^1 \sqcup \mathcal{B}^2 \sqcup \{S_{12}\}$. The blocks $\mathcal{B}_1^i, \mathcal{B}_2^i, \mathcal{B}_3^i$ of each 3-block $\mathcal{B}^i$ have sizes $k, 2k, k$ correspondingly, for some $k > 0$. The value of k is different from the value of x – the block size within a vertex gadget. There are two long intervals L_1, L_2 that terminate at $\mathcal{B}_1^1$ and $\mathcal{B}_2^2$ correspondingly, see Figure A.4.

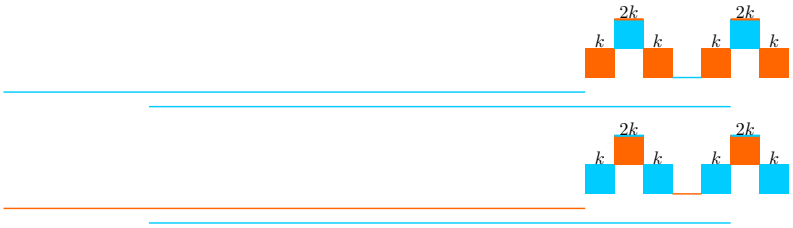


Figure A.4: The two cases of the MaxCut coloring of an edge gadget.

The colors of L_1 and L_2 are enforced by some two vertex gadgets. If we choose k to be significantly smaller than x, then edge gadgets are less influential than vertex gadgets. The colorings of $\mathcal{B}^1$ and $\mathcal{B}^2$ of an edge gadget must be almost alternating and they depend on the colors of L_1 and L_2. The two possible cases are displayed on Figure A.4. If $\text{Color}(L_1) \neq \text{Color}(L_2)$, then the MaxCut value is greater, as S_{12} is connected to blocks of the same color and it can pick the opposite color. If $\text{Color}(L_1) = \text{Color}(L_2)$, then S_{12} is adjacent to blocks of different colors, and the MaxCut value is less than in the first case. This is proved in Lemma A.6.2. Suppose that there is a graph H and a maximum cut partition of H; then, for some induced subgraph H' of H, the *value provided by* H' means the number of cut edges having at least one end in H'.

Lemma A.6.2. *Let $\mathcal{E} = \mathcal{B}^1 \sqcup \mathcal{B}^2 \sqcup \{S_{12}\}$ be an edge gadget of H together with two long intervals L_1, L_2 terminating at $\mathcal{B}_1^1$ and $\mathcal{B}_2^2$, see Figure A.4. Suppose that, for $i \in [2]$, $|\mathcal{B}_1^i| = |\mathcal{B}_3^i| = k$ and $|\mathcal{B}_2^i| = 2k$. Then, for a* MaxCut *partition of H, the colorings of 3-blocks are almost alternating except for at most η intervals of the central blocks; and if $\text{Color}(L_1) \neq \text{Color}(L_2)$, then the maximum cut value provided by $\mathcal{E}$ is greater at least by $k - 2\eta$ than the maximum cut value in the case $\text{Color}(L_1) = \text{Color}(L_2)$.*

Proof. Suppose that $\mathcal{E}$ is overlapped by $\lambda < \eta$ long intervals. By Lemma A.5.1, the MaxCut of $\mathcal{B}^1$ and $\mathcal{B}^2$, is either alternating or almost alternating except for Δ intervals of $\mathcal{B}_2^1$ or $\mathcal{B}_2^2$, where $\Delta < \eta$ is the difference between the numbers of red and blue long overlapping intervals.

Suppose that $\text{Color}(L_1) = \text{Color}(L_2) = B$; then $\text{Color}(\mathcal{B}_1^1) = \text{Color}(\mathcal{B}_3^1) = R$. The color of $\mathcal{B}_2^1$ must be blue except for at most η intervals. For $\mathcal{B}^2$ there are two cases.

1. The color of $\mathcal{B}_2^2$ is red except for at most η intervals, and $\text{Color}(\mathcal{B}_1^2) = \text{Color}(\mathcal{B}_3^2) = B$. Then, for any $\text{Color}(S_{12})$, the cut size is at most $8k^2 + (6 + 4\lambda)k + \frac{1}{4}\Delta^2 + \frac{3}{2}\lambda + 1$.

2. The color of $\mathcal{B}_2^2$ is blue except for at most η intervals, and $\text{Color}(\mathcal{B}_1^2) = \text{Color}(\mathcal{B}_3^2) = R$. Then $\text{Color}(S_{12}) = B$. Then the cut size is at most $8k^2 + (6+4\lambda)k + \frac{1}{2}\Delta^2 + 2\lambda$.

Suppose that $\text{Color}(L_2) = B$ and $\text{Color}(L_1) = R$. Then $\mathcal{B}^2$ is partitioned as RBR. $\mathcal{B}^1$ is also partitioned as RBR. Then the cut size would be at least $8k^2 + (7+4\lambda)k + \frac{1}{2}\Delta^2 + 2$.

The difference between the value provided by $\mathcal{E}$ when $\text{Color}(L_1) \neq \text{Color}(L_2)$ and the value in the case when they are equal is at least $k - 2\lambda$, which is at least $k - 2\eta$. □

Join gadget Similarly to vertex gadgets, a *join gadget* consists of three 3-blocks connected by short link intervals: $\mathcal{J} = \mathcal{B}^1 \sqcup \mathcal{B}^2 \sqcup \mathcal{B}^3 \sqcup \{S_{12}, S_{23}\}$. Every join gadget has 3 long intervals terminating at $\mathcal{B}_1^i$ and 3 long intervals starting from $\mathcal{B}_3^i$, for $i \in [3]$, see Figure A.5. We need such gadgets to connect vertex gadgets to edge gadgets by chains of long intervals.

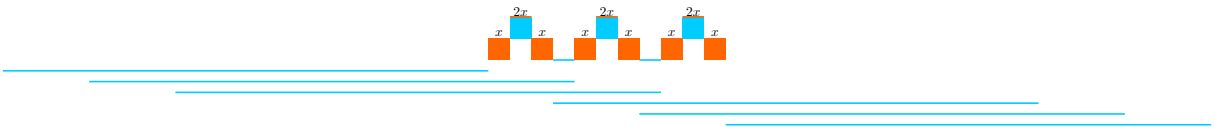


Figure A.5: The join gadget with its MaxCut partition.

By Lemma A.5.1, the coloring of each 3-block $\mathcal{B}^i$ is almost alternating except for some small fraction within its central block $\mathcal{B}_2^i$. All the 3-blocks are colored similarly: either all are BRB or all are RBR. See Lemma A.6.4 for the proof. The colors of the three long intervals that start from a join gadget are the same as the colors of those three that terminate at it. This means that we can keep in our memory the vertex gadget coloring as it is preserved by join gadgets.

Stretch gadget We need to be able to manipulate the positions of long intervals within a join gadget. Either we can increase the number of 3-blocks inside it or we can decrease the distances between blocks without adding any new intersections, as on Figure A.6. Such join gadgets are called *stretch gadgets*. We will never need more than 10 3-blocks within one stretch gadget.

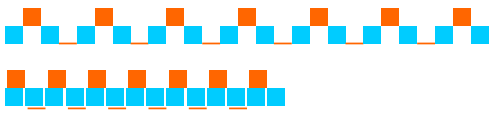


Figure A.6: The two extreme cases to compress a join gadget.

The following lemma provides a lower bound for the distance between the ends of two long intervals of H, it is a consequence of the stretch gadget construction.

Lemma A.6.3. *If the distance between the ends of two long intervals L_1, L_2 is greater than 2, then we can attach them to the same stretch gadget.*

Proof. See Figure A.6. Assume that L_1, L_2 intersect a stretch gadget and do not overlap it. By construction, a long interval can terminate at the first block $\mathcal{B}_1^i$ of some 3-block $\mathcal{B}^i$, and can start from the third one. Consider 3 cases:

1. L_1 and L_2 terminate at (start from) the gadget: L_1 at a, L_2 at b, and $a < b$.

2. L_1 terminates at a and L_2 starts from b and $a < b$.

3. L_1 terminates at a and L_2 starts from b and $a > b$.

If L_1 has its end in a 3-block $\mathcal{B}^i$, and L_2 – in $\mathcal{B}^{i'}$, then we say that the intervals have $|i' - i|$ 3-blocks between them. In particular, if they have their ends in the same block, then they have no 3-blocks between them.

Suppose that there are j 3-blocks between the intervals.

Case 1. The ends cannot be in the same 3-block. The distance between the ends can be in $(2j, 4j]$, for $j \geq 1$.

Case 2. The distance can be in $(2j+1, 4j+3]$, for $j \geq 0$.

