CLERMONT
UNIVERSITÉ

Mémoire d'habilitation

à diriger des recherches

Florent Madelaine

Jury :
Manuel Bodirsky (CNRS École Polytechnique)
Anuj Dawar (Cambridge, rapporteur)
Arnaud Durand (Paris 7, rapporteur)
Lhouari Nourine (Université Blaise Pascal, rapporteur)
Alain Quilliot (Université Blaise Pascal, coordinateur)
Luc Ségoufin (INRIA ENS-Cachan)

Remerciements

Table des matières

1. De la complexité des problèmes de contraintes

Décider si oui ou non l'instance d'un problème de satisfaction de contraintes a une solution est un problème difficile dans le cas général. La *conjecture de la dichotomie* postule qu'un problème de satisfaction de contraintes est soit facile (dans P), soit difficile (NP-complet). Cette conjecture, motivée par Feder et Vardi il y a 20 ans, reste ouverte malgré les efforts importants d'une communauté internationale regroupant des chercheurs issus d'horizons très variés. Nous présentons ici un bref aperçu des résultats obtenus et des techniques utilisées pour étudier cette question centrale en informatique théorique afin d'illustrer cette interdisciplinarité qui mêle algèbre, combinatoire, complexité et logique.

Il est bien connu que les problèmes de contraintes sont très généraux et permettent de modéliser naturellement de nombreux problèmes combinatoires et industriels difficiles. La généralité qu'offre ce cadre de travail et l'existence de méthodes génériques comme la propagation de contraintes est la motivation première de la communauté IA qui s'attache à développer des solveurs efficients tant au niveau de la modélisation que du calcul d'une solution. Ce qui est peut-être moins connu, et qui indique la robustesse des problèmes de contraintes, est leur ubiquité : ils sont étudiés sous d'autres noms en *bases de données* (inclusion de requêtes conjonctives), en *combinatoire et théorie des graphes* (problème d'existence d'homomorphisme, coloriage de graphes) et en *logique* (évaluation de formules primitives positives). La complexité de ce(s) problème(s), en particulier les cas pour lesquels il existe un algorithme polynomial[1] semblent s'expliquer soit de manière combinatoire, soit ce qui est peut-être plus surprenant par des *propriétés algébriques.*

La motivation ici n'est pas de présenter les détails techniques mais de citer certain résultats théoriques importants, d'en expliquer l'intuition et de les remettre dans leur contexte en espérant que le lecteur se tournera vers des *survey* récents en anglais (le livre [32] en regroupe plusieurs). Nous commencerons en §1.1 par des exemples et par deux théorèmes de dichotomie historiquement importants, celui de Schaefer qui concerne le cas Booléen et celui de Hell et Nešetřil qui concerne les graphes non-orientés. Ces théorèmes ont conduit Feder et Vardi à énoncer la conjecture de la dichotomie.

Au delà de l'intérêt évident d'une caractérisation des cas polynomiaux, cette conjecture est motivée par des résultats de *complexité structurelle* puisque la classe des problèmes de contraintes serait « la plus large » pour laquelle on observerait ce phénomène de dichotomie. On abordera cet aspect en §1.2.

Deux approches fondamentalement différentes permettent de restreindre le problème pour obtenir des classes polynomiales. Dans le premier cas, la notion de décomposition arborescente bornée du graphe des contraintes, bien connue depuis les travaux fondateurs de Freuder [47] est centrale et on verra en §1.3 qu'il existe un algorithme polynomial même lorsqu'il n'existe pas directement de telle décomposition. Dans le second cas, on restreint le langage de contraintes, et on verra en §1.4 que l'algèbre joue un rôle prépondérant et permet de caractériser par exemple le fait que pouvoir établir un certain niveau de cohérence suffit pour pouvoir déterminer si une solution existe. La *conjecture de la dichotomie* concerne en fait ce second cas et reste ouverte contrairement au premier cas qui est essentiellement complètement classifié. Finalement, nous nous tournerons en §1.5 vers des variantes, comme les problèmes de contraintes quantifiées, pour lesquelles la méthode algébrique peut être adaptée avec succès.

---

[1] pour le problème de décision : calculer, compter ou énumérer les solutions sont aussi des questions importantes mais on ne les abordera pas ici.

1.1 Préliminaires

Une instance du *problème de satisfaction de contraintes* est traditionnellement définie comme un triplet $(\mathsf{Var}, \mathsf{Dom}, \mathcal{C})$, où Var est un ensemble de variables, Dom est un ensemble fini de valeurs et $\mathcal{C}$ est un ensemble de contraintes. Chaque contrainte est de la forme $(v_{i_1}, \ldots, v_{i_r}, R)$, où r est l'arité de la contrainte et $R \subseteq \mathsf{Dom}^r$ une relation spécifiant tous les r-uplets de valeurs admissibles simultanément par les variables $v_{i_1}, \ldots, v_{i_r}$. Une *solution* de cette instance est une application qui associe une valeur à chaque variable de sorte que chaque contrainte soit satisfaite, à savoir qu'à $v_{i_1}, \ldots, v_{i_r}$ correspond un r-uplet admissible.

Exemple. Le problème de satisfiabilité propositionelle (Sat) peut facilement se coder comme un problème de contraintes. On parle alors de *Generalized Satisfiability* puisque les entrées des deux problèmes sont légèrement différentes. En effet, on aura un domaine booléen $\mathsf{Dom} = \{0, 1\}$, et à la clause $x \vee y \vee \bar{z}$ de Sat on fait correspondre la contrainte ternaire portant sur les variables x, y, z et de relation $\{0, 1\}^3 \setminus \{(0, 0, 1)\}$.

L'exemple précédent montre que le problème de contraintes est NP-complet. La généralité de ce cadre de travail n'est donc pas surprenante puisque si par *modéliser par des contraintes* on entend *réduction polynomiale au problème de satisfaction de contraintes*, tout problème Ω de NP est modélisable en ce sens.

Remarque. Mais est-ce-vraiment la bonne notion de modélisation ? Probablement pas puisque si le problème de départ Ω était facile on voudrait le réduire à une restriction Ω' du problème de contraintes qui soit aussi facile. Nous aborderons plus en détail cette aspect en §1.2.

On reformule souvent le problème de contraintes en tant que *problème d'homomorphisme* puisqu'on peut naturellement séparer une instance en deux structures similaires – la première $\mathcal{A}$ de domaine Var correspond essentiellement au graphe des contraintes, la seconde $\mathcal{B}$ de domaine Dom regroupe les relations listant les r-uplets admissibles – et qu'une solution correspond exactement à un homomorphisme . Un *homomorphisme* de graphe est une application de sommets à sommets qui doit nécessairement envoyer une arête sur une arête. Le cas des graphes orientés est similaire sauf qu'on préservera aussi l'orientation d'un arc. Nous allons expliciter cette reformulation sur un exemple simple ci-dessous. Un exemple plus complexe avec plusieurs contraintes et donc des structures avec plusieurs relations est donné pour le cas de 2-Sat page suivante.

Exemple. On peut facilement coder le problème de 3-colorabilité d'un graphe $\mathcal{G} := (V, E)$ comme problème de contraintes. Les sommets sont les variables, soit $\mathsf{Var} := V$, le domaine correspond aux trois couleurs $\mathsf{Dom} := \{1, 2, 3\}$ et puisque les sommets incidents à une arête doivent prendre une couleur différente, on poste la contrainte $(x_i, x_j, R_{\neq})$ pour chaque arête (x_i, x_j) dans E, où $R_{\neq}$ est la relation binaire $\{1, 2, 3\}^2 \setminus \{(1, 1), (2, 2), (3, 3)\}$ listant les paires de couleurs admissibles.

Sous forme de problème d'homomorphisme, dans notre cas simplifié $\mathcal{A}$ et $\mathcal{B}$ seront des graphes. Pour un graphe G, on aura $\mathcal{A} := \mathcal{G}$ et $\mathcal{B}$ qui sera un triangle $\mathcal{K}_3$
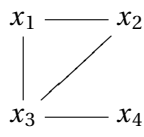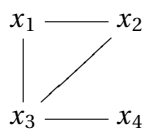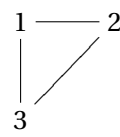
graphe	instance
x_1 —— x_2 x_3 —— x_4	$\mathsf{Var} = \{x_1, x_2, x_3, x_4\}$, $\mathsf{Dom} = \{1,2,3\}$ et $\mathcal{C} = \{((x_1,x_2), R_{\neq}), ((x_1,x_3), R_{\neq}), ((x_3,x_2), R_{\neq}), ((x_3,x_4), R_{\neq})\}$
$\mathcal{A}$ x_1 —— x_2 x_3 —— x_4	$\mathcal{B}$ 1 —— 2 3

FIGURE 1.1: Reformulation de la 3-colorabilité comme problème d'homomorphisme

(*cf.* figure 1.1). Ce problème est NP-complet alors que la 2-*colorabilité* est polynomiale.

En combinatoire, de nombreux problèmes difficiles deviennent faciles lorsque l'instance est un arbre. C'est le cas pour n'importe quel problème exprimable en logique monadique du second ordre, d'après le célèbre méta-théorème de Courcelle [28]. On verra en § 1.2 que notre problème s'exprime dans cette logique.

Théorème 1. *Si on se place dans le cas des graphes et qu'on restreint $\mathcal{A}$ comme étant un arbre, le problème d'homorphisme est polynomial.*

Remarque. On peut remplacer *graphe* par *structure* et *arbre* par *structure de largeur arborescente bornée*. Le résultat de Courcelle est très général et reste trop théorique avec des constantes impraticables (des tours d'exponentielle liées à la construction d'automates d'arbres). Par contre, pour le cas restreint qui nous intéresse, il existe un algorithme pratique dans l'esprit des travaux de Freuder [47] : on résoud localement en progressant du bas vers le haut dans l'arbre. Il s'agit d'un algorithme de type *bucket elimination* [39].

Le problème du $\mathcal{H}$-*coloring* généralise le problème de coloriage de graphe. Il s'agit de savoir étant donné un graphe $\mathcal{G}$ si il existe un homomorphisme de $\mathcal{G}$ dans $\mathcal{H}$ (le k-coloriage correspond au choix d'une k-clique pour $\mathcal{H}$)[2]. Si le graphe $\mathcal{H}$ est biparti, alors ce graphe $\mathcal{H}$ est 2-colorable. C'est-à-dire qu'il existe un homomorphisme de $\mathcal{H}$ vers $\mathcal{K}_2$ (le graphe qui consiste en une seule arête). Si le graphe $\mathcal{H}$ contient au moins une arête, alors il existe un homomorphisme qui envoie l'arête de $\mathcal{K}_2$ sur cette arête de $\mathcal{H}$. Ainsi $\mathcal{H}$ et K_2 sont dit *homomorphiquement équivalent* puisque tout graphe $\mathcal{G}$ qui admet un homomorphisme vers l'un en admet un vers

---

[2]Ce $\mathcal{H}$ correspond à notre $\mathcal{B}$ partout ailleurs, nous gardons le $\mathcal{H}$ ici pour préserver la nomenclature utilisée en combinatoire.

l'autre (car deux homomorphismes se composent naturellement pour donner un homomorphisme). Notez de plus que $\mathcal{K}_2$ est le plus petit sous-graphe de $\mathcal{H}$ qui lui est homomorphiquement équivalent. On dira dans ce cas que K_2 est le *core* de $\mathcal{H}$. Du point de vue du problème de décision, le $\mathcal{H}$-coloring et le 2-coloriage sont exactement le même problème, lorsque $\mathcal{H}$ est biparti et contient au moins une arête. Le problème du $\mathcal{H}$-coloring est donc polynomial si $\mathcal{H}$ est biparti. Une analyse combinatoire subtile permet de montrer qu'il s'agit là des seuls cas faciles. À partir des cas difficiles connus comme les cliques de taille 3 ou plus, la preuve procède par étapes successives afin d'augmenter cette classe progressivement. La preuve se termine puisqu'on impose tellement de conditions sur les graphes restants à classifier qu'on fini par montrer que cette classe est vide.

Théorème 2 (Hell et Nešetřil 1990 [53]). *Le $\mathcal{H}$-coloring est polynomial si $\mathcal{H}$ est biparti et NP-complet sinon.*

En général on considère des structures relationnelles quelconques et non pas juste des graphes. On rappelle qu'une *structure relationnelle $\mathcal{B}$* consiste en un ensemble B (toujours fini) et une liste finie de relations sur B. La notion d'homomorphisme entre deux structures similaires généralise naturellement le cas des graphes : il s'agit d'une fonction qui envoie chaque tuple de chaque relation de la première structure sur un tuple de la relation correspondante dans la seconde structure.

Exemples. On peu coder 2-Sat comme suit. Soit $R^{\mathcal{B}}_{ab} = \{0,1\}^2 \setminus \{(a,b)\}$ et $\mathcal{B} = (\{0,1\}; R^{\mathcal{B}}_{00}, R^{\mathcal{B}}_{01}, R^{\mathcal{B}}_{11})$. Une instance $F = (\neg x \vee \neg z) \wedge (x \vee y) \wedge (y \vee \neg z) \wedge (u \vee x) \wedge (x \vee \neg u)$ correspond à une structure $\mathcal{A}$ de domaine $\{x,y,z,u\}$ et de relations $R^{\mathcal{A}}_{00} = \{(x,y),(u,x)\}$, $R^{\mathcal{A}}_{01} = \{(y,z),(x,u),\}$ et $R^{\mathcal{A}}_{11} = \{(x,z)\}$. Notez la correspondance naturelle entre homomorphisme de $\mathcal{A}$ à $\mathcal{B}$ et assignation valide de F. Ce problème est polynomial (NL-complet pour être précis).

De même, on peut coder *Horn 3-Sat* comme suit. On dispose d'une relation ternaire pour les clauses de Horn non triviales et de deux relations unaires pour les clauses unitaires. On pose $\mathcal{B} = (\{0,1\}; R; \{0\}, \{1\})$ où $R = \{(x,y,z) \mid y \wedge z \rightarrow x\}$. Ce problème est P-complet. Dans ce cas, il est bien connu que les relations de $\mathcal{B}$ sont préservées par l'opération booléenne $\wedge$ et qu'il en est de même pour l'ensemble des solutions d'un problème de type Horn en général. Ceci est la clé de l'approche algébrique que nous expliciterons en §1.4. On dit que $\wedge$ préserve la relation $R = \{0,1\}^3 \setminus \{(0,1,1)\}$ puisque pour tout choix de 2 triplets de R, par exemple $(1,1,0)$ et $(1,0,1)$ en appliquant $\wedge$ à chaque coordonnée on obtient le triplet $(1,0,0)$ qui appartient lui aussi à R.

Pour 2-Sat, l'opération booléenne ternaire h qui retourne toujours la valeur majoritaire (par exemple $h(0,1,1) = 1$) joue un rôle similaire.

Remarque. Notez que $\mathcal{A}$ peut être une structure quelconque. Ainsi, les exemples ci-dessus restent polynomiaux même si $\mathcal{A}$ est une structure non arborescente analogue à une grille par exemple.

Cette vision des problèmes de contraintes via l'homomorphisme est assez populaire car elle permet de séparer de manière naturelle *deux approches fondamentalement différentes* qui permettent d'obtenir des restrictions polynomiales. Les restrictions sur $\mathcal{A}$ sont combinatoires (lien avec les décompositions arborescentes déjà évoquées) alors que les restrictions sur $\mathcal{B}$ sont régies par les propriétés algébriques de fonctions associées à $\mathcal{B}$, comme pour les deux exemples ci-dessus[3].

Un résultat historiquement très important, qui entre dans ce second cas correspond à la classification de *Generalized Satisfiability*. Nous en esquisserons une preuve moderne en §1.4, où le treillis de Post est utilisé de manière explicite.

Théorème 3 (Schaefer 1978 [31]). *Si le domaine de $\mathcal{B}$ a deux éléments alors le problème d'homomorphisme est soit polynomial, soit NP-complet. Si les relations de $\mathcal{B}$ sont* 0-valide, 1-valide, 2-Sat, Horn, dual-Horn, *ou* affine *alors le problème devient polynomial, sinon il est NP-complet.*

À l'instar du problème du $\mathcal{H}$-coloring, pour lequel $\mathcal{H}$ est fixé, on fixe fréquemment la structure $\mathcal{B}$ et seule la structure $\mathcal{A}$ est une entrée du problème. On dénote ce problème par CSP($\mathcal{B}$) et on parle dans ce cas de problème de contrainte *non-uniforme*. On notera par la suite par CSP la classe formée par ces problèmes non-uniformes.

Remarque. Cette vision est quelque peu restrictive et ne permet pas de capturer les *contraintes globales*. Cependant, il est fréquent que les algorithmes « uniformisent ». Nous reviendrons sur cet aspect plus loin.