Case 3. The ends cannot be in the same 3-block. The distance can be in $(2j-2, 4j-3]$, for $j \geq 1$.

One can see that the union of these three intervals, in each case, will contain $(2, +\infty)$. □

We now show similar bounds on the MaxCut for stretch and switch gadgets.

Lemma A.6.4. *Consider a stretch gadget of H consisting of m 3-blocks, each of them being of sizes $x, 2x, x$. Then, for a* MaxCut *partition of H, all the colorings of the 3-blocks are the same and almost alternating except for at most η intervals in the central blocks, and the color of each short link interval is opposite to the color of the two adjacent blocks, as on Figure A.5.*

Proof. By Lemma A.5.1, the coloring of $\mathcal{B}^j$, for $j \in [m]$, must be almost alternating except for η intervals in $\mathcal{B}^j_2$. Suppose that $\mathcal{B}^j$ and $\mathcal{B}^{j+1}$ are colored differently, for some $j \in [m-1]$. Then we invert the coloring of $\mathcal{B}^{j+1}$ so that now they have the same color. Do it for any other $\mathcal{B}^{j'}$, for $k > j+1$, if needed. The possible loss is at most $m \times \eta^2$. The possible gain is at least x, as now, for any short link interval $S_{j,j+1}$, its color is different than the color of both $\mathcal{B}^j_3, \mathcal{B}^{j+1}_1$. As $x \in \Omega(n^6)$ and $\eta < 3n$, this partition is MaxCut. □

Switch gadget A *switch gadget* consists of two parts. The first part that is the most important one is displayed on Figure A.7. It is the only gadget that is not constructed from 3-blocks. It consists of 9 bottom blocks $\mathcal{B}^{bot}_1, \ldots, \mathcal{B}^{bot}_9$ and 4 top blocks $\mathcal{B}^{top}_1, \ldots, \mathcal{B}^{top}_4$. At the bottom, $|\mathcal{B}^{bot}_1| = |\mathcal{B}^{bot}_9| = x$, and, for $2 \leq i \leq 8$, $|\mathcal{B}^{bot}_i| = 2x$. At the top, for any $i \in [4]$, $|\mathcal{B}^{top}_i| = 2x'$, where x' is some value satisfying $x/2 < x' < x$. It also contains two long intervals L_1, L_2 that terminate at the gadget, and two long intervals R_1, R_2 that start from it, where, for each i, the end of L_i is to the left of the end of R_i. L_1 terminates at the bottom left block, L_2 terminates at the top left one. R_1 starts from the top right block, R_2 starts from the bottom right one. The main property of this gadget is that, for any MaxCut partition, we have $\text{Color}(L_1) = \text{Color}(R_2)$ and $\text{Color}(L_2) \neq \text{Color}(R_1)$.

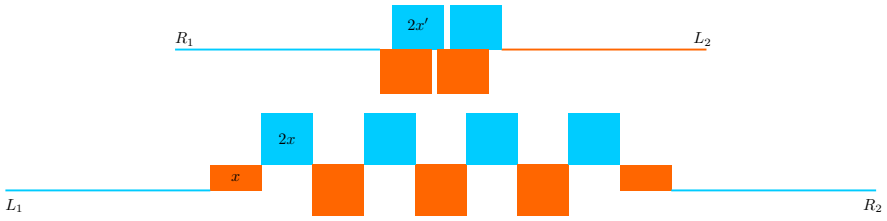


Figure A.7: The first part of a switch gadget.

The second part, see Figure A.8, is just a 3-block that changes the color of the successor of R_1 to be the same as Color(L_2).

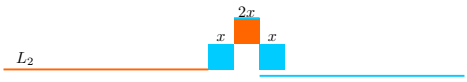


Figure A.8: The second part of a switch gadget.

The idea behind switch gadget is that it "switches" the intervals: the color of R_1 depends on the color of L_2, and the color of R_2 depends on the color of L_1.

A MaxCut partition of the first part of the switch gadget is shown on Figure A.7, this is proved in Lemma A.6.5. If the first switch gadget part is overlapped by many long intervals, then, under a MaxCut partition, the coloring of the blocks will be the same as on Figure A.7 except for at most c intervals in $\mathcal{B}_1^{top}$ or $\mathcal{B}_4^{top}$. This is proved in Lemma A.6.6. Here, c also depends on the size of the input cubic graph. A MaxCut partition of the second part is similar to the one of a 3-block.

Lemma A.6.5. *Consider a graph with an interval model as on Figure A.7. Then, in any* MaxCut *partition of this graph, the following statements hold:*

- *for each of the two levels, block colorings are alternating, and*
- *Color*$(R_1) \neq$ *Color*(L_2) *and Color*$(R_2) =$ *Color*(L_1).

Proof of Lemma A.6.5. At first, let us compute the size of the cut for any partition, where the colors of the blocks on the top and on the bottom alternate. We do not consider the four long intervals at the moment. The value is the same for any of the four such partitions:

$$\text{Cut}_{alter} = 3 \cdot 4x'^2 + 2 \cdot 2x^2 + 6 \cdot 4x^2 + 4 \cdot 4x'x = 12x'^2 + 28x^2 + 16x'x.$$

Suppose now that the alternating partition is not maximal. Then we will introduce the following notations for each block.

- Denote by y_1, y_3 the numbers of blue intervals in $\mathcal{B}_1^{top}$ and $\mathcal{B}_3^{top}$.
- Similarly, y_2, y_4 denote the numbers of red intervals in $\mathcal{B}_2^{top}$ and $\mathcal{B}_4^{top}$.
- Denote by z_1, z_3, z_5, z_7, z_9 the numbers of blue intervals in $\mathcal{B}_1^{bot}, \mathcal{B}_3^{bot}, \mathcal{B}_5^{bot}, \mathcal{B}_7^{bot}, \mathcal{B}_9^{bot}$.
- Similarly, denote by z_2, z_4, z_6, z_8 the numbers of red intervals in $\mathcal{B}_2^{bot}, \mathcal{B}_4^{bot}, \mathcal{B}_6^{bot}, \mathcal{B}_8^{bot}$.

We are going to compute the MaxCut for a general case and to compare it to Cut_{alter} in order to find out when it can be the maximal. We are going to split the computation into several parts in order to be clear:

- The Cut_{inside} part counts the cut-edges inside each of the blocks.
- The Cut_y and Cut_z parts count the cut-edges between different blocks that are both on the same level. There are two levels: the top y and the bottom z.
- The Cut_{inter} part counts the cut-edges between the two levels.

Now we compute the parts one by one.

$$\textsc{Cut}_{inside} = \sum_{i=1}^{4} y_i(2x' - y_i) + z_1(x - z_1) + \sum_{j=2}^{8} z_j(2x - z_j) + z_9(x - z_9) =$$
$$= -\sum_{i=1}^{4} y_i^2 - \sum_{j=1}^{9} z_j^2 + 2x' \sum_{i=1}^{4} y_i + x(z_1 + 2\sum_{j=2}^{8} z_j + z_9).$$

$$\textsc{Cut}_y = \sum_{i=1}^{3} \big(y_i y_{i+1} + (2x' - y_i)(2x' - y_{i+1})\big) =$$
$$= 2\sum_{i=1}^{3} y_i y_{i+1} + 12x'^2 - x'(2y_1 + 4\sum_{i=2}^{3} y_i + 2y_4).$$

$$\textsc{Cut}_z = z_1 z_2 + (x - z_1)(2x - z_2) + \sum_{j=2}^{7} \big(z_j z_{j+1} + (2x - z_j)(2x - z_{j+1})\big) +$$
$$+ z_8 z_9 + (2x - z_8)(x - z_9) = 2\sum_{j=1}^{8} z_j z_{j+1} + 28x^2 - x(2z_1 + 3z_2 + 4\sum_{j=3}^{7} z_j + 3z_8 + 2z_9).$$

Denote by A the set of pairs $\big\{(i,j) \mid \mathcal{B}_i^{top}$ intersects $\mathcal{B}_j^{bot}\big\}$. That is,

$$A = \Big\{(1,4),(1,5),(2,4),(2,5),(3,5),(3,6),(4,5),(4,6)\Big\}.$$

Then

$$\textsc{Cut}_{inter} = \sum_{(i,j)\in A, i+j \text{ is odd}} \big(y_i z_j + (2x' - y_i)(2x - z_j)\big) +$$
$$+ \sum_{(i,j)\in A, i+j \text{ is even}} \big(y_i(2x - z_j) + (2x' - y_i)z_j\big) = 16x'x + 2\sum_{i,j\in A} (-1)^{i+j+1} y_i z_j =$$
$$= 16x'x + 2(y_1 - y_2)(z_4 - z_5) - 2(y_3 - y_4)(z_5 - z_6).$$