1.2 Conjecture de la dichotomie

Un corollaire du théorème de Ladner [60] est que si P est différent de NP, alors il existe des problèmes de complexité intermédiaire *qui ne sont ni dans P, ni NP-complet.*

Dans un article qui a eu beaucoup d'influence [44], Feder et Vardi établissent que trois classes de complexité, qu'on ne définira pas formellement ici, $\mathcal{L}_1$, $\mathcal{L}_2$ et $\mathcal{L}_3$ sont *calculatoirement équivalentes à NP*, c'est-à-dire que pour tout problème Ω de NP, il existe un représentant Ω' dans la classe $\mathcal{L}_i$ telle que Ω se réduit à Ω' en temps polynomial et inversement Ω' se réduit à Ω. Ainsi, structurellement ces trois classes se comportent exactement comme NP et en particulier, par le théorème de Ladner, elles ne peuvent pas avoir de dichotomie (on suppose dorénavant que P est différent de NP).

Ces trois classes de complexité $\mathcal{L}_1$, $\mathcal{L}_2$ et $\mathcal{L}_3$ se définissent de manière logique, et correspondent aux problèmes exprimables dans trois fragments syntaxiques de la logique existentielle du second ordre ESO. Le fragment syntaxique immédiatement en dessous de ces trois fragments, la logique MMSNP, est un fragment de la

---

[3]Notons que même si c'est plutôt contre-intuitif, le théorème de Hell et Nešetřil s'explique lui aussi algébriquement [12].

logique monadique du second ordre. Cette logique MMSNP introduite par Feder et Vardi permet d'exprimer des problème combinatoires de la forme suivante : il existe un coloriage de l'entrée qui interdit localement un nombre fini d'obstructions coloriées. En particulier on peut exprimer les problèmes CSP($\mathcal{B}$), pour tout $\mathcal{B}$.

Exemple. Pour la 3 colorabilité, les obstructions seront simplement les arêtes rouge-rouge, vert-vert et bleue-bleue. La formule de MMSNP sera de la forme suivante :

$\exists$ une partition R, V, B des sommets telle que $\forall$ sommets x, y

$\neg\big(E(x,y) \wedge R(x) \wedge R(y)\big) \wedge \neg\big(E(x,y) \wedge V(x) \wedge V(y)\big) \wedge \neg\big(E(x,y) \wedge B(x) \wedge B(y)\big)$.

La logique MMSNP ne correspond pas exactement à CSP, car elle est trop générale, mais elle a de nombreux points communs avec CSP.

Théorème 4 (Feder et Vardi 1993 [44], Kun[4])**.**

1. *Tout CSP s'exprime par une formule de MMSNP, mais il existe des formules de MMSNP qui n'expriment pas un problème de CSP.*

2. *Par contre, MMSNP est calculatoirement équivalente à CSP.*

Remarques. La logique MMSNP permet d'exprimer un problème comme « pas de triangle » qui n'est pas un problème de contraintes non-uniforme. Il s'agit en fait d'un problème de contrainte non-uniforme *de domaine infini* au sens de Bodirsky [5]. On peut déterminer de manière effective lorsqu'un problème de MMSNP est un CSP à domaine fini ou bien à domaine infini [74]. Pour MMSNP, l'inclusion de 2 problèmes est décidable mais plus complexe que pour CSP : pour CSP, l'inclusion est un CSP uniforme (NP-complet) alors que pour MMSNP on est au moins au second niveau de la hiérarchie polynomiale [65].

Les résultats de dichotomie historiques de Schaefer, et Hell et Nešetřil et des considérations techniques sur la nature foncièrement différente d'une classe $\mathcal{L}_i$ et de MMSNP ont conduit Feder et Vardi à avancer la conjecture suivante en 1993.

Conjecture de la dichotomie. *Tout problème de contrainte non-uniforme est soit polynomial soit NP-complet.*

Notons que cette conjecture implique que MMSNP a une dichotomie puisque MMSNP est calculatoirement équivalent à CSP. L'inverse est trivialement vrai puisque CSP est une restriction de MMSNP. Puisque, les classes de complexité $\mathcal{L}_i$ « immédiatement au dessus » de MMSNP contiennent des problèmes de complexité intermédiaire et que MMSNP est calculatoirement équivalent à CSP, on peut donc voir CSP comme la plus grande classe qui devrait avoir une dichotomie.

---

[4]L'une des réductions reposait sur un lemme dit de Erdös de nature aléatoire, mais ce lemme a été déterminisé par Kun.

De manière générale, on dira qu'une classe de complexité $\mathcal{C}$ est *dichotomie-complète* si $\mathcal{C}$ est calculatoirement équivalent à CSP, puisque démontrer que $\mathcal{C}$ a une dichotomie impliquerait la conjecture de la dichotomie.

Exemples. Le problème d'homomorphisme pour les graphes orientés est dichotomie-complet [44]. Le problème d'homomorphisme pour les algèbres ayant deux fonctions unaires est dichotomie-complet [43].

Lorsqu'on s'intéresse au cas spécial où la formule MMSNP est du premier ordre, c'est-à-dire que les obstructions ne sont pas coloriées comme dans le cas de l'exemple « pas de triangle », le cas où cette formule exprime un CSP fini est étudié sous le nom de *paire duale* en combinatoire. Par exemple, si la formule n'a qu'une obstruction qui est un chemin orienté de longueur n, cela revient à avoir un homomorphisme dans un tournoi transitif à n sommets. L'étude des paires duales et de la dualité est une question importante en combinatoire, voir [13] pour un survey.

On peut aussi se poser la question si cette différence d'expressivité entre MMSNP et CSP s'estompe si on restreint la classe des structures considérées. Ceci est le cas lorsque les structures sont de degré borné, sont planaires, ou plus généralement définissables par mineurs interdits[5] : dans ces cas, une formule de MMSNP, qui définit en général un CSP de domaine infini, devient forcément un CSP de domaine fini [68].

1.3 Graphe des contraintes arborescent

Nous nous tournons dans cette section vers les restrictions dites structurelles, qui concernent uniquement la structure $\mathcal{A}$, la structure $\mathcal{B}$ étant quelconque. Il s'agit de restrictions qui se comportent particulièrement bien et qui impliquent l'existence d'un algorithme polynomial même lorsque $\mathcal{B}$ n'est plus fixe et participe aussi à l'entrée. On dénote ce problème dit uniforme par CSP($\mathscr{C}$, _), pour le cas où l'entrée $\mathcal{A}$ appartient à une classe $\mathscr{C}$ de structure et $\mathcal{B}$ est quelconque.

Question 5. *Quelles sont les classes $\mathscr{C}$ pour lesquelles CSP($\mathscr{C}$, _) est polynomial ?*

Nous avons évoqué qu'un problème de contrainte est facile à résoudre si l'entrée est un arbre. Si l'entrée n'est pas un arbre mais peut être décrite par une décomposition arborescente de largeur au plus k, où k est une constante, alors le problème peut être résolu en temps polynomial n^k. Par contre cet algorithme nécessite d'avoir la décomposition. Décider si une structure $\mathcal{A}$ a une telle largeur arborescente est NP-complet si k est une donnée du problème. Par contre si la largeur k est fixée, il existe un algorithme linéaire en la taille de $\mathcal{A}$ qui calcule, si c'est possible, une telle décomposition.

---

[5]Un mineur d'un graphe $\mathcal{G}$ est un graphe $\mathcal{H}$ obtenu par contraction d'arête(s) et suppression de sommet(s) de $\mathcal{G}$. Le célèbre théorème de Kuratowski caractérise les graphes planaires comme étant les graphes n'ayant ni $\mathcal{K}_5$ ni $\mathcal{K}_{3,3}$ comme mineur. La classe des graphes planaires est un exemple de classe définissable par mineurs interdits.

On travaille sur des structures et non des graphes. On se ramène à un graphe en considérant « le graphe des contraintes ». Ce graphe est connu sous le nom de graphe de Gaifman (de $\mathcal{A}$) en logique. Ce graphe a le même domaine que $\mathcal{A}$ et on a une arête entre deux sommets s'ils participent à un même tuple d'une relation quelconque de $\mathcal{A}$. Ainsi, les sommets de $\mathcal{A}$ participant à un tuple formeront une clique dans ce graphe. La complexité de l'algorithme sera alors $n^{O(k)}$ où k est la largeur arborescente du graphe de Gaifman, le $O(k)$ cachant la complexité du codage d'une structure relationnelle en un graphe. On peut donc prendre pour $\mathscr{C}$ la classe $\mathscr{T}_k$ des structures $\mathcal{A}$ dont le graphe de Gaifman a largeur d'arborescence au plus k.

On peut se demander si cette restriction structurelle est la plus générale possible. En fait, on peut autoriser $\mathcal{A}$ a être non pas une structure de $\mathscr{T}_k$ mais *une structure homomorphiquement équivalente à une structure $\mathcal{A}'$ appartenant à $\mathscr{T}_k$*. On notera cette nouvelle classe par $\mathscr{HT}_k$.

Théorème 6 (Dalmau, Kolaitis, Vardi [34]). *Pour tout k fixé, le problème de contrainte uniforme CSP($\mathscr{HT}_k$, _) est polynomial.*

Exemple. On peut par exemple considérer l'ensemble $\mathscr{B}$ des graphes bipartis. Lorsqu'un graphe biparti a une arête, son core est $\mathcal{K}_2$ et sinon un seul sommet c'est-à-dire $\mathcal{K}_1$. Ces deux graphes sont des arbres, donc $\mathscr{B}$ est inclus dans $\mathscr{HT}_1$. Notez que $\mathscr{B}$ contient les grilles qui sont le prototype même du graphe n'ayant pas une largeur arborescente bornée. CSP($\mathscr{B}$, _) est polynomial pour une raison triviale : pour une entrée $(\mathcal{A}, \mathcal{B})$, il suffit de tester si $\mathcal{B}$ contient une arête ou si $\mathcal{A}$ n'en contient pas.

Pourquoi le problème CSP($\mathscr{HT}_k$, _) est-il polynomial ? Puisque $\mathcal{A}$ et $\mathcal{A}'$ sont homomorphiquement équivalents, ces deux structures seront ou bien simultanément acceptées ou bien simultanément rejetées du fait de leur équivalence (on rappelle que la composition de deux homomorphismes est un homomorphisme). Ainsi, il suffit étant donné $\mathcal{A}$ de calculer $\mathcal{A}'$, de calculer une décomposition de $\mathcal{A}'$ et d'utiliser l'algorithme polynomial reposant sur cette décomposition. Le problème avec cet argument c'est que calculer $\mathcal{A}'$ est une question difficile. En effet, étant donné $\mathcal{A}$, déterminer si un tel $\mathcal{A}'$ existe est NP-complet, même lorsque k est fixé. L'algorithme polynomial pour CSP($\mathscr{HT}_k$, _) ne nécessitera pas le calcul d'une décomposition : il repose sur un jeu associé à l'expressivité dans une logique, qui est un fragment de la logique du premier ordre n'ayant que $k+1$ variables.

Dans le cas des arbres, on a $k = 1$. La logique a seulement 2 variables et dans le jeu associé, chaque joueur a deux pions. Le jeux oppose le *saboteur* au *copieur* qui jouent sur deux structures $\mathcal{A}$ et $\mathcal{B}$: le premier essaye d'exhiber des différences entre $\mathcal{A}$ et $\mathcal{B}$ alors que le second essaye de montrer leur similarité. Le saboteur joue sur $\mathcal{A}$ et le copieur sur $\mathcal{B}$.

Au début, le saboteur pose deux pions p_1 et p_2 où il le veut sur deux sommets de la structure $\mathcal{A}$ (les sommets pouvant coïncider). Le copieur répond en posant deux pions q_1 et q_2 sur deux sommets de $\mathcal{B}$. Ensuite, à chaque tour suivant, le

saboteur ramasse un ou plusieurs pions et les replace sur $\mathcal{A}$ où il le souhaite ; et, le copieur procède de même. Par exemple, si le saboteur déplace son pion p_2, le copieur doit jouer q_2.

Le saboteur gagne la partie si après un tour donné, la fonction qui envoie le sommet de $\mathcal{A}$ sur lequel est placé p_1 sur le sommet de $\mathcal{B}$ associé au premier pion du copieur q_1 et de même pour la paire de sommets correspondant à p_2 et q_2, n'est pas un homomorphisme partiel de $\mathcal{A}$ dans $\mathcal{B}$. Dans le cas des graphes, cela correspond au cas où les pions du saboteur sont adjacents sur $\mathcal{A}$ et ceux du copieur ne le sont pas sur $\mathcal{B}$. Le copieur a une stratégie gagnante s'il peut jouer indéfiniment sans perdre.

Si il existe un homomorphisme de $\mathcal{A}$ dans $\mathcal{B}$, alors le copieur peut utiliser cet homomorphisme pour répondre à n'importe quel coup du saboteur et a donc une stratégie gagnante.

Si $\mathcal{A}$ est un chemin et que le copieur a une stratégie gagnante, on peut construire un homomorphisme h de $\mathcal{A}$ dans $\mathcal{B}$. On utilise les deux pions du saboteur pour « marcher » le long du chemin $\mathcal{A}$ en déplaçant alternativement le premier pion puis le second pion. On regarde la réponse du copieur. L'homomorphisme partiel existant à chaque étape construit progressivement l'homomorphisme de $\mathcal{A}$ dans $\mathcal{B}$. L'argument est similaire si $\mathcal{A}$ est un arbre.

On a donc démontré que si $\mathcal{A}$ est un arbre, alors il y a un homomorphisme de $\mathcal{A}$ dans $\mathcal{B}$ si et seulement si le copieur a une stratégie gagnante. Il se trouve qu'on peut déterminer en temps polynomial si c'est le cas. On a donc pour l'instant une preuve alternative du théorème 1 utilisant la logique et les jeux. Si $\mathcal{A}$ n'est pas un arbre mais que $\mathcal{A}'$ est un arbre homomorphiquement équivalent à $\mathcal{A}$, on peut utiliser ces homomorphismes pour montrer que le copieur a une stratégie gagnante si et seulement si il y a un homomorphisme de $\mathcal{A}$ dans $\mathcal{B}$.

Le théorème précédent caractérise précisément les restrictions structurelles qui permettent d'obtenir un algorithme polynomial, puisque Grohe montre l'implication inverse.

Théorème 7 (Dalmau *et. al.*[34], Grohe [49]). *Sous une hypothèse de complexité paramétrée* [6], *pour toute classe de structures $\mathscr{C}$ qui est récursivement énumérable, $CSP(\mathscr{C}, \_)$ est polynomial si et seulement si $\mathscr{C}$ est inclus dans $\mathscr{HT}_k$, pour k fixé.*

À la lumière de ces résultats, on peut penser que la question des restrictions structurelles est complètement résolue. Ce n'est pas tout à fait le cas puisque le cadre de travail proposé a deux inconvénients. D'une part, il ne permet que de travailler dans le cas où l'arité d'une contrainte est bornée, ce qui est plutôt réducteur si on voit l'importance qu'ont prises les contraintes globales. D'autre part, la polynomialité de $CSP(\mathscr{HT}_k, \_)$ est en quelque sorte inaccessible puisque le méta-problème qui consiste à décider si $\mathcal{A}$ appartient à $\mathscr{HT}_k$ est NP-complet même si k est fixé.

---

[6] $FPT \neq W[1]$: une conjecture analogue à $P \neq NP$ en complexité paramétrée [40].

Ce domaine d'étude reste donc particulièrement actif. Il y a de nombreux travaux portant sur la décomposition d'hypergraphes plutôt que de graphes afin de pouvoir appréhender le cas des contraintes globales. Ces travaux ont beaucoup de liens avec les bases de données puisque le problème de satisfaction de contraintes d'instance $(\mathcal{A}, \mathcal{B})$ correspond au problème d'évaluation d'une requête conjonctive $\varphi_{\mathcal{B}}$ sur une base de données $\mathcal{A}$. Pour plus de détails, voir par exemple [95].

La notion la plus générale dans le cas des hypergraphes est lié à la notion de *fractional hyper-tree width* proposée par Grohe et Marx [50]. Une classification complète par rapport à la complexité paramétrée a été donné par Marx [81].

1.4 Approche algébrique

Dans cette section, on considère que la structure $\mathcal{B}$ de domaine Dom et de relations Γ est fixée. On appelle Γ le *langage des contraintes*. On s'intéresse au pouvoir d'expressivité du langage de contrainte Γ. En particulier, on aimerait pouvoir comparer l'expressivité de différents langages de contraintes. Puisqu'on va être amené à changer Γ mais pas Dom, on notera CSP(Γ) pour CSP($\mathcal{B}$) dans la suite.

Exemple. Supposons que Γ contiennent deux relations binaires R_1 et R_2. Les deux contraintes $((x, z), R_1)$ et $((z, y), R_2)$ induisent une contrainte implicite $((x, y), R_3)$, où $R_3 = R_1 \circ R_2$. La relation R_3 n'est pas forcément dans Γ mais elle est *exprimable* par Γ. Il y a une réduction polynomiale de CSP($\Gamma \cup \{R_3\}$) dans CSP(Γ) puisqu'on peut remplacer chaque contrainte R_3 par un gadget composé de deux contraintes R_1 et R_2 et d'une variable additionnelle. Il existe une réduction triviale dans l'autre direction. Les deux problèmes CSP(Γ) et CSP($\Gamma \cup \{R_3\}$) sont donc polynomialement équivalents.

Il est important de noter que R_3 est l'interprétation de la formule $\exists z \, R_1(x, z) \wedge R_2(z, y)$.

On note par $\langle \Gamma \rangle$ l'ensemble des relations *exprimables* à partir de celles de Γ. Formellement, il s'agit de toutes les relations qu'on peut obtenir en interprétant des formules *primitives positives* sur Γ (il s'agit de formules contenant $\exists, \wedge, =$). Notez que $\langle \Gamma \rangle$ contient une infinité de relations. Jusqu'ici nous ne considérions que des langages de contraintes finis. On dira qu'un langage infini est polynomial[7] ssi tous ses fragments finis le sont. Le théorème suivant nous permet de restreindre notre étude de la complexité aux langage de contraintes tels que $\Gamma = \langle \Gamma \rangle$.

Théorème 8 (Jeavons 1998 [54])**.** *Si Γ_1 et Γ_2 sont des langages de contraintes tels que $\langle \Gamma_1 \rangle \subseteq \langle \Gamma_2 \rangle$ alors CSP(Γ_1) se réduit polynomialement[8] à CSP(Γ_2).*

---

[7]Pour être précis, il s'agit de *local tractability*. En pratique, des algorithmes vont « uniformiser » : par exemple, l'algorithme pour résoudre Horn-3-Sat est le même pour Horn-Sat. Ce second cas est dit *globally tractable*.

[8]En fait, cette réduction est même dans logspace et préservera donc une analyse de complexité plus fine avec des classes comme L ou NL.

On peut considérer ces ensembles clos par $\langle . \rangle$, les *clones relationnels*, par rapport à l'inclusion et chercher parmi eux les langages Γ maximaux (respectivement minimaux) qui sont polynomiaux (respectivement NP-complets). Ceci revient à étudier la frontière entre cas faciles et difficiles dans un treillis connu sous le nom de *treillis des clones relationnels* en algèbre universelle. Dans le cas booléen, lorsque Dom a deux éléments, ce treillis est connu sous le nom de *treillis de Post* et les preuves modernes du théorème de Schaefer s'appuient très directement sur ce treillis. En fait le treillis de Post ne concerne pas des relations mais des fonctions. On passe de l'un à l'autre par la notion de *préservation* déjà entrevue et qu'on ne définira pas formellement ici.

Exemples. Si Γ est l'ensemble des relations correspondant à des clauses de Horn, alors toutes les relations de Γ sont préservées par l'opération booléenne $\wedge$. On dira dans ce cas que $\wedge$ est un *polymorphisme* de Γ.

Si Γ est l'ensemble des relations 0-valides, c'est-à-dire contenant un r-uplet avec seulement des 0, alors la fonction constante 0 est un polymorphisme de Γ.
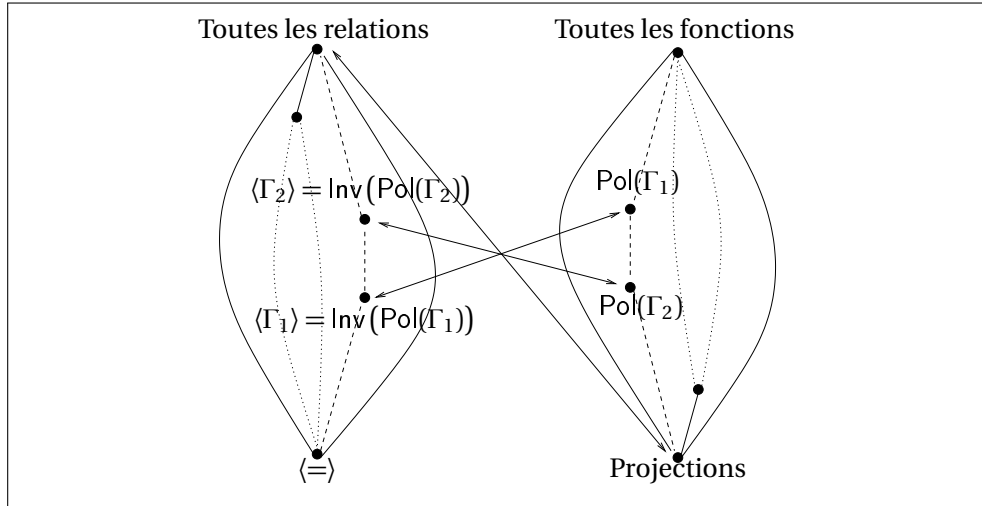
FIGURE 1.2: Correspondance de Galois Pol-Inv

On note par $\mathsf{Pol}(\Gamma)$ l'ensemble de tous les polymorphismes de Γ. Inversement, si F est un ensemble de fonctions[9], on note par $\mathsf{Inv}(F)$ l'ensemble des relations préservées (**inv**ariantes) par toutes les fonctions de F. Les opérateurs Pol et Inv établissent une correspondance de Galois entre deux treillis, l'un étant celui des clones relationnels, l'autre étant celui des *clones (fonctionnels)*. Puisque, plus il y a de relations à préserver, moins il y a de fonctions qui vont les préserver, et inversement, les deux treillis ont exactement la même structure, mais le haut de l'un correspond au bas de l'autre[10] (cf. Figure 1.2). En particulier, le langage de toutes

---

[9] de Dom^r dans Dom, r étant une arité quelconque.
[10] Pol et Inv sont des anti-isomorphismes de treillis.

les relations correspond au clone des fonctions triviales que sont les projections.[11] Le clone de toutes les fonctions ne laisse *grosso modo* invariant que la relation =, et correspond au clone relationnel $\langle = \rangle$.

Intuitivement les polymorphismes contrôlent l'expressivité et donc la complexité. Comme les deux treillis sont structurellement les mêmes on va travailler sur le treillis des clones.
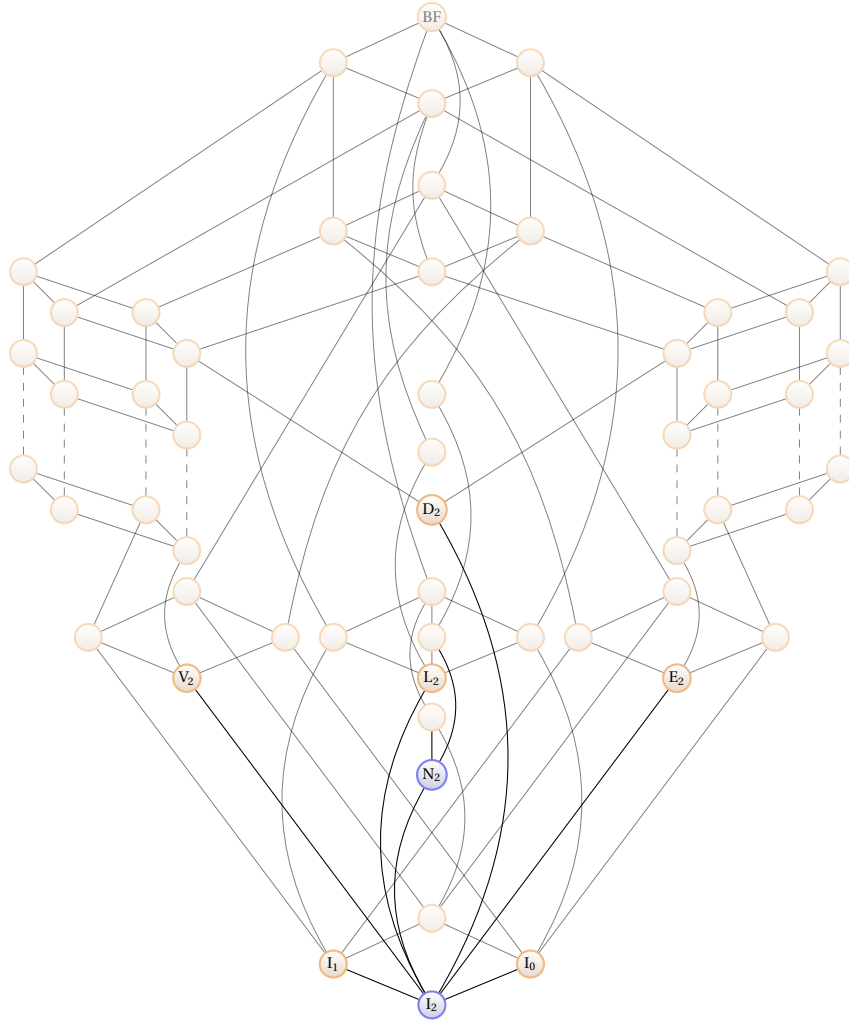
FIGURE 1.3: Treillis de Post (1941) et classification de la complexité des CSP booléens.

Les éléments du treillis de Post (cf. Figure 1.3) correspondent à des clones booléens, c'est-à-dire des ensemble de fonctions booléennes. Sur cette figure, on a tout en haut, toutes les fonctions booléennes (BF), tout en bas les projections (I_2).

---

[11]La *i*ème projection d'arité *r* retourne son *i*ème argument.

La classification de Schaefer (voir théorème 3) peut être vue comme un corollaire du théorème de Jeavons, de la description du treillis de Post et d'une étude de la complexité de 7 cas. Sur le treillis, ces 7 cas sont I_0, I_1, V_2, E_2, L_2, D_2 (qui correspondent aux 6 langages de contraintes polynomiaux et maximaux) et N_2 (qui est NP-complet et minimal).

Remarque. Ce qui est remarquable mais n'apparaît pas dans notre version aseptisée du théorème de Schaefer, c'est que chaque langage de contrainte maximalement polynomial est caractérisé en terme de préservation par une fonction booléenne. Ainsi, on peut disposer d'un algorithme général qui saura *identifier* les cas pour lesquels il existe un algorithme polynomial, et basculer s'il y a lieu en mode polynomial.

Le cas I_0 (préservation par la constante 0) correspond aux relations 0-valides déjà évoqué (I_1 est similaire pour la constante 1) qui est sans intérêt. Le cas E_2 (préservation par $\wedge$) correspond à Horn-Sat, un langage de contrainte intéressant puisque P-complet. Le cas V_2 est dual (préservation par $\vee$, correspond en forme propositionelle CNF à des clauses dites duales Horn, c'est-à-dire avec au plus un littéral négatif). Le cas de L_2 modélise des systèmes d'équations linéaires (préservation par $x \oplus y \oplus z \mod 2$). Le cas de D_2 correspond à 2-Sat (préservation par l'opération de majorité h, voir fin des exemples page 5).

Lorsqu'on se penche sur les cas où le domaine a au moins trois valeurs, on peut de nouveau se ramener à l'étude du treillis des clones, mais contrairement à celui de Post, ces treillis sont grands (ils ne sont plus dénombrables), très complexes et largement inconnus par les algébristes. Jeavons et ses co-auteurs ont beaucoup travaillé à la généralisation des résultats de Schaefer. Ils démontrent par exemple précisément quand *établir la k-cohérence forte* permet de décider si une instance est satisfaisable [55] : à savoir, lorsque le langage des contraintes est préservé par une opération *near-unanimity* d'arité $k + 1$.

Exemple. Le cas du langage 2-Sat déjà évoqué est un example pour $k = 2$: il y a préservation par la fonction majorité ternaire. La 2-cohérence (*path consistency*) est un algorithme complet pour le problème de décision dans ce cas restreint.

Feder et Vardi étudient ce même cas en utilisant la notion de programme Datalog, qui est intrinsèquement exécutable en temps polynomial [44]. Datalog est un cadre plus riche puisqu'il permet aussi d'expliquer le cas d'Horn-Sat de nature différente. Un programme Datalog calcule récursivement des relations auxiliaires. Intuitivement, c'est une forme de propagation de contraintes. On dérive localement de nouvelles contraintes à l'aide des règles du programme (l'optique étant de dériver une contradiction).

Exemple. Le programme suivant décide si une instance de Horn-3-Sat est insatisfaisable.

$$
\begin{aligned}
\lambda(x) &\leftarrow 1(x) \\
\lambda(x) &\leftarrow \lambda(y), \lambda(z), R(x, y, z) \\
\gamma &\leftarrow \lambda(x), 0(x)
\end{aligned}
$$

Le prédicat constant γ est le *but du programme* : si ce dernier « est activé » alors l'instance est rejetée.

Malgré ces avancés importantes, les résultats sont longtemps restés parcellaires et l'analyse sur le treillis ne permettait pas d'obtenir de classification complète. Il a fallu attendre Bulatov pour une classification complète du cas où il y a seulement *trois valeurs* [16]. Ce résultat peut sembler très incrémental mais conceptuellement il contient des idées très importantes qui permettent de s'abstraire du treillis et de travailler directement sur des variétés[12]. En effet, un langage de contrainte correspond à une algèbre et localement cette algèbre ne peut prendre que 5 types (par exemple elle se comporte comme un treillis ou encore comme une algèbre booléenne) : l'analyse de ces types permet d'établir la complexité associée au langage de contrainte. Ce renouveau de l'approche algébrique a permis à Bulatov d'établir une classification complète, *pour n'importe quel domaine fini*, de la complexité des langages de contraintes dits *conservatifs* [17]. Il s'agit du cas où on peut associer à chaque variable un domaine, une hypothèse qui semble être prévalente en intelligence artificielle.

Ce second résultat très important confirme ce que Feder et Vardi avaient suggéré dans leur papier fondateur [44], à savoir la pauvreté relative de « l'arsenal algorithmique polynomial ». Il existe essentiellement deux algorithmes, pour résoudre un problème non-uniforme : l'un généralise la cohérence et correspond à l'existence d'un programme Datalog, et l'autre découle du fait qu'on peut dans certain cas représenter de manière compacte l'ensemble des solutions[13].

Il existe une reformulation algébrique de la conjecture de la dichotomie et de nombreux raffinements pour d'autres classes de complexité. Depuis quelques années, l'approche algébrique des problèmes de contraintes a contribué à redynamiser l'algèbre universelle et à apporter de nouvelles questions dans cette discipline. Après une longue série de travaux, on sait par exemple maintenant que l'assertion suivante implique la conjecture de la dichotomie.

Conjecture 9. *Si* $\mathsf{Pol}(\Gamma)$ *contient un* terme de Siggers *alors* Γ *est polynomial.*

Un terme de Siggers est une opération d'arité 4 satisfaisant les identités $f(x,x,x,x) = x$ et $f(y,x,y,z) = f(x,y,z,x)$.

Pour plus de détails sur la dichotomie de la conjecture et la nouvelle approche algébrique, voir [15].

1.5 Autres questions

L'approche algébrique permet d'attaquer d'autres questions de complexité que la conjecture de la dichotomie. En particulier, dans le cas booléen, Creignou *et.*

---

[12]au sens de Birkhoff : il s'agit d'une classe d'algèbres clos par image homomorphique, produits et sous-algèbre.

[13]un exemple typique est la résolution d'un système d'équation linéaires.

al. montrent qu'on peut s'appuyer sur le treillis de Post pour caractériser la complexité de nombreux problèmes [33]. Parfois, on dispose d'un théorème similaire au théorème de Jeavons et on peut travailler *a priori* sur le treillis, c'est la cas pour la *circumscription* ou l'*abduction*. Dans d'autres cas, on peut s'appuyer sur le treillis pour réaliser l'étude mais on ne remarque qu'*a posteriori* que le résultat est compatible avec la structure du treillis, par exemple pour le *dénombrement* ou *l'énumération* des solutions de Sat. D'autres questions importantes ne sont pas compatibles avec la structure du treillis, c'est le cas par exemple de MaxSat, de la complexité paramétrée de Sat ou encore de son approximation.

On peut reformuler CSP comme problème de *model checking* pour les formules primitives positives (le fragment syntaxique de la logique du premier ordre autorisant $\exists, \wedge$ à l'exclusion de tout autre symbole, qu'on dénotera par $\{\exists, \wedge\}$-FO). Le *model checking problem* prend en entrée une formule φ (ici toujours du premier ordre) et en paramètre une structure $\mathcal{B}$ et demande si $\mathcal{B}$ est un modèle de φ. Comme pour le CSP non-uniforme, on considère que $\mathcal{B}$ est fixé et on cherche à connaître la complexité du problème pour chaque $\mathcal{B}$.