Our goal is to prove that $f = \textsc{Cut}_{inside} + \textsc{Cut}_y + \textsc{Cut}_z + \textsc{Cut}_{inter} - \textsc{Cut}_{alter}$ is less or equal to 0 and the equality is reached only when the colors alternate. Think of f as of a polynomial of x' and x of degree 2:

$$f = f(x', x) = f_0 + f_1 x' + f_2 x + f_3 x'^2 + f_4 x' x + f_5 x^2.$$

Clearly, $f_3 = f_4 = f_5 = 0$. Now compute the terms f_0, $f_1 x'$, and $f_2 x$:

$$f_0 = \underbrace{-\sum_{i=1}^{4} y_i^2 - \sum_{j=1}^{9} z_j^2}_{\textsc{Cut}_{inside}} + \underbrace{2\sum_{i=1}^{3} y_i y_{i+1}}_{\textsc{Cut}_y} + \underbrace{2\sum_{j=1}^{8} z_j z_{j+1}}_{\textsc{Cut}_z} +$$
$$+ \underbrace{2(y_1 - y_2)(z_4 - z_5) - 2(y_3 - y_4)(z_5 - z_6)}_{\textsc{Cut}_{inter}};$$

$$f_1x' = \underbrace{2x'\sum_{i=1}^{4} y_i}_{\textsc{Cut}_{inside}} \underbrace{- x'(2y_1 + 4\sum_{i=2}^{3} y_i + 2y_4)}_{\textsc{Cut}_y} = -2x'\sum_{i=2}^{3} y_i = -\sum_{i=2}^{3} y_i^2 - \sum_{i=2}^{3}(2x' - y_i)y_i;$$

$$f_2x = \underbrace{x(z_1 + 2\sum_{j=2}^{8} z_j + z_9)}_{\textsc{Cut}_{inside}} \underbrace{- x(2z_1 + 3z_2 + 4\sum_{j=3}^{7} z_j + 3z_8 + 2z_9)}_{\textsc{Cut}_z} =$$

$$= -x(z_1 + z_2 + 2\sum_{j=3}^{7} z_j + z_8 + z_9) =$$

$$= -z_1^2 - \frac{z_2^2}{2} - \sum_{j=3}^{7} z_j^2 - \frac{z_8^2}{2} - z_9^2 -$$

$$- (x - z_1)z_1 - \left(x - \frac{z_2}{2}\right) z_2 - \sum_{j=3}^{7}(2x - z_j)z_j - \left(x - \frac{z_8}{2}\right) z_8 - (x - z_9)z_9.$$

We have extracted the negative squares of y_i, z_j from f_1x' and f_2x in order to combine them together with the negative squares of the part of f_0 provided by $\textsc{Cut}_{inside}$. The other summands of f_1x' and f_2x are almost always negative, except for the minimal and maximal values of y_i, z_j. These extreme cases happen exactly when a block is either all red or all blue. Then

$$f = f_0 + f_1x' + f_2x = 2\sum_{i=1}^{3} y_iy_{i+1} - y_1^2 - 2\sum_{i=2}^{3} y_i^2 - y_4^2 +$$

$$+ 2\sum_{j=1}^{8} z_jz_{j+1} - 2z_1^2 - \frac{3}{2}z_2^2 - 2\sum_{j=3}^{7} z_j^2 - \frac{3}{2}z_8^2 - 2z_9^2 +$$

$$+ 2(y_1 - y_2)(z_4 - z_5) - 2(y_3 - y_4)(z_5 - z_6) -$$

$$- \sum_{i=2}^{3}(2x' - y_i)y_i -$$

$$- (x - z_1)z_1 - \left(x - \frac{z_2}{2}\right) z_2 - \sum_{j=3}^{7}(2x - z_j)z_j - \left(x - \frac{z_8}{2}\right) z_8 - (x - z_9)z_9 =$$

$$= -\sum_{i=1}^{3}(y_i - y_{i+1})^2 - 2\left(z_1 - \frac{z_2}{2}\right)^2 - \sum_{j=2}^{7}(z_j - z_{j+1})^2 - 2\left(\frac{z_8}{2} - z_9\right)^2 +$$

$$+ 2(y_1 - y_2)(z_4 - z_5) - 2(y_3 - y_4)(z_5 - z_6) -$$

$$- \sum_{i=2}^{3}(2x' - y_i)y_i -$$

$$- (x - z_1)z_1 - \left(x - \frac{z_2}{2}\right) z_2 - \sum_{j=3}^{7}(2x - z_j)z_j - \left(x - \frac{z_8}{2}\right) z_8 - (x - z_9)z_9 =$$

$$= -(y_1 - y_2 - z_4 + z_5)^2 - (y_2 - y_3)^2 - (y_3 - y_4 + z_5 - z_6)^2 -$$
$$- 2\left(z_1 - \frac{z_2}{2}\right)^2 - (z_2 - z_3)^2 - (z_3 - z_4)^2 - (z_6 - z_7)^2 - (z_7 - z_8)^2 - 2\left(\frac{z_8}{2} - z_9\right)^2 -$$
$$- \sum_{i=2}^{3}(2x' - y_i)y_i -$$
$$- (x - z_1)z_1 - \left(x - \frac{z_2}{2}\right)z_2 - \sum_{j=3}^{7}(2x - z_j)z_j - \left(x - \frac{z_8}{2}\right)z_8 - (x - z_9)z_9.$$

Clearly, $f \leq 0$. We are going to find all the cases when $f = 0$. Observe that, for any $i \in \{2, 3\}, j \in [9]$, there is a summand $-y_i(2x' - y_i)$ and a corresponding one for z_j. Thus, we need to consider only the cases when $y_2, y_3 \in \{0, 2x'\}$, $z_1, z_9 \in \{0, x\}$ and $z_2, \ldots, z_8 \in \{0, 2x\}$. Suppose that $y_1 \notin \{0, 2x'\}$; then $0 < |y_1 - y_2| < 2x'$ and thus $y_1 - y_2 - z_2 + z_3 \neq 0$, for any $z_2, z_3 \in \{0, 2x\}$, as $x' < x$. So $y_1 = y_2$, and thus $z_2 = z_3$. Similarly, $y_3 = y_4$, and so $z_5 = z_6$. Using other clauses of the expression, we conclude that $y_1 = \ldots = y_4$, $z_2 = \ldots = z_8$, and also we have $z_1 = z_9 = z_2/2 = z_8/2$. This means that the colorings of the blocks alternate.

Now we need to prove that $\mathrm{Color}(L_2) \neq \mathrm{Color}(R_1)$. Suppose the opposite; then the cut between the blocks and the long intervals will be at most the sum: $4x$ provided by L_2 plus $4x$ provided by R_1, plus x provided by L_1 plus x provided by R_2 and plus $2x'$ provided by one of L_2, R_1. In total it makes $10x + 2x'$. On the other hand, if $\mathrm{Color}(L_2) \neq \mathrm{Color}(R_1)$, then together L_2 and R_1 provide $7x + 4x'$, for each of two possible colorings of the bottom blocks, and L_1 and R_2 provide $2x$, in total it is $9x + 4x'$. This case is reachable, because we can choose the leftmost bottom block to have a color opposite to $\mathrm{Color}(L_1)$. As $\mathrm{Color}(R_2)$ is not fixed, we can choose the right one that will add x as well. So we have to satisfy the inequality

$$10x + 2x' < 9x + 4x'.$$

We assumed that $2x' > x$ so the inequality holds and thus the second case is optimal. We should note that, in the second case, L_1 and R_2 have the same color because the leftmost and the rightmost bottom blocks are of the same color. We have shown that L_1 and R_2 are colored similarly and that L_2 and R_1 are colored differently. □

Lemma A.6.6. *Consider some switch gadget of H. For any maximum cut partition of H, the colorings of both levels will be alternating except for at most η intervals within $\mathcal{B}_1^{top}$ or $\mathcal{B}_4^{top}$.*

Proof. Let the number of red and blue long intervals overlapping the switch gadget be r and b respectively, and denote the quantity $(r - b)$ by Δ. Suppose that $r > b$. Let y_i, z_j,

for $i \in [4], j \in [9]$, be the same as in Lemma A.6.5. Compute the size of the cut:

$$\begin{aligned}
\textsc{Cut} = 12x'^2 + 28x^2 + 16x'x - \\
- (y_1 - y_2 - z_4 + z_5)^2 - (y_3 - y_4 + z_5 - z_6)^2 - \\
- 2\left(z_1 - \frac{z_2}{2}\right)^2 - (z_2 - z_3)^2 - (z_3 - z_4)^2 - (z_6 - z_7)^2 - (z_7 - z_8)^2 - 2\left(\frac{z_8}{2} - z_9\right)^2 - \\
- (y_2 - y_3)^2 - \sum_{i=2}^{3}(2x' - y_i)y_i - \\
- (x - z_1)z_1 - \left(x - \frac{z_2}{2}\right)z_2 - \sum_{j=3}^{7}(2x - z_j)z_j - \left(x - \frac{z_8}{2}\right)z_8 - (x - z_9)z_9 + \\
- \Delta\left(\sum_{i=1}^{4}(-1)^i y_i + \sum_{j=1}^{9}(-1)^j z_j\right) + (8x + 4x')(r + b) = \\
\\
= 12x'^2 + 28x^2 + 16x'x - \\
- (y_1 - y_2 - z_4 + z_5 - \Delta/4)^2 - (y_3 - y_4 + z_5 - z_6 - \Delta/4)^2 - \\
- 2\left(z_1 - \frac{z_2}{2} - \Delta/4\right)^2 - (z_2 - z_3 + \Delta/4)^2 - (z_3 - z_4 - \Delta/4)^2 - \\
- (z_6 - z_7 + \Delta/4)^2 - (z_7 - z_8 - \Delta/4)^2 - 2\left(\frac{z_8}{2} - z_9 + \Delta/4\right)^2 - \\
- (y_2 - y_3 + \Delta/4)^2 - \sum_{i=2}^{3}(2x' - y_i)y_i - \\
- (x - z_1)z_1 - \left(x - \frac{z_2}{2}\right)z_2 - \sum_{j=3}^{7}(2x - z_j)z_j - \left(x - \frac{z_8}{2}\right)z_8 - (x - z_9)z_9 + \\
+ \frac{\Delta}{2}(y_1 - y_4) + (8x + 4x')(r + b) + \frac{11}{16}\Delta^2.
\end{aligned}$$

Consider only those summands that participate in the size of the cut when the coloring of the blocks alternates, i.e. when, for all i, j, $y_i = z_j = 0$:

$$\textsc{Cut}_{alter} = 12x'^2 + 28x^2 + 16x'x + (8x + 4x')(r + b).$$

If we choose $x > \frac{(2\Delta+1)^2}{2} + \frac{11}{4}\Delta^2$, then the distance between some variable (except for y_1, y_4) and the closest end of its domain cannot be greater than Δ. For example, suppose that $\min(z_1, x - z_1) > \Delta$. Then $(x - z_1)z_1 > \Delta x + \frac{11}{16}\Delta^2$, and so

$$\textsc{Cut} \leq \textsc{Cut}_{alter} - (x - z_1)z_1 + \frac{\Delta}{2}(y_1 - y_4) + \frac{11}{16}\Delta^2 < \textsc{Cut}_{alter}.$$

So, in this case it will not be a maximum cut.

Suppose that $|z_j - z_{j+1}| > \Delta$, then either $|z_j - z_{j+1}| > 2x - 2\Delta$ when $j \notin \{1, 4, 5, 8\}$, or $|z_j - z_{j+1}| > x - 2\Delta$ when $j \in \{1, 8\}$. But then one of the clauses will be greater than $(x - 2\Delta - \Delta/4)^2$, hence, greater than $\Delta x + \frac{11}{16}\Delta^2$. Similarly, $|y_2 - y_3| < \Delta$. Consider the variables z_4, z_5, z_6. Choose x' between $x/2$ and x such that $(2x' - 2x + 2\Delta + \Delta/4)^2$ is

greater than $\Delta x + \frac{11}{16}\Delta^2$, as $x > \Delta^2$, it is easy to choose a convenient value for x'. For such x', we will have $|z_4 - z_5|, |z_5 - z_6| < \Delta$.

Denote $d_{12} = y_1 - y_2$ and $d_{34} = y_3 - y_4$. Then $y_1 - y_4 < |d_{12}| + \Delta + |d_{34}|$. Then we can choose only those d_{12}, d_{34} that satisfy

$$-(d_{12} - z_4 + z_5 - \Delta/4)^2 - (d_{34} + z_5 - z_6 - \Delta/4)^2 + \frac{\Delta}{2}(|d_{12}| + \Delta + |d_{34}|) + \frac{11}{16}\Delta^2 > 0$$

as, otherwise, it will not be a maximal cut. Denote $w = -z_4 + z_5 - \Delta/4, w' = z_5 - z_6 - \Delta_4$, we know that $|w|, |w'| \leq 5\Delta/4$. Then the expression can be written in the following form:

$$-d_{12}^2 + 2d_{12}w - w^2 - d_{34}^2 + 2d_{34}w' - w'^2 + \Delta/2(|d_{12}| + |d_{34}|) + \frac{19}{16}\Delta^2 > 0 \Rightarrow$$

$$\Rightarrow -(d_{12} - w - \Delta/4)^2 - (d_{34} - w' - \Delta/4)^2 + \frac{\Delta}{2}(|w| + |w'|) + \frac{2\Delta^2}{16} + \frac{19}{16}\Delta^2 > 0.$$

Take $|d_{12}| \geq 4\Delta$. Then

$$-(d_{12} - w - \Delta/4)^2 - (d_{34} - w' - \Delta/4)^2 + \frac{\Delta}{2}(|w| + |w'|) + \frac{2\Delta^2}{16} + \frac{19}{16}\Delta^2 \leq$$

$$\leq -\left(\frac{5\Delta}{2}\right)^2 + \frac{5\Delta^2}{4} + \frac{21\Delta^2}{16} < 0.$$

The same is true for $|d_{34}|$. We now can imply that $|y_1 - y_4| < 9\Delta$, otherwise the cut is not maximal. This is an important result because we can show now that all the blocks except for y_1, y_4 are monochromatic. Take any $x > \frac{9}{2}\Delta^2 + \frac{11}{16}\Delta^2 = \frac{83}{16}\Delta^2$, then suppose that for some variable except for y_1, y_4, say z_1 for example: $z_1 \notin \{0, x\}$. Then $(x - z_1)z_1 \geq x - 1 > \frac{9}{2}\Delta^2 + \frac{11}{16}\Delta^2$. So the cut will not be maximal in this case. Same is true for any variable other than y_1, y_4. So we can conclude that $2z_1 = z_2 = \ldots = z_8 = 2z_9 \in \{0, 2x\}$, and $y_2 = y_3 \in \{0, 2x'\}$.

Suppose without loss of generality that all the variables except for y_1, y_4 are equal to 0. Then $\textsc{Cut} - \textsc{Cut}_{alter}$ equals

$$-(y_1 - \Delta/4)^2 - (y_4 + \Delta/4)^2 + \frac{\Delta}{2}(y_1 - y_4) + \frac{\Delta^2}{8} = -y_1(y_1 - \Delta) - y_4(y_4 + \Delta).$$

One can see now that the cut will be optimal if $y_1 = \Delta/2, y_4 = 0$ for $\Delta > 0$. For the case when $\Delta < 0$ the optimal cut will be when $y_1 = 0$ and $y_4 = -\Delta/2$. Recall that η denotes the total number of long intervals. As $\Delta = (r - b)$, we have $|\Delta/2| < \eta$. □

A.7 Reduction

After describing each gadget of H we can go one level higher and explain how they are positioned in the graph. We first consider an operation that is used many times in H, the switch procedure. It allows to change the relative positions of two long interval chains. If we need to connect two chains to the same edge gadget and if there are other chains between these two, then this procedure becomes extremely helpful. Further, we provide the explicit construction of H and, at the end, prove Theorem A.1.1.

Switch procedure Consider some adjacent vertices g_i, g_j of the cubic graph G, they correspond to two vertex gadgets $\mathcal{V}_i, \mathcal{V}_j$ and to an edge gadget $\mathcal{E}_{ij}$ in the interval model of H. It might be that there is another vertex gadget $\mathcal{V}'$ between $\mathcal{V}_i$ and $\mathcal{V}_j$. Consequently, there is a join gadget corresponding to $\mathcal{V}'$ between any pair of join gadgets of $\mathcal{V}_i, \mathcal{V}_j$. We do not want any gadget to intersect $\mathcal{E}_{ij}$, so we have to switch some long interval chain starting from $\mathcal{V}_i$ with all three long interval chains that start from $\mathcal{V}'$, and similarly for any other vertex gadget between $\mathcal{V}_i$ and $\mathcal{V}_j$.