Dans le cadre de CSP, toutes les variables sont contrôlables. Il arrive parfois qu'on veuille modéliser des problèmes à l'aide de contraintes où certaines variables dépendent de l'environnement, voir sont contrôlées par un adversaire. Pour se faire on considère le problème de model-checking, où on s'autorise de surcroît le symbole $\forall$, c'est-à-dire pour les formules dites positives Horn (le fragment syntaxique associé sera dénoté par $\{\exists, \forall, \wedge\}$-FO) sous le nom de *problème de contraintes quantifiées* (QCSP). On peut adapter la méthode algébrique et travailler avec des clones relationnels particuliers, correspondant à une clôture par interprétation avec les symboles $\exists, \wedge, =$ mais aussi avec $\forall$. Puisque les clones relationnels ont plus de relations, ceci signifie qu'on aura moins de polymorphismes. Dans ce cas, les *polymorphismes surjectifs* caractérisent la complexité [10] lorsqu'on autorise toutes les constantes.

Pour le cas des QCSP booléens, la classification est complète, toujours à l'aide du treillis de Post, que ce soit lorsque les constantes sont autorisées [97], ou bien lorsqu'elles ne le sont pas [35, 31].

Théorème 10. *Si le domaine de $\mathcal{B}$ a deux éléments alors le problème de contrainte quantifié QCSP($\mathcal{B}$) est soit polynomial, soit* Pspace*-complet. Si les relations de $\mathcal{B}$ sont* 2-Sat*,* Horn*,* dual-Horn*, ou* affine *alors le problème devient polynomial, sinon il est* Pspace*-complet.*

En général, la classification de QCSP reste incomplète et est au moins aussi difficile que la conjecture de la dichotomie. En effet, si on considère une structure $\mathcal{B}'$ qui consiste en une structure $\mathcal{B}$ à laquelle on ajoute un somme isolé, alors toute formule comportant une variable universelle qui apparaît dans un atome est forcément non satisfaite par $\mathcal{B}'$. La complexité de QCSP($\mathcal{B}'$) est donc la même que celle de CSP($\mathcal{B}$).

On observe jusqu'à présent une *trichotomie* avec les complexités P, NP-complet et Pspace-complet [10, 22, 23, 79, 78]. Des prototypes de solveurs pour

les QCSP existent, par exemple QeCode un solveur *open source* qui est une extension du solveur CSP GeCode. QeCode autorise même de l'optimisation [3].

Barnaby Martin a initié l'étude systématique de la complexité du problème de model checking pour les fragments syntaxiques de la logique du premier ordre [76] induit par le choix de symboles dans $\{\forall, \exists, \wedge, \vee, \neg, =, \neq\}$. Il a établi que :

- soit un fragment est facilement classifiable,
- soit il se réduit au théorème de Schaefer, ou bien
- soit il s'agit d'un fragment riche du point de vue de la complexité.

Ces trois fragments plus riches sont : CSP $(\exists, \wedge, =)$, QCSP $(\forall, \exists, \wedge, =)$ et *QCSP avec disjonction* $(\forall, \exists, \wedge, \vee)$. Dans le cas de CSP ou QCSP la présence ou l'absence du symbole $=$ n'a pas d'effet sur la complexité. Par contre dans ce dernier cas, l'absence d'égalité est cruciale (en présence de l'égalité le problème est facilement classifiable, il est Pspace-complet si $\mathcal{B}$ a deux éléments ou plus et dans P sinon).

D'autres correspondances de Galois que la correspondance $\mathsf{Pol} - \mathsf{Inv}$ ont étés étudiées en algèbre universelle. Il n'y a pas vraiment de méthode générique pour démontrer les théorèmes techniques nécessaires mais Ferdinand Börner propose des recettes assez complètes de mise en œuvre [9].

fragment $\mathscr{L}$ de FO	« fonction »
absence d'$\exists$	partielle
présence d'$\exists$	totale
présence de $\forall$	« surjective »
présence de $\vee$	unaire
présence de $=$	fonction
absence de $=$	hyperfonction (l'image d'un élément est un sous-ensemble du domaine)
présence de $\neq$	injective
absence de $\neg$ au niveau atomique	préservation faible (image d'un tuple est un tuple)
présence de $\neg$ au niveau atomique	préservation forte[14] (image d'un tuple est un tuple et image d'un non-tuple est un non-tuple)

TABLE 1.1: Recette de Ferdinand Börner pour trouver la bonne notion de « fonction » pour la correspondance de Galois pour une clôture d'une ensemble de relation par un fragment syntaxique de la logique du premier ordre.

Exemples. Si on considère le fragment $\{\wedge, =\}$-FO, on aura une notion de préservation par des polymorphismes partiellement définis (ce qui a un intérêt si on veut analyser la valeur de c pour des algorithmes exponentiels $O(c^n)$ pour les problèmes de contraintes, voir [89]).

---

[14]En anglais, on trouve les adjectifs *strong* ou *full* selon les auteurs. On utilisera *full*.

Si on considère le fragment $\{\exists, \forall, \wedge, =\}$-FO, on retrouve bien les polymorphismes surjectifs.

Si on considère $\{\exists, \forall, \wedge, \vee, \neg\}$-FO, on aura une notion de préservation par des hyper-endomorphismes surjectifs forts.

Si on considère $\{\exists, \forall, \wedge, \vee\}$-FO, on aura une notion de préservation par des hyper-endomorphismes surjectifs (faibles). Nous explicitons ce dernier cas ci-dessous.

Il s'avère que dans le cas de *QCSP avec disjonction* $(\forall, \exists, \wedge, \vee)$ puisqu'on ne dispose plus de l'égalité, on ne travaille plus avec une notion de préservation par une fonction comme pour les polymorphismes avec CSP mais avec des *hyper-fonctions* : c'est-à-dire des fonctions de Dom dans l'ensemble de ses parties, ce qui rend les choses plus complexes. D'un autre côté, la présence de $\vee$ signifie qu'on peut se contenter des hyper-fonctions qui sont unaires ; et, celle de $\forall$, comme pour le cas de QCSP, signifie qu'on doivent imposer une notion adéquate de *surjectivité*.

L'expressivité de $\mathcal{B}$ et donc sa complexité pour le problème de QCSP avec disjonction est caractérisée par ses *hyper-endomorphismes surjectifs*. Pour chaque valeur finie de Dom, on se ramène à l'étude d'un treillis *fini*. Pour le cas booléen, celui-ci est trivial et on observe une dichotomie entre Pspace-complet et Logspace. Pour le cas à trois éléments, on peut calculer les éléments importants du treillis *à la main* et on obtient une tetrachotomie entre Logspace, NP-complete, Co-NP-complete et Pspace-complete. De plus, la complexité s'explique de manière uniforme par la possibilité d'*éliminer les quantificateurs* (par exemple, élimination des $\forall$ mais pas des $\exists$ donne un problème NP-complet) [71]. Pour quatre éléments, l'explosion combinatoire empêche de faire ce calcul à la main et en passant à une preuve assistée par ordinateur, on peut démontrer une tetrachotomie [80].

Finalement, en définissant une notion adéquate de *core quantifié*, en terme de propriétés de relativisation, et en le caractérisant algébriquement, on est capable de faire abstraction du treillis et de se ramener à l'étude de cas génériques semblables au cas à 4 éléments. On obtient ainsi une tetrachotomie pour n'importe quel domaine [70].

1.6 Remarques

Classifier la complexité des problèmes de satisfaction de contraintes est une question centrale en informatique théorique qui a des liens forts avec les mathématiques (algèbre, combinatoire, logique) et qui en informatique dépasse largement le cadre de l'intelligence artificielle du fait des liens importants avec les bases de données. Mathématiquement très riche et très actif, ce domaine dépasse le cadre de la question de la dichotomie, même si cette conjecture est importante.

Nous avons fait un rapide survol de différents résultats théoriques. Nous avons vu que deux types de restrictions permettant d'obtenir un problème polynomial pour une instance $(\mathcal{A}, \mathcal{B})$: soit le graphe des contraintes $\mathcal{A}$ est arborescent, soit

le langage des contraintes $\mathcal{B}$ présente de bonnes propriétés algébriques. Plus récemment, la question des *classes polynomiales hybrides* a pris de l'importance : il s'agit d'identifier des classes polynomiales nouvelles où $\mathcal{A}$ et $\mathcal{B}$ sont restreints simultanément (voir par exemple [27]).

Même si du point de vue pratique on pourrait reprocher à certaines hypothèses d'être irréalistes, de nombreux concepts ont un impact concret. Par exemple, la notion de core d'une structure et la notion de paire duale ont des applications en bases de données dans le cadre de l'intégration d'information [98, 99]. Par ailleurs, on voit apparaître des résultats qui font des hypothèses plus réalistes, par exemple sur la manière dont les contraintes sont codées [21], ou qui prennent en compte des contraintes globales [14].

1.7 Plan de ce manuscript

Dans une première partie, j'analyse de manière systématique la complexité du model checking pour les fragments syntaxiques de la logique du premiere ordre. Je commence par introduire le problème et la méthodologie employée en § 2. Ensuite en § 3, je présente les différentes notions de core qu'on obtient pour des fragments de la logique du premier ordre autre que $\{\exists, \wedge\}$-FO. Aux § 4 et § 5, je classifie la complexité de tous les fragments syntaxiques sauf pour $\{\exists, \wedge\}$-FO et $\{\exists, \forall, \wedge\}$-FO qui correspondent à CSP et QCSP. Le second chapitre étant dédié à la preuve de la tetrachotomie de $\{\exists, \forall, \wedge, \vee\}$-FO.

Dans une seconde partie, je présente quelques résultats concernant la complexité descriptive des problèmes de contraintes. Au § 7, je rappelle les résultats de Feder et Vardi concernant la logique SNP et son fragment monotone monadique sans $\neq$ (MMSNP), ainsi que les problèmes de motifs coloriés interdits, le pendant combinatoire de cette logique. Ces problèmes généralisent strictement les problèmes de contrainte, ce sont en fait des problèmes de contraintes avec un domaine infini. Je rappelle également quelques résultats de préservation par homomorphisme. Au § 8, je montre que l'inclusion des problèmes de motifs coloriés interdits est caractérisée par la notion de recoloriage si les problèmes sont donnée dans une forme normale (une question restée ouverte dans ma thèse [64]). Finalement au § 9, je rappelle des résultats concernant les problèmes de motifs interdits (non coloriés) qui sont des problèmes de contraintes (à domaine fini). Il s'agit de résultats de "dualité par homomorphisme" en combinatoire. Je montre ensuite qu'on peut relever ces résultats dans un contexte colorié, ce qui permet entre autre de conclure que tout problème de motif coloriés interdits est un problème de contrainte (à domaine fini) lorsqu'on le restreint à des structures qui sont suffisamment peu denses (comme les graphes de degré borné, les graphes planaires ou encore les graphes de largeur d'arborescence bornée).

1.8   Liens avec mes travaux

Le chapitre 1 reprend presque intégralement un article qui est une sorte de méta-*survey* que j'ai présenté à la conférence française des contraintes [66].

Les chapitres de la première partie suivent essentiellement un manuscript en cours de soumission à une revue qui clotûre un série de travaux avec Barnaby Martin sur la complexité de $\{\exists, \forall, \wedge, \vee\}$-FO. Nous avions commencé ce travail en présentant une classification à LICS 2009 pour le cas à deux ou trois élèments [69], un travail qui est paru en version longue à ACM TOCL [71] avec des résultats de Barny sur le cas à 4 élèments. Ensuite, nous avions présenté notre tetrachotomie pour un domaine arbitraire à LICS 2011 [70]. Le chapitre 3 contient aussi des résultats concernant la notion de *containment* pour QCSP, qui sont repris d'un article plus ancien présenté à LICS 2008 [24] ainsi que ceux plus récents d'un article accepté à CP 2012 [72] concernant la notion de core pour QCSP.

Concernant la seconde partie, le § 7 reprend des élèments de ma thèse (publiée dans [74]) nécessaire à la compréhension du § 8, ce dernier reprenant un travail que j'ai présenté à CP 2010 [65]. Finalement le § 9 reprouve de manière plus simple des résultats que j'ai présenté dans l'article [68] ainsi que quelques autres résultats plus récents non publiés fruits d'une collaboration avec Manuel Bodirsky (§ 9.5).

PART I

MODEL CHECKING FOR SYNTACTIC FRAGMENTS OF FIRST-ORDER LOGIC
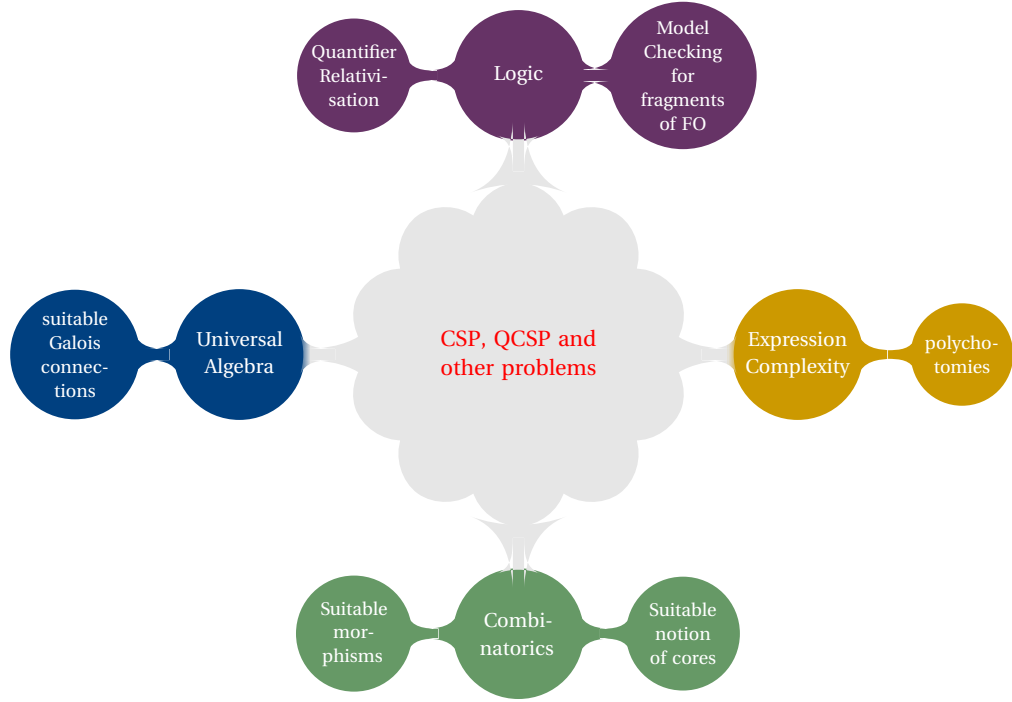
The constraint satisfaction problem CSP can be defined as the model checking problem for a fragment of first-order logic. Its extension the quantified constraint satisfaction problem QCSP can express uncertainty and models problems in which not all variables are controlled. The QCSP can also be defined as a model checking problem. We explore systematically the expression complexity of the model checking problem for all syntactic fragments $\mathscr{L}$ of first-order logic: that is, the input is a sentence φ of $\mathscr{L}$ and the structure $\mathcal{D}$ is a fixed parameter. We use suitable notion of morphisms to characterise equivalence: that is, when two structures satisfy the same sentences of $\mathscr{L}$. In particular we concentrate on cores, i.e. minimal equivalent substructures. We use suitable Galois connections to harness expressivity for a fixed domain: i.e. what kind of relation over D can one simulate using the relations in $\mathcal{D}$ and the logic $\mathscr{L}$. These tools suffices to classify all fragments but those corresponding to the CSP and the QCSP. In particular, we show that positive equality-free first-order logic, which can be seen as the extension of QCSP with disjunction, follows a tetrachotomy between L, NP-complete, co-NP-complete and Pspace-complete.

2. Introduction

The *model checking problem* over a logic $\mathscr{L}$ takes as input a structure $\mathcal{D}$ and a sentence φ of $\mathscr{L}$, and asks whether $\mathcal{D} \models \varphi$. The problem can also be parameterised, either by the sentence φ, in which case the input is simply $\mathcal{D}$, or by the model $\mathcal{D}$, in which case the input is simply φ. Vardi has studied the complexity of this problem, principly for logics which subsume First Order logic (FO) [100]. He describes the complexity of the unrestricted problem as the *combined complexity*, and the complexity of the parameterisation by the sentence (respectively, model) as the *data complexity* (respectively, *expression complexity*). For the majority of his logics, the expression and combined complexities are comparable, and are one exponential higher than the data complexity (see Table 2.1).