We solve this issue by the *switch procedure*. It is displayed on Figure A.9. The line $\mathbb{R}$ is split into zones and buffers between zones. Each zone (or buffer) is a disjoint union of fragments of the same size, where the distance between two neighbor fragments is the same and depends on the long interval length α. Every zone corresponds to some vertex gadget, it contains the vertex gadget and most of the join gadgets of link chains starting from it. Every fragment of Buffer$(i, i+1)$ is placed between the fragments of Zone(i) and Zone$(i+1)$, for any pair of neighbor vertex gadgets $\mathcal{V}_i$ and $\mathcal{V}_{i+1}$. Buffer$(i, i+1)$ contains switch gadgets that are used to let a chain starting from $\mathcal{V}_i$ pass all three chains starting from $\mathcal{V}_{i+1}$, this is exactly what is shown on Figure A.9. By repeatedly iterating the switching, we pass all the vertex gadgets between $\mathcal{V}_i$ and $\mathcal{V}_j$ and eventually connect the corresponding long interval chains to a common edge gadget $\mathcal{E}_{ij}$. In order to justify the correctness of Figure A.9, we describe precise positions of each gadget and long interval within each zone during this procedure.

Suppose that $\alpha = 53n - 3$. All the intervals are going to be inside $[0, +\infty)$. This ray is divided into zones.

- Zones that correspond to vertices of G. For a vertex $g_i \in V(G)$, denote the corresponding zone by Zone(i) and define it as

$$\text{Zone}(i) = \bigcup_{j \in \mathbb{Z}} \big[53i + j(\alpha+3), 53i + j(\alpha+3) + 32\big].$$

  This zone usually contains the vertex and the join gadgets corresponding to $g_i \in V(G)$. The size of its fragments is 32. The distance between the start of the ith interval and the start of the $(i+1)$th interval is called the *phase*.

- Buffer zones. For two vertices $g_i, g_{i+1} \in V(G)$ (and also g_{n-1}, g_0), we introduce a buffer zone between Zone(i), Zone$(i+1)$. It is denoted by Buffer$(i, i+1)$ (Buffer$(n-1, 0)$ is denoted similarly) and is defined as follows:

$$\text{Buffer}(i, i+1) = \bigcup_{j \in \mathbb{Z}} \big[53i + j(\alpha+3) + 32, 53(i+1) + j(\alpha+3)\big].$$

  We need buffer zones in order to do the switching. Every fragment o f a buffer zone has size 21.

We need to write $\alpha + 3$ instead of α because a long interval must start from the rightmost block of some 3-block and must terminate at the leftmost block of another 3-block. So, the length of long intervals must be lesser than the length of the phase by 3.

Call by a point x of Buffer$(i, i+1)$ a set of points $\{53i + (j+3)\alpha + 32 + x \mid j \in \mathbb{Z}\}$. As the minimal distance between two points of this set is exactly α and as we always attach just one long interval, it can be uniquely understood which point of the set is considered at the moment.

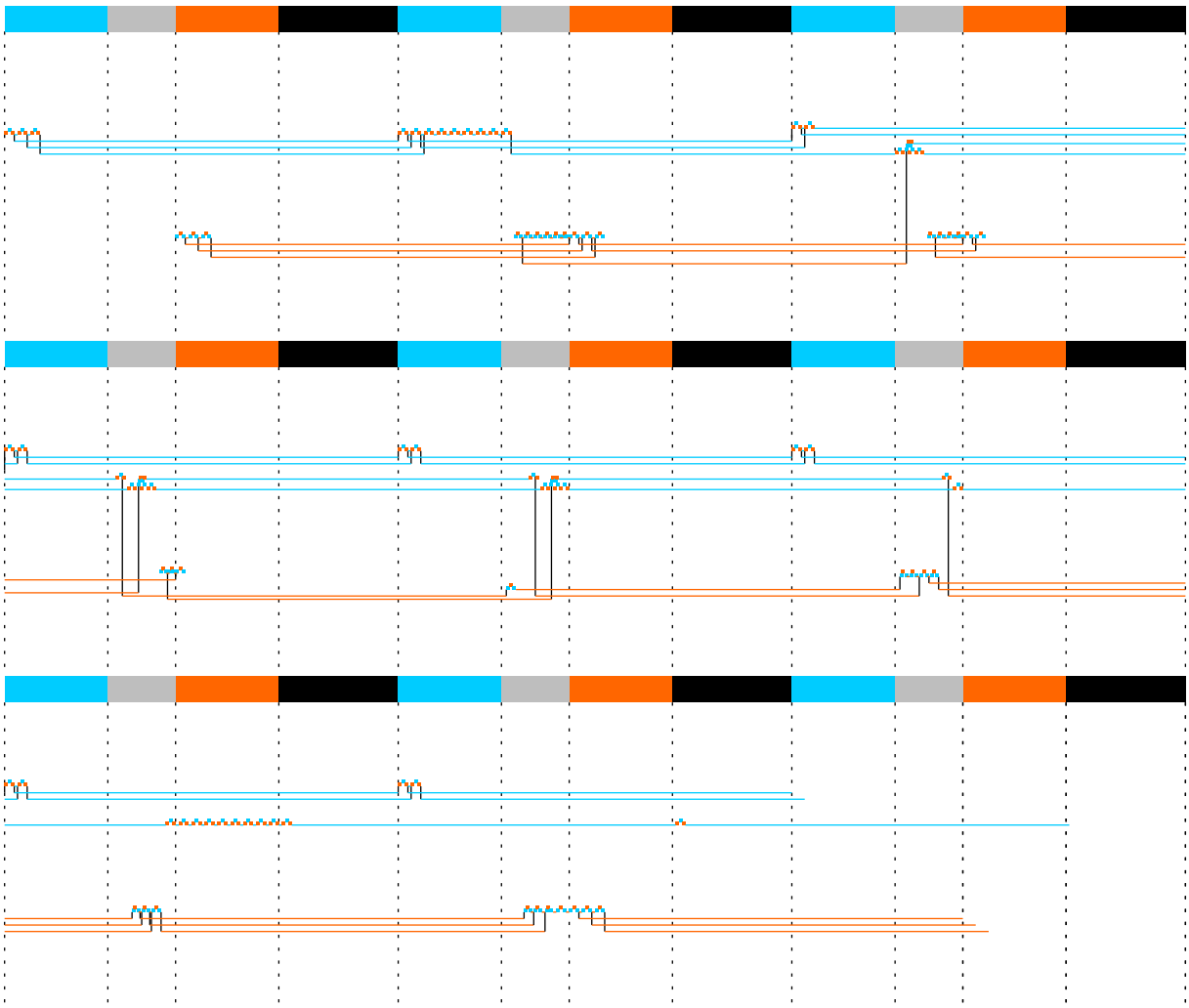


Figure A.9: The switch procedure. Blue and red fragments correspond to Zone(i) and Zone($i+1$). Grey fragment is Buffer($i, i+1$) between these zones. Black fragment contains all other zones and buffers. One should read this figure from left to right and from top to bottom, like a book.

Say that a gadget is in the fragment $[a, b]$ of some zone if its leftmost and rightmost points are in a and b. We are going to show that no gadgets intersect each other by considering each of nine buffer zones on Figure A.9.

1. The first buffer does not contain any gadgets.

2. The second buffer intersects two stretch gadgets. The blue one is in $[0, 32]$ of Zone(i) and in $[0, 3]$ of Buffer($i, i+1$). Its long intervals terminate at $\{0, 4, 8\}$ of Zone(i) and start from $\{3, 7\}$ of Zone(i) and from 3 of Buffer($i, i+1$). The red one is in $[4, 21]$ of Buffer($i, i+1$) and in $[0, 11]$ of Zone($i+1$). Its long intervals terminate at $\{0, 4, 8\}$ of Zone($i+1$) and start from 6.5 of Buffer($i, i+1$) and from $\{3, 7\}$ of Zone($i+1$).

3. The third buffer contains the first switch gadget in $[0, 9]$. Its long intervals terminate at $\{0, 3.5\}$ and start from $\{5.5, 9\}$ of Buffer($i, i+1$). It also contains the red stretch gadget in $[10, 21]$ of Buffer($i, i+1$) and in $[0, 7]$ of Zone($i+1$). Its long intervals terminate at $\{0, 4\}$ of Zone($i+1$) and start from 12.5 of Buffer($i, i+1$) and from 3 of Zone($i+1$).

4. The fourth buffer contains the second switch gadget in $[6, 15]$. Its long intervals terminate at $\{6, 9.5\}$ and start from $\{11.5, 15\}$. It also contains the second part of the first switch gadget in $[2.5, 5.5]$. Its long intervals terminate at 2.5 and start from 4.5. It also contains a red stretch gadget in $[16, 21]$ of Buffer$(i, i+1)$ and in $[0, 3]$ of Zone$(i+1)$. Its long intervals terminate at 21 and start from 18.5 of Buffer$(i, i+1)$.

5. The fifth buffer contains the second part of the second switch gadget in $[8.5, 11.5]$. Its long intervals terminate at 8.5 and start from 10.5. It also contains the third switch gadget in $[12, 21]$. Its long intervals terminate at $\{12, 15.5\}$ and start from $\{17.5, 21\}$. It also contains a red join gadget in $[1.5, 4.5]$. Its long intervals terminate at 1.5 and start from 4.5.

6. The sixth buffer contains a blue join gadget in $[18, 21]$. Its long intervals terminate at 18 and start from 21. It also contains the second part of the third switch gadget in $[14.5, 17.5]$. Its long intervals terminate at 14.5 and start from 16.5. It also contains a red stretch gadget in $[1.5, 13.5]$. Its long intervals terminate at $1.5, 7.5$ and start from $10.5, 13.5$.

7. The seventh buffer intersects a blue stretch gadget that is in $[18, 21]$ of Buffer$(i, i+1)$, in $[0, 32]$ of Zone$(i+1)$ and in $[0, 4]$ of Buffer$(i+1, i+2)$. Its long intervals terminate at 18 of Buffer$(i, i+1)$ and start from 4 of Buffer$(i+1, i+2)$. It also contains a red stretch gadget in $[7.5, 16.5]$. Its long intervals terminate at $\{7.5, 10.5, 13.5\}$ and start from $\{10, 13, 16.5\}$.

8. The eighth buffer intersects a red stretch gadget. It is in $[7, 21]$ of Buffer$(i, i+1)$ and in $[0, 11]$ of Zone$(i+1)$. Its long intervals terminate at $\{7, 10, 13.5\}$ of Buffer$(i, i+1)$ and start from $\{3, 7, 11\}$ of Zone$(i+1)$.

9. The ninth buffer is empty.

Observe that the distance between any two ends of long intervals is always at least 2.5 if one end is left and the other is right, and is at least 3. Lemma A.6.3 allows us to do this. This is the end of the switching procedure.

Construction of H Let G be a cubic graph of size n. Let the long interval length to be $\alpha := 53n - 3$. All the intervals of H are going to be inside $[0, +\infty)$. This ray is split into zones and buffers in the following order:

$$\text{Zone}(1), \text{Buffer}(1, 2), \text{Zone}(2), \text{Buffer}(2, 3), \ldots, \text{Zone}(n), \text{Buffer}(n, 1), \text{Zone}(1), \ldots$$

For $i \in [n]$, put a vertex gadget $\mathcal{V}_i$ into the leftmost fragment of Zone(i). For each $\mathcal{V}_i$ there are 3 long intervals starting from it; they terminate at a join gadget that has 3 new long intervals starting from it. This produces 3 chains of long intervals. Each such chain eventually terminates at some edge gadget.

For every edge $g_i g_j \in E(G)$, where $i < j$, we choose a chain starting from $\mathcal{V}_i$ and repeatedly apply the switch procedure to this chain until it is in Zone(j). Once it happens, this chain of $\mathcal{V}_i$ and one of the chains of $\mathcal{V}_j$ terminate at the same edge gadget $\mathcal{E}_{ij}$. No long interval starts from $\mathcal{E}_{ij}$. Then we choose another edge of G and repeat this operation until all edges of G are treated. If a chain does not participate in some switch procedure, then, during this procedure, the long intervals of this chain are joined by join gadgets that look the same as vertex gadgets. The composition of H is displayed on Figure A.10.

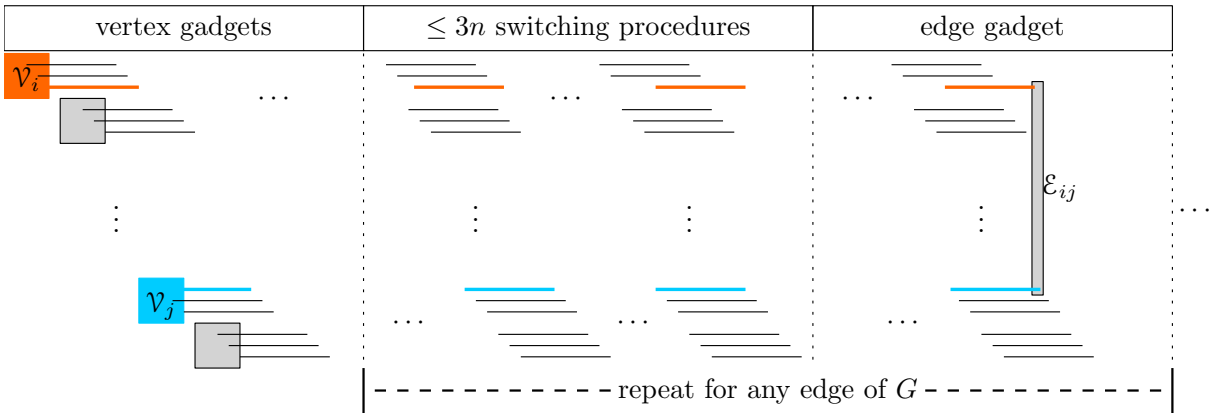


Figure A.10: The composition of H.

Reduction At first, we return to the three conditions about H that we stated in Appendix A.3 and that we assumed to hold. We need to verify that the graph H that we have just constructed indeed satisfies them. Clearly, we are free to choose x, k to be in $\Omega(n^6)$. In total, there are $3n$ long intervals starting from vertex gadgets, so every block is intersected by at most $3n$ long intervals. Every long interval overlaps $n-1$ zone and n buffers, each zone contains a join gadget that has at most 10 3-blocks, each buffer contains at most one switch gadget that has 16 blocks. Thus, those 3 conditions hold for H.

Let us say that a partition of $V(H)$ in 2 classes $\{R, B\}$ is *good* if the following conditions hold.

1. For any gadget, its coloring is alternating or almost alternating except for $\mathcal{B}_2^i$ of a 3-block $\mathcal{B}^i$ or one of $\mathcal{B}_1^{top}, \mathcal{B}_4^{top}$ of a switch gadget that contain at most η intervals of the other color.

2. For any long interval L that starts from or terminates at a block $\mathcal{B}$ of size x', x, or $2x$, we have $\text{Color}(L) \neq \text{Color}(\mathcal{B})$.

Lemma A.7.1. *Any* MaxCut *partition of H is good.*

Proof. The first condition is provided by Lemma A.6.1, Lemma A.6.2, Lemma A.6.4, and Lemma A.6.6.

Assume that the second condition does not hold for some long interval L, suppose that it starts from a block $\mathcal{B}$ of size at least $x' > \frac{x}{2}$ and $\text{Color}(L) = \text{Color}(\mathcal{B})$. L is a part of some long interval chain that connects a vertex gadget $\mathcal{V}$ to an edge gadget $\mathcal{E}$. Invert $\text{Color}(L)$ and modify the colors of all gadgets and long intervals of this chain that are between L and $\mathcal{E}$ so that this particular chain satisfies the second condition. The gain after this operation is at least x because $\text{Color}(L) \neq \text{Color}(\mathcal{B})$. Denote by l the number of long intervals in the chain, it is at most $3n \times 9 \times \frac{3n}{2}$. The loss is at most the sum of the following values:

- k, if both intervals terminating at $\mathcal{E}$ have the same color;

- $\eta\mu l$ – the cut between the gadgets of H overlapped by the intervals of the chain, where η is an upper bound for the number of long intervals overlapping a gadget and μ is an upper bound on blocks that a long interval intersects;

- $2\eta l$ – the cut between the long intervals of H intersected by the intervals of the chain;
- $\eta^2 l$ – the cut between the gadgets of the chain and the long intervals of H overlapping them (the number of gadgets in the chain also equals l).

As $\eta, \mu \in O(n)$, and $l \in O(n^2)$, the loss is in $O(k + n^4)$. We are free to choose x to be large enough for the gain to be strictly greater than the maximal possible loss. Thus, for such value of x, the second condition also holds. □

A good partition $p\colon V(H) \to \{R, B\}$ corresponds to some partition $q\colon V(G) \to \{R, B\}$: assign to a vertex $g_i \in V(G)$ the same color that is assigned to the chains starting from the vertex gadget $\mathcal{V}_i$. For such p and q, we will write $p \sim q$. Clearly, the other direction is also true: for every $q\colon V(G) \to \{R, B\}$ there exists a good partition p such that $p \sim q$.