Complexity of the Model Checking problem for some well-known logics.			
Logic	Complexity		
	Data input $\mathcal{D}$, fixed φ	Expression fixed $\mathcal{D}$, input φ	Combined input $\mathcal{D}$ and φ
quantifier free FO	L	L	L
$\{\exists, \wedge\}$-FO	L	NP	NP
FO	L	Pspace	Pspace
TC	NL	(N)Pspace	(N)Pspace
LFP	P	E	E
ESO	NP	NE	NE

Table 2.1: Notation for the logics: TC is FO augmented with a transitive closure operator, LFP is FO augmented with a least-fixed-point operator, and ESO is existential Second Order logic (see [41] for definitions). Notation for the complexity classes: L is logarithmic space, NL is non-deterministic logspace, P is polynomial time, NP is non-deterministic polynomial time, Pspace is polynomial space (which coincides with NPspace which is non-deterministic polynomial space), E is exponential time, NE is non-deterministic exponential time (see [90] for definitions).

In this part, we will be interested in taking syntactic fragments $\mathscr{L}$ of FO, induced by the presence or absence of quantifiers and connectives, and studying the complexities of the parameterisation of the model checking problem by the model $\mathcal{D}$, that is the expression complexities for certain $\mathcal{D}$. When $\mathscr{L}$ is the *primitive positive* fragment of FO, $\{\exists, \wedge\}$-FO, the model checking problem is equivalent to the much-studied *constraint satisfaction problem* (CSP). The parameterisation of this

problem by the model $\mathcal{D}$ is equivalent to what is sometimes described as the *non-uniform* constraint satisfaction problem, CSP($\mathcal{D}$) [56]. The dichotomy conjecture is tantamount to the condition that the expression complexity for $\{\exists, \wedge\}$-FO on $\mathcal{D}$ is always either in P or is NP-complete.

When $\mathscr{L}$ is *positive Horn*, $\{\exists, \forall, \wedge\}$-FO, the model checking problem is equivalent to the well-studied *quantified constraint satisfaction problem* (QCSP). No overarching polychotomy has been conjectured for the non-uniform QCSP($\mathcal{D}$), although the only known attainable complexities are P, NP-complete and Pspace-complete.

Owing to the natural duality between $\exists, \vee$ and $\forall, \wedge$, we consider also various dual fragments. For example, the dual of $\{\exists, \wedge\}$-FO is *positive universal disjunctive* FO, $\{\forall, \vee\}$-FO. It is straightforward to see that this class of expression complexities exhibits dichotomy between P and co-NP-complete if, and only if, the class of CSPs exhibits dichotomy between P and NP-complete. Table 2.2 summarises known results regarding the complexity of the model checking for syntactic fragments of first-order logic, up to this duality.

In the case of primitive positive logic, it makes little difference whether or not equality is allowed, that is the expression complexities for $\{\exists, \wedge\}$-FO and $\{\exists, \wedge, =\}$-FO are equivalent. This is because equality may be propagated out in all but trivial instances. The same is not true of positive universal conjunctive FO; while a classification of the expression complexities over $\{\forall, \vee\}$-FO is equivalent to the unproven CSP dichotomy conjecture, though we are able to give a full dichotomy for the expression complexities over $\{\forall, \vee, =\}$-FO. The reason for this is that the equality relation in the latter simulates a disequality relation in the former. If the model $\mathcal{D}$ has $k \geq 3$ elements then $\{\exists, \wedge, \neq\}$-FO can simulate k-colourability; and, otherwise we have a Boolean model and Schaefer's dichotomy theorem provides the classification. A similar phenomenon occurs at a higher level when $\forall$ is also present.

Other fragments can be easily classified, as the model checking problem is always hard except for pathological and rather trivial models, with the notable exception of *positive equality-free first-order logic* $\{\exists, \forall, \wedge, \vee\}$-FO. For this outstanding fragment, the corresponding model checking problem can be seen as an extension of QCSP in which disjunction is returned to the mix. Note that the absence of equality is here important, as there is no general method for its being propagated out by substitution. Indeed, we will see that evaluating the related fragment $\{\exists, \forall, \wedge, \vee, =\}$-FO is Pspace-complete on any structure $\mathcal{D}$ of size at least two.

For a given fragment $\mathscr{L}$, it will be often useful to know and to characterise when two structures $\mathcal{A}$ and $\mathcal{B}$ are equivalent; that is, for any sentence φ in $\mathscr{L}$, $\mathcal{A} \models \varphi$ if, and only if, $\mathcal{B} \models \varphi$. This can be achieved by giving a characterisation of the notion of *containment* – for any sentence φ in $\mathscr{L}$, if $\mathcal{A} \models \varphi$ then $\mathcal{B} \models \varphi$ – in combinatorial terms. We will also be interested in a minimal structure $\mathcal{B}$ which is equivalent to $\mathcal{A}$. In many cases, the structure $\mathcal{B}$ turns out to be unique up to isomorphism and an induced substructure of $\mathcal{A}$.

For example, when $\mathscr{L}$ is $\{\exists, \wedge\}$-FO then containment is characterised by the ex-

Polychotomies for the expression complexity of the model checking problem		
Fragment	Dual	Classification?
$\{\exists, \vee\}$ $\{\exists, \vee, =\}$	$\{\forall, \wedge\}$ $\{\forall, \wedge, \neq\}$	Trivial (in L).
$\{\exists, \wedge, \vee\}$ $\{\exists, \wedge, \vee, =\}$	$\{\forall, \wedge, \vee\}$ $\{\forall, \wedge, \vee, \neq\}$	Trivial (in L) if the core of $\mathcal{D}$ has one element and NP-complete otherwise.
$\{\exists, \wedge, \vee, \neq\}$	$\{\forall, \wedge, \vee, =\}$	Trivial (in L) if $|D| = 1$ and NP-complete otherwise.
$\{\exists, \wedge\}$ $\{\exists, \wedge, =\}$	$\{\forall, \vee\}$ $\{\forall, \vee, \neq\}$	**CSP dichotomy** conjecture: P or NP-complete.
$\{\exists, \wedge, \neq\}$	$\{\forall, \vee, =\}$	Trivial if $|D| = 1$; in P if $|D| = 2$ and $\mathcal{D}$ is affine or bijunctive; and, NP-complete otherwise.
$\{\exists, \forall, \wedge\}$ $\{\exists, \forall, \wedge, =\}$	$\{\exists, \forall, \vee\}$ $\{\exists, \forall, \vee, \neq\}$	a **QCSP trichotomy** is observed: P, NP-complete, or Pspace-complete.
$\{\exists, \forall, \wedge, \neq\}$	$\{\exists, \forall, \vee, =\}$	Trivial if $|D| = 1$; in P if $|D| = 2$ and $\mathcal{D}$ is affine or bijunctive; and, Pspace-complete otherwise.
$\{\forall, \exists, \wedge, \vee\}$		**Positive equality free tetrachotomy**: P, NP-complete, co-NP-complete or Pspace-complete
$\{\neg, \exists, \forall, \wedge, \vee\}$		Trivial when $\mathcal{D}$ contains only trivial relations (empty or all tuples, and Pspace-complete otherwise.
$\{\forall, \exists, \wedge, \vee, =\}$ $\{\forall, \exists, \wedge, \vee, \neq\}$ $\{\neg, \exists, \forall, \wedge, \vee, =\}$		Trivial when $|D| = 1$, Pspace-complete otherwise.

Table 2.2: Expression complexity of the model checking problem according to the model for syntactic fragments of FO(L stands for logarithmic space, P for polynomial time, NP for non-deterministic polynomial time, co-NP for its dual and Pspace for polynomial space).

istence of a homomorphism from $\mathcal{A}$ to $\mathcal{B}$ and two structures have the same model checking problem w.r.t. $\{\exists, \wedge\}$-FO if, and only if, they are homomorphically equivalent. The minimal structure $\mathcal{B}$ equivalent to $\mathcal{A}$ is known as *the core of $\mathcal{A}$* and it is the smallest induced substructure $\mathcal{B}$ of $\mathcal{A}$ such that there is an endomorphism of $\mathcal{A}$ the image of which is $\mathcal{B}$.

We will study in detail the notions of containment, equivalence and cores for fragments of interest in § 3.

2.1 Basic Definitions

Unless otherwise stated, we shall work with finite relational structures that share the same finite relational signature σ. Let $\mathcal{D}$ be such a structure. We will denote its domain by D. We denote the size of such a set D by $|D|$. The *complement* $\overline{\mathcal{D}}$ of a structure $\mathcal{D}$ consists of relations that are exactly the set-theoretic complements

of those in $\mathcal{D}$. I.e., for an a-ary R, $R^{\overline{\mathcal{D}}} := D^a \setminus R^{\mathcal{D}}$. For graphs this leads to a slightly non-standard notion of complement, as it includes self-loops.

A *homomorphism* (resp. *full homomorphism*) from a structure $\mathcal{D}$ to a structure $\mathcal{E}$ is a function $h : D \to E$ that preserves (resp. preserves fully) the relations of $\mathcal{D}$, i.e. for all a_i-ary relations R_i, and for all $x_1, \ldots, x_{a_i} \in D$, $R_i(x_1, \ldots, x_{a_i}) \in \mathcal{D}$ implies $R_i(h(x_1), \ldots, h(x_{a_i})) \in \mathcal{E}$ (resp. $R_i(x_1, \ldots, x_{a_i}) \in \mathcal{D}$ iff $R_i(h(x_1), \ldots, h(x_{a_i})) \in \mathcal{E}$). $\mathcal{D}$ and $\mathcal{E}$ are *homomorphically equivalent* if there are homomorphisms both from $\mathcal{D}$ to $\mathcal{E}$ and from $\mathcal{E}$ to $\mathcal{D}$.

Let $\mathscr{L}$ be a fragment of FO. Let $\mathcal{D}$ be a fixed structure. The decision problem $\mathscr{L}(\mathcal{D})$ has:

- Input: a sentence φ of $\mathscr{L}$.

- Question: does $\mathcal{D}$ models φ?

2.2 Methodology

We will be concerned with syntactic fragments $\mathscr{L}$ of FO defined by allowing or disallowing symbols from $\{\exists, \forall, \wedge, \vee, \neq, =, \neg\}$. Clearly, given any sentence φ in $\mathscr{L}$, we may compute in logarithmic space an equivalent sentence φ' in prenex normal form, with negation pushed inwards at the atomic level. Since we will not be concerned with complexities beneath L, we assume hereafter that all inputs are in this form.

In general Pspace membership of FO($\mathcal{D}$) follows by a simple evaluation procedure inward through the quantifiers. Similarly, the expression complexity of the existential fragment $\{\exists, \wedge, \vee, \neq, =\}$-FO is at most NP; and, that of its dual fragment $\{\forall, \vee, \wedge, =, \neq\}$-FO is at most co-NP (in both cases, we may even allow atomic negation) [100]. We introduce formally below this principle of duality.

Let $\mathscr{L}$ be a syntactic fragment of FO defined by allowing or disallowing symbols from $\{\exists, \forall, \wedge, \vee, \neq, =\}$. We denote by $\overline{\mathscr{L}}$ its dual fragment by de Morgan's law: $\wedge$ is dual to $\vee$, $\exists$ to $\forall$ and $=$ to $\neq$.

Proposition 11. *Let $\mathscr{L}$ be a syntactic fragment of* FO *defined by allowing or disallowing symbols from* $\{\exists, \forall, \wedge, \vee, \neq, =\}$. *The problem $\mathscr{L}(\mathcal{D})$ belongs to a complexity class* C *if, and only if, the problem $\overline{\mathscr{L}}(\overline{\mathcal{D}})$ belongs to the dual complexity class* co-C.

Proof. For any sentence φ in $\mathscr{L}$, we may rewrite its negation $\neg\varphi$ by pushing the negation inwards until all atoms appear negatively, denoting the sentence hence obtained by ψ (which is logically equivalent to $\neg\varphi$). Next, we replace every occurrence of a negated relational symbol $\neg R$ by R to obtain a sentence of $\overline{\mathscr{L}}$ which we denote by $\overline{\varphi}$. The following chain of equivalences holds

$$\mathcal{D} \models \varphi \iff \mathcal{D} \models \neg(\neg\varphi) \iff \mathcal{D} \models \neg(\psi) \iff \mathcal{D} \not\models \psi \iff \overline{\mathcal{D}} \not\models \overline{\varphi}.$$

Clearly, $\overline{\varphi}$ can be constructed in logspace from φ and the result follows. $\qquad\square$

We will use this principle of duality to only classify one fragment or its dual: for example we will study $\{\exists, \wedge\}$-FO and ignore its dual $\{\forall, \vee\}$-FO. We will also use this principle to classify the self-dual fragment $\{\exists, \forall, \wedge, \vee\}$-FO.

We assume at least one quantifier and one binary connective (weaker fragments being trivial). By the duality principle, we may consider only purely existential fragments, or fragments with both quantifiers. Regarding connectives, we have three possibilities: purely disjunctive fragments, purely conjunctive fragments and fragments with both connectives. Regarding equality and disequality, we should have the four possible subsets of $\{=, \neq\}$ but it will become clear that cases with both follow the same complexity delineation as the case with $\neq$ only. Moreover, for fragments with both quantifiers, we may use the duality principle between $\{\exists, \forall, \wedge\}$ and $\{\forall, \exists, \vee\}$ to simplify our task. This means that we would need to consider 3×3 positive existential fragments and 2×3 positive fragments with both quantifiers. Actually, we can decrease this last count by one, due to the duality between $\{\exists, \forall, \wedge, \vee, \neq\}$-FO and $\{\exists, \forall, \wedge, \vee, =\}$-FO. Regarding fragments with $\neg$, since we necessarily have both connectives and both quantifiers, we only have to consider two fragments: FO and $\{\exists, \forall, \wedge, \vee, \neg\}$-FO. However, we shall see that the complexity of FO agrees with that of $\{\exists, \forall, \wedge, \vee, \neq\}$-FO (and its dual $\{\exists, \forall, \wedge, \vee, =\}$-FO).

This makes a grand total of *15 fragments to classify*, which are listed below; the fragments marked with a $\star$ correspond to the CSP and QCSP and are still open. We will settle all other listed fragments.

The 15 relevant fragments can be organised broadly in the following four classes.

First Class

This consists of the following trivial fragments: for such a fragment $\mathscr{L}$, the problem $\mathscr{L}(\mathcal{D})$ is trivial (in L) for any structure $\mathcal{D}$.

- $\{\exists, \vee\}$-FO   (see Proposition 55)
- $\{\exists, \vee, =\}$-FO   (see Proposition 55)
- $\{\exists, \vee, \neq\}$-FO   (see Proposition 55)

Second Class

This consists of the following fragments which exhibit a simple dichotomy: for such a fragment $\mathscr{L}$, the problem $\mathscr{L}(\mathcal{D})$ is trivial (in L) when the $\mathscr{L}$-core (defined in the next chapter) of $\mathcal{D}$ has one element and hard otherwise (NP-complete for existential fragments, Pspace-complete for fragments that allow both quantifiers). For this class, tractability amounts to the relativisation of all quantifiers to some constant.

- $\{\exists, \wedge, \vee\}$-FO, $\{\exists, \wedge, \vee, =\}$-FO   (see Proposition 59.)
- $\{\exists, \wedge, \vee, \neq\}$-FO   (see Proposition 63.)
- $\{\exists, \forall, \wedge, \vee, \neg\}$-FO   (see Proposition 58.)
- $\{\exists, \forall, \wedge, \vee, \neq\}$-FO   (see Proposition 57.)

Third Class

This exhibits more richeness complexity-wise, tractability can not be explained simply by $\mathscr{L}$-core size and relativisation of quantifiers.

- $\{\exists, \wedge, \neq\}$-FO   (see Proposition 64.)

- $\{\exists, \forall, \wedge, \neq\}$-FO   (see Proposition 65.)

- ⋆ $\{\exists, \wedge\}$-FO, $\{\exists, \wedge, =\}$-FO

- ⋆ $\{\exists, \forall, \wedge\}$-FO, $\{\exists, \forall, \wedge, =\}$-FO

Fourth Class

The last class consists of a single fragment and is rich complexity-wise, though we will see that a drop in complexity is always witnessed by relativisation of quantifiers.

- $\{\exists, \forall, \wedge, \vee\}$-FO   (see Theorem 66.)

The method

Our complexity analysis for the expression complexity of the model checking problem for a fragment $\mathscr{L}$ proceeds as follows. First, we consider only as a parameter a structure $\mathcal{D}$ that is an $\mathscr{L}$-core. Secondly, in some cases, we introduce the suitable Galois connection that underpins $\mathscr{L}$-expressibility and analyse the associated lattice to classify complexity.

In the next chapter, we discuss containment, equivalence and core. In § 4, we establish the complexity classification for the framents in the first, second and third class. Chapter 5 deals with the fourth class ($\{\exists, \forall, \wedge, \vee\}$-FO). The case of $\{\exists, \wedge, \vee\}$-FO, dealt with in § 4.3, serves as a good warm-up for understanding our approach of $\{\exists, \forall, \wedge, \vee\}$-FO, dealt with in § 5, as in both cases, we introduce and use a suitable Galois connection.

3. Containment, Equivalence and Core

It is well known that conjunctive query containment is characterised by the presence of homomorphism between the corresponding canonical databases (this goes back to Chandra and Merlin [19], see also [48, chapter 6]). For exactly the same reason, a similar result holds for $\{\exists, \wedge\}$-FO-containment (see Theorem 14), which we state and prove for pedagogical reason, before moving on to other fragments.

First, let us define the keywords of this chapter's title.

Definition 12. *Let $\mathcal{A}$ and $\mathcal{B}$ be two structures and $\mathscr{L}$ a fragment of* FO. *We say that $\mathcal{A}$ is $\mathscr{L}$-contained in $\mathcal{B}$ (respectively, $\mathscr{L}$-equivalent) if, and only if, for any φ in $\mathscr{L}$, $\mathcal{A} \models \varphi$ implies (respectively, iff) $\mathcal{B} \models \varphi$. If $\mathcal{B}$ is a minimal substructure w.r.t. inclusion such that $\mathcal{B}$ and $\mathcal{A}$ are $\mathscr{L}$-equivalent, then we say that $\mathcal{B}$ is an $\mathscr{L}$-core of $\mathcal{A}$.*

The $\{\exists, \wedge\}$-FO-core is unique up to isomorphism and is better known as *the core*. We proceed to characterise notions of containment, equivalence and core for other fragments of FO, which we will use to study the complexity of the associated model-checking problems. The results of this chapter are summarised in Table 3.1 on the next page.

In most cases, we will find that the $\mathscr{L}$-core is an *induced* substructure and that is unique up to isomorphism. It is not necessarily so in the case of $\{\exists, \forall, \wedge\}$-FO: there are examples where the $\{\exists, \forall, \wedge\}$-FO-core – which we call the *Q-core* – is not induced, and we do not know whether it is unique up to isomorphism. We have proved that it behaves well in a number of cases [72]. In the case of $\{\exists, \vee\}$-FO, the core is neither an induced substructure nor unique up to isomorphism, though $\{\exists, \vee\}$-FO is a rather contrived fragment.

More importantly, we will see that the $\{\exists, \forall, \wedge, \vee\}$-FO-core, which we will call the *U-X-core*, is induced by the union of two subsets U and X of the domain, one for each quantifier, and can be recast in terms of quantifier relativisation. This notion will be instrumental in deriving our tetrachotomy for $\{\exists, \forall, \wedge, \vee\}$-FO in § 5.

Remark 13. We could actually have used the following stronger notion. If $\mathcal{B}$ is a minimal (both w.r.t. size and inclusion) structure such that $\mathcal{B}$ and $\mathcal{A}$ are $\mathscr{L}$-equivalent, then we say that $\mathcal{B}$ is a strong $\mathscr{L}$-*core* of $\mathcal{A}$. The two notions coincide for all fragments under consideration but $\{\exists, \forall, \wedge\}$-FO, where we do not know whether the two notions coincide, and $\{\exists, \vee\}$-FO where it does not.

Fragment $\mathscr{L}$	$\mathscr{L}$-containment	$\mathscr{L}$-equivalence	$\mathscr{L}$-core
$\{\exists,\wedge\}$-FO $\{\exists,\wedge,=\}$-FO $\{\exists,\wedge,\vee\}$-FO $\{\exists,\wedge,\vee,\neq\}$-FO	homomorphism	homomorphic equivalence	(classical) core
$\{\exists,\forall,\wedge,\vee\}$-FO	surjective hypermorphism	surjective hypermorphism equivalence	U-X-core
$\{\exists,\forall,\wedge,\vee,\neg\}$-FO	Full surjective hypermorphism	Full surjective hypermorphism	quotient by $\sim$
contains $\{\exists,\wedge,\neq\}$-FO or contains $\{\forall,\vee,=\}$-FO	isomorphism	isomorphism	each structure
$\{\exists,\forall,\wedge,=\}$-FO	surjective homomorphism from a power	surjective homomorphism from a power in both directions	Q-core (not necessarily an induced substructure)
$\{\exists,\vee\}$-FO	preservation of coarsest "equality type" of tuples	same coarsest "equality types" of tuples	not unique up to isomorphism

Table 3.1: The various notions of containment, equivalence and core for syntactic fragments of FO.

3.1 Fragments from $\{\exists,\wedge\}$-FO to $\{\exists,\wedge,\vee,=\}$-FO

Given a primitive positive sentence φ in $\{\exists,\wedge\}$-FO, we denote by $\mathcal{D}_\varphi$ its *canonical database*, that is the structure with domain the variables of φ and whose tuples are precisely those that are atoms of φ. In the other direction, given a finite structure $\mathcal{A}$, we write $\varphi_\mathcal{A}$ for the so-called *canonical conjunctive query*[1] of $\mathcal{A}$, the quantifier-free formula that is the conjunction of the positive facts of $\mathcal{A}$, where the variables $v_1,\ldots,v_{|A|}$ correspond to the elements $a_1,\ldots,a_{|A|}$ of $\mathcal{A}$. It is well known that $\mathcal{D}_\varphi$ is homomorphic to a structure $\mathcal{A}$ if, and only if, $\mathcal{A} \models \varphi$. Moreover, a winning strategy for $\exists$ in the Hintikka $(\mathcal{A},\varphi)$-game is precisely a homomorphism from $\mathcal{D}_\varphi$ to $\mathcal{A}$. Note also that $\mathcal{A}$ is isomorphic to the canonical database of $\exists v_1 \exists v_2 \ldots v_{|A|}\varphi_\mathcal{A}$.

Theorem 14 (**Containment for $\{\exists,\wedge\}$-FO**)**.** *Let $\mathcal{A}$ and $\mathcal{B}$ be two structures. The following are equivalent.*

(i) *For every sentence φ in $\{\exists,\wedge\}$-FO, if $\mathcal{A} \models \varphi$ then $\mathcal{B} \models \varphi$.*

(ii) *There exists a homomorphism from $\mathcal{A}$ to $\mathcal{B}$.*

---

[1]Most authors consider the canonical query to be the sentence which is the existential quantification of $\varphi_\mathcal{A}$.

(iii) $\mathcal{B} \models \varphi_{\mathcal{A}}^{\{\exists, \wedge\}\text{-FO}}$ *where* $\varphi_{\mathcal{A}}^{\{\exists, \wedge\}\text{-FO}} := \exists v_1 \exists v_2 \ldots v_{|A|} \varphi_{\mathcal{A}}.$

Proof. As we observed above, a homomorphism corresponds to a winning strategy in the $(\mathcal{A}, \varphi)$-game and (ii) and (iii) are equivalent.

Clearly, (i) implies (iii) since $\mathcal{A} \models \exists v_1 \exists v_2 \ldots v_{|A|} \varphi_{\mathcal{A}}.$

We now prove that (ii) implies (i). Let h be a homomorphism from $\mathcal{A}$ to $\mathcal{B}$. If $\mathcal{A} \models \varphi$, then there is a homomorphism g from $\mathcal{D}_\varphi$ to $\mathcal{A}$. By composition, $h \circ g$ is a homomorphism from $\mathcal{D}_\varphi$ to $\mathcal{B}$. In other words, $h \circ g$ is a winning strategy for $\exists$ in the $(\mathcal{B}, \varphi)$-game. $\qquad \square$

We may add $\vee$ and $=$ without affecting this result.

Proposition 15. *Let $\mathcal{A}$ and $\mathcal{B}$ be two relational structures. The following are equivalent.*

  (i) *There is a homomorphism from $\mathcal{A}$ to $\mathcal{B}$.*

 (ii) *$\mathcal{A}$ is $\{\exists, \wedge\}$-FO-contained in $\mathcal{B}$.*

(iii) *$\mathcal{A}$ is $\{\exists, \wedge, =\}$-FO-contained in $\mathcal{B}$.*

(iv) *$\mathcal{A}$ is $\{\exists, \wedge, \vee\}$-FO-contained in $\mathcal{B}$.*

 (v) *$\mathcal{A}$ is $\{\exists, \wedge, \vee, =\}$-FO-contained in $\mathcal{B}$.*

Proof. The equivalence of (i) and (ii) are stated in Theorem 14 and are equivalent to $\mathcal{B} \models \exists v_1 \exists v_2 \ldots v_{|A|} \varphi_{\mathcal{A}}$, a sentence of $\{\exists, \wedge\}$-FO. This takes care of the implications from (v), (iv) and (iii) to (i). Trivially (v) implies both (iv) and (iii).

It suffices to prove (i) implies (v). As in the proof of Theorem 14, it can be easily checked that a homomorphism can be applied to a winning strategy for $\exists$ in the $(\mathcal{A}, \varphi)$-game to obtain a winning strategy for $\exists$ in the $(\mathcal{B}, \varphi)$-game. To see this, write the quantifier-free part ψ of φ in conjunctive normal form as a disjunction of conjunction-of-positive-atoms ψ_i. We may even propagate equality out by substitution such that each ψ_i is equality-free (if some ψ_i contained no extensional symbol other than equality, the sentence φ would trivially holds on any structure as we only ever consider structures with at least one element). A winning strategy in the $(\mathcal{A}, \varphi)$-game corresponds to a homomorphism from some $\mathcal{D}_{\psi_i}$ to $\mathcal{A}$. By composition with the homomorphism from $\mathcal{A}$ to $\mathcal{B}$, we get a homomorphism from $\mathcal{D}_{\psi_i}$ to $\mathcal{B}$, *i.e.* a winning strategy in the $(\mathcal{B}, \varphi)$-game as required. $\qquad \square$

Corollary 16. *Let $\mathcal{A}$ and $\mathcal{B}$ be two relational structures. The following are equivalent.*

  (i) *$\mathcal{A}$ and $\mathcal{B}$ are homomorphically equivalent.*

 (ii) *$\mathcal{A}$ and $\mathcal{B}$ have isomorphic cores.*

(iii) *$\mathcal{A}$ is $\{\exists, \wedge\}$-FO-equivalent to $\mathcal{B}$.*

(iv) *$\mathcal{A}$ is $\{\exists, \wedge, =\}$-FO-equivalent to $\mathcal{B}$.*

 (v) *$\mathcal{A}$ is $\{\exists, \wedge, \vee\}$-FO-equivalent to $\mathcal{B}$.*

(vi) *$\mathcal{A}$ is $\{\exists, \wedge, \vee, =\}$-FO-equivalent to $\mathcal{B}$.*

3.2 Fragments containing $\{\exists, \wedge, \neq\}$

Proposition 17 (**Containment for** $\{\exists, \wedge, \neq\}$-FO)**.** *Let $\mathcal{A}$ and $\mathcal{B}$ be two structures. The following are equivalent.*

(i) *For every sentence φ in $\{\exists, \wedge, \neq\}$-FO, if $\mathcal{A} \models \varphi$ then $\mathcal{B} \models \varphi$.*

(ii) *There exists an injective homomorphism from $\mathcal{A}$ to $\mathcal{B}$.*

(iii) *$\mathcal{B} \models \varphi_{\mathcal{A}}^{\{\exists, \wedge, \neq\}\text{-FO}}$ where $\varphi_{\mathcal{A}}^{\{\exists, \wedge, \neq\}\text{-FO}} := \exists v_1 \dots v_{|A|} \varphi_{\mathcal{A}} \wedge \bigwedge_{1 \leq i < j \leq |A|} v_i \neq v_j$.*

Proof. Similar to Theorem 14. □

Corollary 18. *Let $\mathscr{L}$ be a fragment of FO such that $\mathscr{L}$ or its dual $\overline{\mathscr{L}}$ contains $\{\exists, \wedge, \neq\}$-FO. Let $\mathcal{A}$ and $\mathcal{B}$ be two structures. The following are equivalent.*

(i) *$\mathcal{A}$ and $\mathcal{B}$ are isomorphic.*

(ii) *$\mathcal{A}$ is $\mathscr{L}$-equivalent to $\mathcal{B}$.*

Proof. For the case when $\mathscr{L}$ contains $\{\exists, \wedge, \neq\}$-FO, the result follows from the previous proposition and the fact that we deal with finite structures only.

For the case when $\overline{\mathscr{L}}$ contains $\{\exists, \wedge, \neq\}$-FO, we apply the duality principle and the previous case, and equivalently $\overline{\mathcal{A}}$ and $\overline{\mathcal{B}}$ are isomorphic. This is in turn equivalent to $\mathcal{A}$ being isomorphic to $\mathcal{B}$. □

3.3 Some Definitions

Hintikka Games

Before moving on to the equality-free fragments $\{\exists, \forall, \wedge, \vee\}$-FO and $\{\exists, \forall, \wedge, \vee, \neg\}$-FO, let us recall first basic definitions and notations regarding Hintikka Games. Let φ be a sentence of FO in prenex form with all negations pushed to the atomic level. A *strategy* for $\exists$ in the (Hintikka) $(\mathcal{A}, \varphi)$-game is a set of mappings $\{\sigma_x : `\exists x' \in \varphi\}$ with one mapping σ_x for each existentially quantified variable x of φ. The mapping σ_x ranges over the domain A of $\mathcal{A}$; and, its domain is the set of functions from Y_x to A, where Y_x denotes the universally quantified variables of φ preceding x.

We say that $\{\sigma_x : `\exists x' \in \varphi\}$ is *winning* if for any assignment π of the universally quantified variables of φ to A, when each existentially quantified variable x is set according to σ_x applied to $\pi|_{Y_x}$, then the quantifier-free part ψ of φ is satisfied under this overall assignment h. When ψ is a conjunction of positive atoms, this amounts to h being a homomorphism from $\mathcal{D}_\psi$ to $\mathcal{A}$.

Hyper-morphisms

For the equality-free fragments $\{\exists, \forall, \wedge, \vee, \neg\}$-FO and $\{\exists, \forall, \wedge, \vee\}$-FO, the correct concept to transfer winning strategies involves unary hyper-operations, that is functions to the power-set.

A *hyper-operation f* from a set A to a set B is a function from A to the power-set of B. For a subset S of A, we will define its image $f(A)$ under the hyper-operation

f as $\bigcup_{s \in S} f(s)$. When we wish to stress that an element may be sent to $\emptyset$, we speak of a *partial hyper-operation*; and otherwise we assume that f is *total*, that is for any a in A, $f(a) \neq \emptyset$. We say that f is *surjective* whenever $f(A) = B$. The *inverse* of a partial hyper-operation f from A to B, denoted by f^{-1}, is the (partial) hyper-operation from B to A defined for any b in B as $f^{-1}(b) := \{a \in A \mid b \in f(a)\}$. We call an element of $f^{-1}(b)$ an *antecedent* of b under f. Let f be a hyper-operation from A to B and g a hyper-operation from B to C. The hyper-operation $g \circ f$ is defined naturally as $g \circ f(x) := g(f(x))$ (recall that $f(x)$ is a set).

When f is a (total) surjective hyper-operation from A to A, we say that f is a *shop* of A. Note that the inverse of a shop is a shop (surjectivity and totality being dual concept under the inverse operation) and that the composition of two shops is a also a shop. Observing further that shop composition is associative and that the identity shop (which sends an element x of A to the singleton $\{x\}$) is the identity with respect to composition, we may consider the monoid generated by a set of shops. A shop f is a *sub-shop* of a shop g whenever, for every x in A, $f(x) \subseteq g(x)$. In our context, we will be interested in a particular monoid which will be closed further under sub-shops, a so-called *down-shop-monoid* (DSM). [2] We denote by $\langle F \rangle_{DSM}$ the DSM generated by a set F of shops.

Let f be a shop of A. When for a subset U of A we have $f(U) = A$, we say that f is *U-surjective*. Observing that the totality of f may be rephrased as $f^{-1}(A) = A$, we say more generally that f is *X-total* for a subset X of A whenever $f^{-1}(X) = A$. Note that for shops U-surjectivity and X-totality are dual to one another, that is the inverse of a U-surjective shop is an X-total shop with $X = U$ and vice versa. Somewhat abusing terminology, and when it does not cause confusion, we will drop the word surjective and by U- or U'-shop we will mean a U- or U'-surjective shop. Similarly, we will speak of an X- or X'-shop in the total case and of a U-X-shop in the case of a shop that is both U-surjective and X-total. Suitable compositions of U-shops and X-shops preserve these properties.

Lemma 19. *Let f and g be two shops.*

  (i) *If f is a U-shop then $g \circ f$ is a U-shop.*

 (ii) *If g is a X-shop then $g \circ f$ is a X-shop.*

(iii) *If both f is a U-shop and g is a X-shop then $g \circ f$ is a U-X-shop.*

(iv) *If both f and g are U-X-shops then $g \circ f$ is a U-X-shop.*

 (v) *The iterate of a U-X-shop is a U-X-shop.*

Proof. We prove (i). Since $f(U) = A$, we have $g(f(U)) = g(A)$. By surjectivity of g, we know that $g(A) = A$. It follows that $g(f(U)) = A$ and we are done. (ii) is dual to (i), and (iii) follows directly from (i) and (ii). (iv) is a restriction of (iii) and is only stated here as we shall use it often. (v) follows by induction on the order of iteration using (iv). $\qquad\square$

---

[2]The "down" comes from *down-closure*, here under sub-shops; a nomenclature inherited from [11].