Lemma A.7.2. *Let $p\colon V(H) \to \{R, B\}$ be a* MaxCut *partition of H, and $q\colon V(G) \to \{R, B\}$ be the corresponding partition of G, i.e., $p \sim q$. Then q is a* MaxCut *partition of G.*

Proof. Suppose that q is not a MaxCut partition, that is, there is another $q'\colon V(G) \to \{R, B\}$ which is maximal. Let $p'\colon V(H) \to \{R, B\}$ be a good partition that satisfies $p' \sim q'$.

For simplicity, we denote by E the set of intervals that belong to edge gadgets, and C denotes the set of intervals that belong to chains or to other gadgets. Clearly, $V(H) = C \sqcup E$.

The difference between the cut values of p and p', for edges induced by E, is at most $\eta^3 \in O(n^3)$ because there are $\frac{\eta}{2}$ edge gadgets, and each of them changes the cut by at most $2\eta^2$. The difference between the cut values, for edges induced by C, is at most

$$\underbrace{\eta^2 \cdot 10 \cdot l \cdot \eta}_{\text{within each gadget}} + \underbrace{\eta \cdot l \cdot \eta}_{\text{between long intervals}} + \underbrace{2\eta \cdot \mu \cdot l \cdot \eta}_{\text{between gadgets and long intervals}}$$

which belongs to $O(n^5)$. Finally, as q' has a strictly greater cut value that q, we know that the number of cut edges between C and E for p' is greater than the corresponding number for p by at least $(k - 2\eta) - \eta \cdot 2\eta \cdot \frac{\eta}{2}$, by Lemma A.6.2 and because an edge gadget is overlapped by at most η long intervals, each of them adds at most 2η to the cut, and there are $\frac{\eta}{2}$ edge gadgets. As we choose k to be in $\Omega(n^6)$, the cut for p' is greater than the cut for p, it is a contradiction. □

Lemma A.7.2 implies Theorem A.1.1.

Bibliography

[ABMR21] Ranendu Adhikary, Kaustav Bose, Satwik Mukherjee, and Bodhayan Roy. Complexity of maximum cut on interval graphs. In *37th International Symposium on Computational Geometry, SoCG 2021, June 7-11, 2021, Buffalo, NY, USA (Virtual Conference)*, volume 189 of *LIPIcs*, pages 7:1–7:11. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

[Alo86] Noga Alon. Eigenvalues, geometric expanders, sorting in rounds, and ramsey theory. *Comb.*, 6(3):207–219, 1986.

[Ats08] Albert Atserias. On digraph coloring problems and treewidth duality. *Eur. J. Comb.*, 29(4):796–820, 2008.

[Awo10] Steve Awodey. *Category Theory*. Oxford University Press, Inc., 2nd edition, 2010.

[Bar83] Francisco Barahona. The max-cut problem on graphs not contractible to k5. *Operations Research Letters*, 2(3):107–111, 1983.

[BBKO21] Libor Barto, Jakub Bulín, Andrei A. Krokhin, and Jakub Oprsal. Algebraic approach to promise constraint satisfaction. *J. ACM*, 68(4):28:1–28:66, 2021.

[BD13] Manuel Bodirsky and Víctor Dalmau. Datalog and constraint satisfaction with infinite templates. *J. Comput. Syst. Sci.*, 79(1):79–100, 2013.

[BdFG+04] Hans L. Bodlaender, Celina M. H. de Figueiredo, Marisa Gutierrez, Ton Kloks, and Rolf Niedermeier. Simple max-cut for split-indifference graphs and graphs with few p_4's. In *Experimental and Efficient Algorithms, Third International Workshop, WEA 2004, Angra dos Reis, Brazil, May 25-28, 2004, Proceedings*, volume 3059 of *Lecture Notes in Computer Science*, pages 87–99. Springer, 2004.

[BDJN15] Jakub Bulin, Dejan Delic, Marcel Jackson, and Todd Niven. A finer reduction of constraint problems to digraphs. *Log. Methods Comput. Sci.*, 11(4), 2015.

[BES17] Arman Boyaci, Tínaz Ekim, and Mordechai Shalom. A polynomial-time algorithm for the maximum cardinality cut problem in proper interval graphs. *Inf. Process. Lett.*, 121:29–33, 2017.

[BFJ+22] Jan Bok, Jirí Fiala, Nikola Jedlicková, Jan Kratochvíl, and Pawel Rzazewski. List covering of regular multigraphs. *CoRR*, abs/2204.04280, 2022.

[Big74] Norman Biggs. *Algebraic graph theory*. Cambridge Tracts in Mathematics. Cambridge University Press, 2 edition, 1974.

[BJ94] Hans L. Bodlaender and Klaus Jansen. On the complexity of the maximum cut problem. In *STACS 94, 11th Annual Symposium on Theoretical Aspects of Computer Science, Caen, France, February 24-26, 1994, Proceedings*, volume 775 of *Lecture Notes in Computer Science*, pages 769–780. Springer, 1994.

[BJK05] Andrei A. Bulatov, Peter Jeavons, and Andrei A. Krokhin. Classifying the complexity of constraints using finite algebras. *SIAM J. Comput.*, 34(3):720–742, 2005.

[BK99] Piotr Berman and Marek Karpinski. On some tighter inapproximability results (extended abstract). In *Automata, Languages and Programming, 26th International Colloquium, ICALP'99, Prague, Czech Republic, July 11-15, 1999, Proceedings*, volume 1644 of *Lecture Notes in Computer Science*, pages 200–209. Springer, 1999.

[BK17] Libor Barto and Marcin Kozik. Absorption in universal algebra and CSP. In *The Constraint Satisfaction Problem: Complexity and Approximability*, volume 7 of *Dagstuhl Follow-Ups*, pages 45–77. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

[BKM12] Manuel Bodirsky, Jan Kára, and Barnaby Martin. The complexity of surjective homomorphism problems - a survey. *Discret. Appl. Math.*, 160(12):1680–1690, 2012.

[BKN99] Hans L. Bodlaender, Ton Kloks, and Rolf Niedermeier. SIMPLE MAX-CUT for unit interval graphs and graphs with few p_4s. *Electron. Notes Discret. Math.*, 3:19–26, 1999.

[BKO+17] Libor Barto, Michael Kompatscher, Miroslav Olsák, Trung Van Pham, and Michael Pinsker. The equivalence of two dichotomy conjectures for infinite domain constraint satisfaction problems. In *32nd Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2017, Reykjavik, Iceland, June 20-23, 2017*, pages 1–12. IEEE Computer Society, 2017.

[BKS20] Manuel Bodirsky, Simon Knäuer, and Florian Starke. ASNP: A tame fragment of existential second-order logic. In *Beyond the Horizon of Computability - 16th Conference on Computability in Europe, CiE 2020, Fisciano, Italy, June 29 - July 3, 2020, Proceedings*, volume 12098 of *Lecture Notes in Computer Science*, pages 149–162. Springer, 2020.

[BMM18] Manuel Bodirsky, Florent R. Madelaine, and Antoine Mottet. A universal-algebraic proof of the complexity dichotomy for monotone

monadic SNP. In *Proceedings of the 33rd Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2018, Oxford, UK, July 09-12, 2018*, pages 105–114. ACM, 2018.

[BNP10] Richard N. Ball, Jaroslav Nešetřil, and Aleš Pultr. Dualities in full homomorphisms. *Eur. J. Comb.*, 31(1):106–119, 2010.

[Bod21] Manuel Bodirsky. *Complexity of Infinite-Domain Constraint Satisfaction.* Lecture Notes in Logic. Cambridge University Press, 2021.

[BtCLW14] Meghyn Bienvenu, Balder ten Cate, Carsten Lutz, and Frank Wolter. Ontology-based data access: A study through disjunctive datalog, csp, and MMSNP. *ACM Trans. Database Syst.*, 39(4):33:1–33:44, 2014.

[Bul03] Andrei A. Bulatov. Tractable conservative constraint satisfaction problems. In *18th IEEE Symposium on Logic in Computer Science (LICS 2003), 22-25 June 2003, Ottawa, Canada, Proceedings*, page 321. IEEE Computer Society, 2003.

[Bul06] Andrei A. Bulatov. A dichotomy theorem for constraint satisfaction problems on a 3-element set. *J. ACM*, 53(1):66–120, 2006.

[Bul17] Andrei A. Bulatov. A dichotomy theorem for nonuniform csps. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 319–330. IEEE Computer Society, 2017.

[CE12] Bruno Courcelle and Joost Engelfriet. *Graph Structure and Monadic Second-Order Logic - A Language-Theoretic Approach*, volume 138 of *Encyclopedia of mathematics and its applications.* Cambridge University Press, 2012.

[CSS99] Gregory Cherlin, Saharon Shelah, and Niandong Shi. Universal graphs with forbidden subgraphs and algebraic closure. *Advances in Applied Mathematics*, 22(4):454–491, 1999.