A *hyper-morphism f* from a structure $\mathcal{A}$ to a structure $\mathcal{B}$ is a hyper-operation from A to B that satisfies the following property.

- (**preserving**) if $R(a_1, \ldots, a_i)$ holds in $\mathcal{A}$ then $R(b_1, \ldots, b_i)$ holds in $\mathcal{B}$, for all $b_1 \in f(a_1), \ldots, b_i \in f(a_i)$.

When $\mathcal{A}$ and $\mathcal{B}$ are the same structure, we speak of a *hyper-endomorphism*. We say that f is *full* if moreover

- (**fullness**) $R(a_1, \ldots, a_i)$ holds in $\mathcal{A}$ iff $R(b_1, \ldots, b_i)$ holds in $\mathcal{B}$, for all $b_1 \in f(a_1), \ldots, b_i \in f(a_i)$.

Note that the inverse of a full surjective hyper-morphism is also a full surjective hyper-morphism.

3.4 Equality-free first-order logic ($\{\exists, \forall, \wedge, \vee, \neg\}$-FO)

Lemma 20 (strategy transfer for $\{\exists, \forall, \wedge, \vee, \neg\}$-FO). *Let $\mathcal{A}$ and $\mathcal{B}$ be two structures such that there is a full surjective hyper-morphism from $\mathcal{A}$ to $\mathcal{B}$. Then, for every sentence φ in $\{\exists, \forall, \wedge, \vee, \neg\}$-FO, if $\mathcal{A} \models \varphi$ then $\mathcal{B} \models \varphi$.*

Proof. Let h be a full surjective hyper-morphism from $\mathcal{A}$ to $\mathcal{B}$ and φ be a sentence of $\{\exists, \forall, \wedge, \vee, \neg\}$-FO such that $\mathcal{A} \models \varphi$. We fix an arbitrary linear order over A and write $\min h^{-1}(b)$ to denote the smallest antecedent of b in A under h.

Let $\{\sigma_x : \text{`}\exists x\text{'} \in \varphi\}$ be a winning strategy in the $(\mathcal{A}, \varphi)$-game. We construct a strategy $\{\sigma'_x : \text{`}\exists x\text{'} \in \varphi\}$ in the $(\mathcal{B}, \varphi)$-game as follows. Let $\pi_B : Y_x \to B$ be an assignment to the universal variables Y_x preceding an existential variable x in φ, we select for $\sigma'_x(\pi)$ an arbitrary element of $h(\sigma(\pi_A))$ where $\pi_A : Y_x \to A$ is an assignment such that for any universal variable y preceding x, we have $\pi_A(y) := \min h^{-1}(\pi_B(y))$. This strategy is well defined since h is surjective (which means that π_A is well defined) and total (which means that $h(\sigma(\pi_A)) \neq \emptyset$). Note moreover that using min in the definition of π_A means that a branch in the tree of the game on $\mathcal{B}$ will correspond to a branch in the tree of the game on $\mathcal{A}$. It remains to prove that $\{\sigma'_x : \text{`}\exists x\text{'} \in \varphi\}$ is winning. We will see that it follows from the fact that h is full and preserving.

We assume that negations have been pushed to the atomic level and write the quantifier-free part ψ of φ in disjunctive normal form as a disjunction of conjunctions-of-atoms ψ_i. If ψ_i has contradictory positive and negative atoms (as in $E(x, y) \wedge \neg E(x, y)$) then we may discard the sentence ψ_i as false. Moreover, for each pair of atoms $R(v_1, v_2, \ldots, v_r)$ and $\neg R(v_1, v_2, \ldots, v_r)$ (induced by the choice of a relational symbol R and the choice of r variables $v_1, v_2, \ldots, v_r$ occuring in ψ_i) such that neither is present in ψ_i, we may replace ψ_i by the logically equivalent $(\psi_i \wedge R(v_1, v_2, \ldots, v_r)) \vee (\psi_i \wedge \neg R(v_1, v_2, \ldots, v_r))$. After this completion process, note that every conjunction of atoms ψ_i corresponds naturally to a structure $\mathcal{D}_{\psi_i}$ (take only the positive part of ψ_i which is now maximal).

Assume first that ψ is disjunction-free. The winning condition of the $(\mathcal{B}, \varphi)$-game can be recast as a full homomorphism from $\mathcal{D}_\psi$ to $\mathcal{B}$. Composing with h the

full homomorphism from $\mathcal{D}_\psi$ to $\mathcal{A}$ (induced by the sequence of compatible assignments π_A to the universal variables and the strategy $\{\sigma_x : \text{'}\exists x\text{'} \in \varphi\}$), we get a full hyper-morphism from $\mathcal{D}_\psi$ to $\mathcal{B}$. The map from the domain of $\mathcal{D}_\psi$ to $\mathcal{B}$ induced by the sequence of assignments π_B and the strategy $\{\sigma'_x : \text{'}\exists x\text{'} \in \varphi\}$ is a range restriction of this full hyper-morphism and is therefore a full homomorphism (we identify hyper-morphism to singletons with homomorphisms). In general when the quantifier-free part of φ has several disjuncts ψ_i, most likely after the completion process of the previous paragraph, the winning condition can be recast as a full homomorphism from some $\mathcal{D}_{\psi_i}$. The above argument applies and the result follows. $\qquad\square$

We shall see that the converse of Lemma 20 holds. Consequently, it turns out that containment and equivalence coincide for $\{\exists, \forall, \wedge, \vee, \neg\}$-FO, since the inverse of a full surjective hyper-morphism is a full surjective hyper-morphism.

For $\{\exists, \forall, \wedge, \vee, \neg\}$-FO, we define an equivalence relation $\sim$ over the structure elements in the spirit of the Leibnitz-rule for equality. For propositions P and Q, let $P \longleftrightarrow Q$ be an abbreviation for $(P \wedge Q) \vee (\neg P \wedge \neg Q)$. For the sake of clarity, we deal with the case of digraphs first and write $x \sim y$ as an abbreviation for $\forall z (E(x,z) \longleftrightarrow E(y,z)) \wedge (E(z,x) \longleftrightarrow E(z,y))$. It is straightforward to verify that $\sim$ induces an equivalence relation over the vertices (which we denote also by $\sim$). In general, for each r-ary symbol R, let ψ_R stands for

$$\big(R(x,z_1,\ldots,z_{r-1}) \longleftrightarrow R(y,z_1,\ldots,z_{r-1})\big) \wedge \big(R(z_1,x,z_2,\ldots,z_{r-1}) \longleftrightarrow R(z_1,y,z_2,\ldots,z_{r-1})\big)$$
$$\wedge \ldots \wedge \big(R(z_1,z_2,\ldots,z_{r-1},x) \longleftrightarrow R(z_1,z_2,\ldots,z_{r-1},y)\big).$$

We write $x \sim y$ for $\bigwedge_{R \in \sigma} \forall z_1, z_2, \ldots, z_{r-1} \psi_R$.

We write $\mathcal{A}/\sim$ for the quotient structure defined in the natural way. Note that there is a full surjective homomorphism from $\mathcal{A}$ to $\mathcal{A}/\sim$. As observed earlier, its inverse (viewing the homomorphism as an hyper-morphism) is a full surjective hyper-morphism from $\mathcal{A}/\sim$ to $\mathcal{A}$. Thus, it follows from Lemma 20 that $\mathcal{A}$ and $\mathcal{A}/\sim$ are $\{\exists, \forall, \wedge, \vee, \neg\}$-FO-equivalent.

Let $\varphi_\mathcal{A}^+$ denotes the (quantifier-free) canonical conjunctive query of $\mathcal{A}$ (denoted earlier as $\varphi_\mathcal{A}$) and $\varphi_\mathcal{A}^-$ denotes the similar sentence which lists the negative atoms of $\mathcal{A}$ instead of the positive atoms.

Proposition 21 (**Containment and equivalence for** $\{\exists, \forall, \wedge, \vee, \neg\}$-FO)**.** *Let $\mathcal{A}$ and $\mathcal{B}$ be two structures. The following are equivalent.*

(i) *For every sentence φ in $\{\exists, \forall, \wedge, \vee, \neg\}$-FO, if $\mathcal{A} \models \varphi$ then $\mathcal{B} \models \varphi$.*

(ii) *There exists a full surjective hyper-morphism from $\mathcal{A}$ to $\mathcal{B}$.*

(iii) $\mathcal{B} \models \varphi_\mathcal{A}^{\{\exists, \forall, \wedge, \vee, \neg\}\text{-FO}}$ *where*

$$\varphi_\mathcal{A}^{\{\exists, \forall, \wedge, \vee, \neg\}\text{-FO}} := \exists v_1 \exists v_2 \ldots v_{|A|} \varphi_\mathcal{A}^+ \wedge \varphi_\mathcal{A}^- \wedge \forall w \bigvee_{1 \le i \le |A|} w \sim v_i.$$

(iv) *for every sentence φ in $\{\exists, \forall, \wedge, \vee, \neg\}$-FO, $\mathcal{A} \models \varphi$ iff $\mathcal{B} \models \varphi$.*

(v) $\mathcal{A}/\sim$ and $\mathcal{B}/\sim$ are isomorphic.

Proof. The implication (i) to (iii) is clear since by construction $\mathcal{A}$ models the canonical sentence $\varphi_{\mathcal{A}}^{\{\exists,\forall,\wedge,\vee,\neg\}\text{-FO}}$.

We prove that (iii) implies (ii). Assume that $\mathcal{B} \models \varphi_{\mathcal{A}}^{\{\exists,\forall,\wedge,\vee,\neg\}\text{-FO}}$. We construct a full and total surjective hyper-morphism h as follows. Let $b_1, b_2, \ldots, b_{|A|}$ be witnesses in B for $v_1, v_2, \ldots, v_{|A|}$. We set $h(a_i) \ni b_i$ for $1 \le i \le |A|$ (totality). For each b in B, we set the universal variable w to b and pick some j such that $w \sim v_j$ holds and set $h(a_j) \ni b$ (surjectivity). By construction, h is preserving and full.

The implication (ii) to (i) is proved as Lemma 20.

The equivalence of (i), (ii), (iii) with (iv) follows from our earlier observation that the inverse f^{-1} of a full surjective hyper-morphism f from $\mathcal{A}$ to $\mathcal{B}$ is a full surjective hyper-morphism from $\mathcal{B}$ to $\mathcal{A}$.

To see that (v) implies (ii), compose the quotient map from $\mathcal{A}$ to $\mathcal{A}/\sim$ (which is a full surjective homomorphism) with the inverse of the quotient map from $\mathcal{B}$ to $\mathcal{B}/\sim$ (which is a full surjective hyper-morphism).

For the direction (ii) to (v), the natural quotient $f/\sim$ of a full surjective hyper-morphism f from $\mathcal{A}$ to $\mathcal{B}$ is a full surjective homomorphism. Since we deal with finite structures, it is an isomorphism and we are done. $\square$

Note that no smaller structure can be $\{\exists,\forall,\wedge,\vee,\neg\}$-FO-equivalent to $\mathcal{A}' := \mathcal{A}/\sim$. Indeed, a full surjective hyper-morphism f from a smaller structure $\mathcal{B}$ to $\mathcal{A}'$ would have to satisfy $\{a_1', a_2'\} \subseteq f(b)$ for some b in B and some distinct a_1', a_2' in A'. But this would imply that $a_1' \sim a_2'$ which is not possible. Moreover, any structure that is $\{\exists,\forall,\wedge,\vee,\neg\}$-FO-equivalent and of the same size as $\mathcal{A}'$ will be isomorphic (a full surjective hyper-morphism must induce an isomorphism by triviality of $\sim$ over $\mathcal{A}'$). Thus, $\mathcal{A}/\sim$ is the (up to isomorphism unique) $\{\exists,\forall,\wedge,\vee,\neg\}$-FO-core of $\mathcal{A}$.

3.5 Positive Equality-free first-order logic ($\{\exists,\forall,\wedge,\vee\}$-FO)

Containment

Lemma 22 (**strategy transfer for** $\{\exists,\forall,\wedge,\vee\}$**-FO**)**.** *Let $\mathcal{A}$ and $\mathcal{B}$ be two structures such that there is a surjective hyper-morphism from $\mathcal{A}$ to $\mathcal{B}$. Then, for every sentence φ in $\{\exists,\forall,\wedge,\vee\}$-FO, if $\mathcal{A} \models \varphi$ then $\mathcal{B} \models \varphi$.*

Proof. The proof is exactly the same as that of Lemma 20, except that we no longer need to preserve atomic negation, and may drop the assumption of fullness. $\square$

We extend the notion of canonical conjunctive query of a structure $\mathcal{A}$. Given a tuple of (not necessarily distinct) elements $\mathbf{r} := (r_1, \ldots, r_l) \in A^l$, define the quantifier-free formula $\varphi_{\mathcal{A}(\mathbf{r})}(v_1, \ldots, v_l)$ to be the conjunction of the positive facts of $\mathbf{r}$, where the variables $v_1, \ldots, v_l$ correspond to the elements $r_1, \ldots, r_l$. That is, $R(v_{\lambda_1}, \ldots, v_{\lambda_i})$ appears as an atom in $\varphi_{\mathcal{A}(\mathbf{r})}$ iff $R(r_{\lambda_1}, \ldots, r_{\lambda_i})$ holds in $\mathcal{A}$. When $\mathbf{r}$ enumerates the elements of the structure $\mathcal{A}$, this definition coincides with the usual definition of canonical conjunctive query. Note also that in this case there is a

full homomorphism from the canonical database $\mathcal{D}_{\varphi_{\mathcal{A}(\mathbf{r})}}$ to $\mathcal{A}$ given by the map $v_{\lambda_i} \mapsto r_i$.

Definition 23 (**Canonical** $\{\exists, \forall, \wedge, \vee\}$-FO **sentence**). *Let $\mathcal{A}$ be a structure and $m > 0$. Let $\mathbf{r}$ be an enumeration of the elements of $\mathcal{A}$.*

$$\theta_{\mathcal{A},m}^{\{\exists, \forall, \wedge, \vee\}\text{-FO}} := \exists v_1, \ldots, v_{|A|} \varphi_{\mathcal{A}(\mathbf{r})}(v_1, \ldots, v_{|A|}) \wedge \forall w_1, \ldots, w_m \bigvee_{\mathbf{t} \in A^m} \varphi_{\mathcal{A}(\mathbf{r},\mathbf{t})}(\mathbf{v}, \mathbf{w}).$$

Observe that $\mathcal{A} \models \theta_{\mathcal{A},m}^{\{\exists, \forall, \wedge, \vee\}\text{-FO}}$. Indeed, we may take as witness for the variables $\mathbf{v}$ the corresponding enumeration $\mathbf{r}$ of the elements of $\mathcal{A}$; and, for any assignment $\mathbf{t} \in A^m$ to the universal variables $\mathbf{w}$, it is clear that $\mathcal{A} \models \varphi_{\mathcal{A}(\mathbf{r},\mathbf{t})}(\mathbf{r}, \mathbf{t})$ holds.

Lemma 24. *Let $\mathcal{A}$ and $\mathcal{B}$ be two structures. If $\mathcal{B} \models \theta_{\mathcal{A},|\mathcal{B}|}^{\{\exists, \forall, \wedge, \vee\}\text{-FO}}$ then there is a surjective hyper-morphism from $\mathcal{A}$ to $\mathcal{B}$.*

Proof. Let $\mathbf{b}' := b_1', \ldots, b_{|A|}'$ be witnesses for $v_1, \ldots, v_{|A|}$. Assume that an enumeration $\mathbf{b} := b_1, b_2, \ldots, b_{|B|}$ of the elements of $\mathcal{B}$ is chosen for the universal variables $w_1, \ldots w_{|B|}$. Let $\mathbf{t} \in A^m$ be the witness s.t. $\mathcal{B} \models \varphi_{\mathcal{A}(\mathbf{r})}(\mathbf{b}') \wedge \varphi_{\mathcal{A}(\mathbf{r},\mathbf{t})}(\mathbf{b}', \mathbf{b})$.

Let f be the map from the domain of $\mathcal{A}$ to the power set of that of $\mathcal{B}$ which is the union of the following two partial hyper-operations h and g (*i.e.* $f(a_i) := h(a_i) \cup g(a_i)$ for any element a_i of $\mathcal{A}$), which guarantee totality and surjectivity, respectively.

- $h(a_i) := b_i'$ (totality.)