[dFdMdSOS21] Celina M. H. de Figueiredo, Alexsander Andrade de Melo, Fabiano de S. Oliveira, and Ana Silva. Maximum cut on interval graphs of interval count four is np-complete. In *46th International Symposium on Mathematical Foundations of Computer Science, MFCS 2021, August 23-27, 2021, Tallinn, Estonia*, volume 202 of *LIPIcs*, pages 38:1–38:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

[dFdMdSOS22] Celina M. H. de Figueiredo, Alexsander Andrade de Melo, Fabiano de S. Oliveira, and Ana Silva. Maxcut on permutation graphs is np-complete. *CoRR*, abs/2202.13955, 2022.

[DK07] Josep Díaz and Marcin Kaminski. MAX-CUT and MAX-BISECTION are np-hard on unit disk graphs. *Theor. Comput. Sci.*, 377(1-3):271–276, 2007.

[Erd59] P. Erdős. Graph theory and probability. *Canadian Journal of Mathematics*, 11:34–38, 1959.

[Fag74] Ronald Fagin. Generalized first-order spectra, and polynomial-time recognizable sets. *SIAM-AMS Proc.*, 7:43–73, 01 1974.

[FH08] Tomás Feder and Pavol Hell. On realizations of point determining graphs, and obstructions to full homomorphisms. *Discret. Math.*, 308(9):1639–1652, 2008.

[FHKM03] Tomás Feder, Pavol Hell, Sulamita Klein, and Rajeev Motwani. List partitions. *SIAM J. Discret. Math.*, 16(3):449–478, 2003.

[FHS14] Tomás Feder, Pavol Hell, and Oren Shklarsky. Matrix partitions of split graphs. *Discret. Appl. Math.*, 166:91–96, 2014.

[FHSS11] Tomás Feder, Pavol Hell, David G. Schell, and Juraj Stacho. Dichotomy for tree-structured trigraph list homomorphism problems. *Discret. Appl. Math.*, 159(12):1217–1224, 2011.

[FHX07] Tomás Feder, Pavol Hell, and Wing Xie. Matrix partitions with finitely many obstructions. *Electron. J. Comb.*, 14(1), 2007.

[FRW88] Peter Frankl, Vojtech Rödl, and Richard M. Wilson. The number of submatrices of a given type in a hadamard matrix and related results. *J. Comb. Theory, Ser. B*, 44(3):317–328, 1988.

[FV98] Tomás Feder and Moshe Y. Vardi. The computational structure of monotone monadic SNP and constraint satisfaction: A study through datalog and group theory. *SIAM J. Comput.*, 28(1):57–104, 1998.

[FV03] Tomás Feder and Moshe Y. Vardi. Homomorphism closed vs. existential positive. In *18th IEEE Symposium on Logic in Computer Science (LICS 2003), 22-25 June 2003, Ottawa, Canada, Proceedings*, pages 311–320. IEEE Computer Society, 2003.

[GJ90] Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., USA, 1990.

[Gro07] Martin Grohe. The complexity of homomorphism and constraint satisfaction problems seen from the other side. *J. ACM*, 54(1):1:1–1:24, 2007.

[Gur99] Venkatesan Guruswami. Maximum cut on line and total graphs. *Discret. Appl. Math.*, 92(2-3):217–221, 1999.

[Had75] F. Hadlock. Finding a maximum cut of a planar graph in polynomial time. *SIAM J. Comput.*, 4(3):221–225, 1975.

[Hel14] Pavol Hell. Graph partitions with prescribed patterns. *Eur. J. Comb.*, 35:335–353, 2014.

[HN90] Pavol Hell and Jaroslav Nešetřil. On the complexity of H-coloring. *J. Comb. Theory, Ser. B*, 48(1):92–110, 1990.

[HN04] Pavol Hell and Jaroslav Nešetřil. *Graphs and homomorphisms*, volume 28 of *Oxford lecture series in mathematics and its applications*. Oxford University Press, 2004.

[HN07] Pavol Hell and Jarik Nesetril. On the density of trigraph homomorphisms. *Graphs Comb.*, 23(Supplement-1):275–281, 2007.

[HN15] Jan Hubicka and Jaroslav Nešetřil. Universal structures with forbidden homomorphisms. In *Logic Without Borders - Essays on Set Theory, Model Theory, Philosophical Logic and Philosophy of Mathematics*, volume 5 of *Ontos Mathematical Logic*, pages 241–264. De Gruyter, 2015.

[Hod97] Wilfrid Hodges. *A Shorter Model Theory*. Cambridge University Press, USA, 1997.

[HR11] Pavol Hell and Arash Rafiey. The dichotomy of list homomorphisms for digraphs. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2011, San Francisco, California, USA, January 23-25, 2011*, pages 1703–1713. SIAM, 2011.

[Kar72] Richard M. Karp. Reducibility among combinatorial problems. In *Proceedings of a symposium on the Complexity of Computer Computations, held March 20-22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, USA*, The IBM Research Symposia Series, pages 85–103. Plenum Press, New York, 1972.

[KMN20] Jan Kratochvíl, Tomás Masarík, and Jana Novotná. U-bubble model for mixed unit interval graphs and its applications: The maxcut problem revisited. In *45th International Symposium on Mathematical Foundations of Computer Science, MFCS 2020, August 24-28, 2020, Prague, Czech Republic*, volume 170 of *LIPIcs*, pages 57:1–57:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

[Kun13] Gábor Kun. Constraints, MMSNP and expander relational structures. *Comb.*, 33(3):335–347, 2013.

[KZ13] Vladimir Kolmogorov and Stanislav Zivný. The complexity of conservative valued csps. *J. ACM*, 60(2):10:1–10:38, 2013.

[KZ17] Andrei A. Krokhin and Stanislav Zivný. The complexity of valued csps. In *The Constraint Satisfaction Problem: Complexity and Approximability*, volume 7 of *Dagstuhl Follow-Ups*, pages 233–266. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

[Lad75] Richard E. Ladner. On the structure of polynomial time reducibility. *J. ACM*, 22(1):155–171, 1975.

[Lib04] Leonid Libkin. *Elements of Finite Model Theory*. Texts in Theoretical Computer Science. An EATCS Series. Springer, 2004.

[Mad09] Florent R. Madelaine. Universal structures and the logic of forbidden patterns. *Log. Methods Comput. Sci.*, 5(2), 2009.

[Mad10] Florent R. Madelaine. On the containment of forbidden patterns problems. In *Principles and Practice of Constraint Programming - CP 2010 - 16th International Conference, CP 2010, St. Andrews, Scotland, UK, September 6-10, 2010. Proceedings*, volume 6308 of *Lecture Notes in Computer Science*, pages 345–359. Springer, 2010.

[Mar17] Barnaby Martin. Quantified constraints in twenty seventeen. In *The Constraint Satisfaction Problem: Complexity and Approximability*, volume 7 of *Dagstuhl Follow-Ups*, pages 327–346. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

[MS07] Florent R. Madelaine and Iain A. Stewart. Constraint satisfaction, logic and forbidden patterns. *SIAM J. Comput.*, 37(1):132–163, 2007.

[MS10] Gary MacGillivray and Jacobus Swarts. The complexity of locally injective homomorphisms. *Discret. Math.*, 310(20):2685–2696, 2010.

[Pap94] Christos H. Papadimitriou. *Computational complexity.* Addison-Wesley, 1994.

[PF79] Nicholas Pippenger and Michael J. Fischer. Relations among complexity measures. *J. ACM*, 26(2):361–381, 1979.

[PRS88] Pavel Pudlák, Vojtech Rödl, and Petr Savický. Graph complexity. *Acta Informatica*, 25(5):515–535, 1988.

[Sch78] Thomas J. Schaefer. The complexity of satisfiability problems. In *Proceedings of the 10th Annual ACM Symposium on Theory of Computing, May 1-3, 1978, San Diego, California, USA*, pages 216–226. ACM, 1978.

[Sta00] Richard P. Stanley. *Enumerative combinatorics*, volume Volume 1 of *Cambridge Studies in Advanced Mathematics*. Wadsworth and Brooks / Cole, Chapman and Hall, 1st edition, 2000.

[Wal23] J. L. Walsh. A closed set of normal orthogonal functions. *American Journal of Mathematics*, 45(1):5–24, January 1923.

[Zhu20] Dmitriy Zhuk. A proof of the CSP dichotomy conjecture. *J. ACM*, 67(5):30:1–30:78, 2020.

[ZM20] Dmitriy Zhuk and Barnaby Martin. QCSP monsters and the demise of the chen conjecture. In *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 91–104. ACM, 2020.