- $g(t_i) \ni b_i$ (surjectivity.)

It remains to show that f is preserving. This follows from $\mathcal{B} \models \varphi_{\mathcal{A}(\mathbf{r},\mathbf{t})}(\mathbf{b}', \mathbf{b})$.

Let R be a r-ary relational symbol such that $R(a_{i_1}, \ldots, a_{i_r})$ holds in $\mathcal{A}$. Let $b_{i_1}'' \in f(a_{i_1}), \ldots, b_{i_r}'' \in f(a_r)$. We will show that $R(b_{i_1}'', \ldots, b_{i_r}'')$ holds in $\mathcal{B}$. Assume for clarity of the exposition and w.l.o.g. that from i_1 to i_k the image is set according to h and from i_{k+1} to i_r according to g: i.e. for $1 \leq j \leq k$, $h(a_{i_j}) = b_{i_j}' = b_{i_j}''$ and for $k+1 \leq j \leq r$, there is some l_j such that $t_{l_j} = a_{i_j}$ and $g(t_{l_j}) \ni b_{i_j}'' = b_{l_j}$. By definition of $\mathcal{A}(\mathbf{r}, \mathbf{t})$ the atom $R(v_{i_1}, \ldots, v_{i_k}, w_{l_{k+1}}, \ldots, w_r)$ appears in $\varphi_{\mathcal{A}(\mathbf{r},\mathbf{t})}(\mathbf{v}, \mathbf{w})$. It follows from $\mathcal{B} \models \varphi_{\mathcal{A}(\mathbf{r},\mathbf{t})}(\mathbf{b}', \mathbf{b})$ that $R(b_{i_1}'', \ldots, b_{i_r}'')$ holds in $\mathcal{B}$. $\square$

Theorem 25 (**Containment for** $\{\exists, \forall, \wedge, \vee\}$-FO). *Let $\mathcal{A}$ and $\mathcal{B}$ be two structures. The following are equivalent.*

(i) *For every sentence φ in $\{\exists, \forall, \wedge, \vee\}$-FO, if $\mathcal{A} \models \varphi$ then $\mathcal{B} \models \varphi$.*

(ii) *There exists a surjective hyper-morphism from $\mathcal{A}$ to $\mathcal{B}$.*

(iii) $\mathcal{B} \models \theta_{\mathcal{A},|B|}^{\{\exists, \forall, \wedge, \vee\}\text{-FO}}$.

Proof. By construction $\mathcal{A} \models \theta_{\mathcal{A},|B|}^{\{\exists, \forall, \wedge, \vee\}\text{-FO}}$, so (i) implies (iii). By Lemma 22, (ii) implies (i). By Lemma 24, (iii) implies (i). $\square$

The Notion of a $\{\exists, \forall, \wedge, \vee\}$-FO-**core**

The property of a (classical) core can be rephrased in the logical context as the minimal $X = \widetilde{A} \subseteq A$ such that a primitive positive sentence φ is true on $\mathcal{A}$ iff it is true on $\mathcal{A}$ with the (existential) quantifiers relativised to $X = \widetilde{A}$. Let us say in this case that $\mathcal{A}$ has *X-relativisation* with respect to $\{\exists, \wedge\}$-FO.

Thus, the notion of a core can be recast in the context of $\{\exists, \wedge\}$-FO in a number of equivalent ways, as a minimal induced substructure $\widetilde{\mathcal{A}}$ of $\mathcal{A}$,

  (i) that satisfies the same $\{\exists, \wedge\}$-FO sentences;

 (ii) that is induced by minimal $X \subseteq A$ such that $\mathcal{A}$ has X-relativisation w.r.t. $\{\exists, \wedge\}$-FO; or,

(iii) that is induced by minimal $X \subseteq A$ such that $\mathcal{A}$ has an endomorphism with image X.

We are looking for a useful characterisation of the analogous concept of core for $\{\exists, \forall, \wedge, \vee\}$-FO. As we now have both quantifiers, two sets U and X, one for each quantifier, will emerge naturally, hence we will call a $\{\exists, \forall, \wedge, \vee\}$-FO-core, a *U-X*-core. As we shall see shortly, there are two equivalent ways of defining a *U-X*-core – one is logical, the other algebraic – as a minimal substructure $\widetilde{\mathcal{A}}$ of $\mathcal{A}$, induced by minimal $U, X \subseteq A$ such that:

 (ii) $\mathcal{A}$ has $\forall U$-$\exists X$-*relativisation* w.r.t. $\{\exists, \forall, \wedge, \vee\}$-FO; or,

(iii) $\mathcal{A}$ has a U-surjective X-total hyper-endomorphism.

Recall that a surjective hyper-endomorphism f of $\mathcal{A}$ is U-*surjective* if $f(U) = A$ and X-*total* if $f^{-1}(X) = A$.

We will show that the sets U and X are unique up to isomorphism and that within a minimal induced substructure $\widetilde{\mathcal{A}}$, the sets U and X are uniquely determined. This will reconcile our definition of a *U-X*-core with the following natural definition, in which U and X are not explicit:

  (i) as a minimal induced substructure $\widetilde{\mathcal{A}}$ of $\mathcal{A}$ that satisfies the same sentences of $\{\exists, \forall, \wedge, \vee\}$-FO.

In our definition of $\{\exists, \forall, \wedge, \vee\}$-FO-core, we ask for a minimal structure, i.e. not necessarily an induced substructure. We shall see that it is equivalent to the above.

Relativisation

Given a formula φ, we denote by $\varphi_{[\forall u / \forall u \in U, \exists x / \exists x \in X]}$ the formula obtained from φ by relativising simultaneously every universal quantifier to U and every existential quantifier to X. When we only relativise universal quantifiers to U, we write $\varphi_{[\forall u / \forall u \in U]}$, and when we only relativise existential quantifiers to X, we write $\varphi_{[\exists x / \exists x \in X]}$.

Definition 26. *Let $\mathcal{A}$ be a finite structure over a set A, and U, X be two subsets of A. We say that $\mathcal{A}$ has $\forall U$-$\exists X$-relativisation if, for all sentences φ in $\{\exists, \forall, \wedge, \vee\}$-FO the following are equivalent*

  (i) $\mathcal{A} \models \varphi$

(ii) $\mathcal{A} \models \varphi_{[\forall u / \forall u \in U]}$

(iii) $\mathcal{A} \models \varphi_{[\exists x / \exists x \in X]}$

(iv) $\mathcal{A} \models \varphi_{[\forall u / \forall u \in U, \exists x / \exists x \in X]}$

Lemma 27. *Let $\mathcal{A}$ be a finite structure over a set A, and U, X be two subsets of A. If $\mathcal{A}$ has a U-surjective X-total hyper-endomorphism then $\mathcal{A}$ has $\forall U$-$\exists X$-relativisation.*

Proof. Note that in Definition 26, we have $(iii) \Rightarrow (i) \Rightarrow (ii)$ and $(iii) \Rightarrow (iv) \Rightarrow (ii)$ trivially. It suffices to prove that $(ii) \Rightarrow (i)$ and $(i) \Rightarrow (iii)$ to complete the proof. To do so, we will consider the well known Hintikka game corresponding to Case (i), called the *unrelativised game* hereafter; and, the relativised Hintikka games corresponding to the relativised formulae from Cases (ii), (iii) and (iv) (the relativised game considered being clear from context).

Let h be a U-surjective X-total surjective hyper-endomorphism of $\mathcal{D}$. The proof follows the line of that of Lemma 22.

$((ii) \Rightarrow (i))$. Assume that we have a winning strategy in the universally relativised game. We produce a winning strategy in the unrelativised game using h. When taking the antecedent of a universal variable, we make sure to pick an antecedent in U which we can do by U-surjectivity of h. To be more precise, the linear order over $\mathcal{A}$ used in the proof of Lemma 22 starts with the elements of U.

$((i) \Rightarrow (iii))$. Assume that we have a winning strategy in the unrelativised game. We produce a winning strategy in the existentially relativised game using h. When taking the image of an existential variable, we no longer pick an arbitrary element but one in X, which we can do by X-totality of h. $\square$

Proposition 28. *The following are equivalent.*

(i) $\mathcal{A}$ *has* $\forall U$-$\exists X$*-relativisation.*

(ii) $\overline{\mathcal{A}}$ *has* $\forall X$-$\exists U$*-relativisation.*

Proof. It suffices to prove one implication. We prove (ii) implies (i). Let φ be a sentence of $\{\exists, \forall, \wedge, \vee\}$-FO. We use the duality principle and prove that $\mathcal{A} \models \varphi \iff \mathcal{A} \models \varphi_{[\forall u / \forall u \in U]}$. The other cases are similar and are omitted.

We follow the same notation as in Proposition 11: ψ is the sentence logically equivalent to $\neg \varphi$ with negation pushed at the atomic level, and $\overline{\varphi}$ is the sentence obtained from ψ by replacing every occurrence of a negative atom $\neg R$ by R. Recall the following chain of equivalence.

$$\mathcal{A} \models \varphi \iff \mathcal{A} \models \neg(\neg\varphi) \iff \mathcal{A} \models \neg(\psi) \iff \mathcal{A} \not\models \psi \iff \overline{\mathcal{A}} \not\models \overline{\varphi}.$$

By assumption $\overline{\mathcal{A}} \not\models \overline{\varphi} \iff \overline{\mathcal{A}} \not\models \overline{\varphi}_{[\exists u / \exists u \in U]}$. Using the above chain of equivalence backward and propagating the relativisation we obtain the following chain of equivalence.

$$\overline{\mathcal{A}} \not\models \overline{\varphi}_{[\exists u / \exists u \in U]}. \iff \mathcal{A} \not\models \psi_{[\exists u / \exists u \in U]} \iff \mathcal{A} \models \neg(\psi_{[\exists u / \exists u \in U]})$$
$$\iff \mathcal{A} \models \neg(\neg\varphi_{[\forall u / \forall u \in U]}) \iff \mathcal{A} \models \varphi_{[\forall u / \forall u \in U]}.$$

$\square$

Lemma 29. *Let $\mathcal{A}$ be a finite structure over a set A, and U, X be two subsets of A. If $\mathcal{A}$ has $\forall U$-$\exists X$-relativisation then $\mathcal{A}$ has a U-surjective X-total hyper-endomorphism.*

Proof. Using the fact that the identity (defined as $i(x) := \{x\}$ for every x in $\mathcal{A}$) is a surjective hyper-endomorphism of $\mathcal{A}$ and applying Theorem 25, we derive that $\mathcal{A} \models \theta_{\mathcal{A}, |A|}^{\{\exists, \forall, \wedge, \vee\}\text{-FO}}$. By assumption, we may equivalently relativise only the existential quantifiers to X (Definition 26 $(i) \Rightarrow (iii)$) and $\mathcal{A} \models \theta_{\mathcal{A}, |A|[\exists x / \exists x \in X]}^{\{\exists, \forall, \wedge, \vee\}\text{-FO}}$. Proceeding as in the proof of Lemma 22 but over this relativised sentence, we derive the existence of an X-total surjective hyper-operation g.

Using Proposition 28 and working over $\overline{\mathcal{A}}$, we derive similarly that $\overline{\mathcal{A}}$ has a U-total surjective hyper-operation. Let f be the inverse of this hyper-operation. Observe that it is a U-surjective hyper-operation.

By Lemma 19, the composition of these operations $g \circ f$ is a X-total U-surjective hyper-endomorphism as required. $\square$

Together, the two previous lemmata establish an algebraic characterisation of relativisation.

Theorem 30. *Let $\mathcal{A}$ be a finite structure over a set A, and U, X be two subsets of A. The following are equivalent.*

(i) *The structure $\mathcal{A}$ has $\forall U$-$\exists X$-relativisation.*

(ii) *The structure $\mathcal{A}$ has a X-total U-surjective hyper-endomorphism.*

Corollary 31. *Let $\mathcal{A}$ be a finite structure that has a U-surjective X-total hyper-endomorphism. Let $\widetilde{\mathcal{A}}$ be the substructure of $\mathcal{A}$ induced by $U \cup X$. The following holds.*

(i) *$\mathcal{A}$ and $\widetilde{\mathcal{A}}$ are $\{\exists, \forall, \wedge, \vee\}$-FO-equivalent.*

(ii) *$\widetilde{\mathcal{A}}$ has $\forall U$-$\exists X$-relativisation.*

Proof. Let f be the U-surjective X-total hyper-endomorphism of $\mathcal{A}$. Its range restriction g to $\widetilde{A} = U \cup X$ is a surjective hyper-morphism from $\mathcal{A}$ to $\widetilde{\mathcal{A}}$. The inverse g^{-1} of g is a surjective hyper-morphism from $\widetilde{\mathcal{A}}$ to $\mathcal{A}$, by X-totality of f. Appealing to Lemma 22 twice, once with g and once with g^{-1}, we obtain (i).

The restriction of g to $\widetilde{A}$ is a U-surjective X-total hyper-endomorphism of $\widetilde{\mathcal{A}}$, and (ii) follows from Lemma 27. $\square$

The U-X Core

Given a structure $\mathcal{D}$, we consider all minimal subsets X of D such that there is an X-total surjective hyper-endomorphism g of $\mathcal{D}$, and all minimal subsets U such that there is a U-surjective hyper-endomorphism f of $\mathcal{D}$. Such sets always exist as totality and surjectivity of surjective hyper-endomorphisms mean that in the worst case we may choose $U = X = D$. Since $g \circ f$ is a X-total U-surjective hyper-endomorphisms of $\mathcal{D}$ by Lemma 19, we may furthermore require that among all

minimal sets satisfying the above, we choose a set U and a set X with $U \cap X$ maximal. Let $\widetilde{\mathcal{D}}$ be the substructure of $\mathcal{D}$ induced by $U \cup X$. We call $\widetilde{\mathcal{D}}$ a U-X-core of $\mathcal{D}$.

Remark 32. Assume that there is an X_1-shop h_1 and an X_2-shop h_2 that preserves $\mathcal{D}$ such that $|X_1| > |X_2|$. We consider images of $h_1 \circ h_2$. For each element in X_2, pick a single element x_1' of X_1 in $h_1(X_2) \cap X_1$ such that $x_1' \in h_1(x_2)$. Let X_1' denote the set of picked elements. Since $|X_1| > |X_2|$ then $h_1 \circ h_2$ is an X_1'-shop that preserves $\mathcal{D}$ with $|X_1'| \leq |X_2|$. Diagrammatically, this can be written as,

$$D \xrightarrow{h_2} X_2 \xrightarrow{h_1} X_1' \subseteq h_1(X_2) \cap X_1 \subseteq X_1 \subseteq h_1 \circ h_2(D).$$

This means that we may look for an X-shop where the set X is minimal with respect to inclusion, or equivalently, for a set with minimal size $|X|$. So, in order to find an X-shop with a minimal set $|X|$, we may proceed greedily, removing elements from D while we have an X-shop until we obtain a set X such that there is no X'-shop for $X' \subsetneq X$. The dual argument applies to U-shops, and consequently to U-X-shops.

This further explains why minimising U and X, and then maximising their intersection, necessarily leads to a minimal $\widetilde{D} := U \cup X$ also. Because, would we find $U' \cup X'$ of smaller size, we might look within U' and X' for potentially smaller sets of cardinality $|U|$ and $|X|$, thus contradicting minimality.

Note that the sets U and X are not necessarily unique. However, as we shall see later *the U-X-core is unique up to isomorphism* (see Theorem 39). Moreover, within $\widetilde{\mathcal{D}}$, the sets U and X are uniquely determined. We delay until later the proof of this second result (see Theorem 75).

Uniqueness of the U-X-core

Throughout this section, let $\mathcal{D}$ be a finite structure and $\mathcal{M}$ its associated DSM; i.e. $\mathcal{M}$ is the set of surjective hyper-endomorphisms of $\mathcal{D}$. Let U and X be subsets of D such that the substructure $\widetilde{D}$ of $\mathcal{D}$ induced by $\widetilde{D} = U \cup X$ is a U-X-core of $\mathcal{D}$. We will progress through various lemmata and eventually derive the existence of a canonical U-X-shop in $\mathcal{M}$ which will be used to prove that the U-X-core is unique up to isomorphism. Uniqueness of the U-X-core has no real bearing on our classification program but the canonical shop will allow us to characterise all other shops in $\mathcal{M}$, which will be instrumental in the hardness proofs for $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{D})$.

Lemma 33. *Let f be a shop in $\mathcal{M}$. For any element z in D, $f(z)$ contains at most one element of the set U, that is $|f(z) \cap U| \leq 1$.*

Proof. Assume for contradiction that there is some z and some distinct elements u_1 and u_2 of U such that $f(z) \supseteq \{u_1, u_2\}$. Let $z_3, z_4, \ldots$ be any choice of antecedents under f of the remaining elements $u_3, u_4, \ldots$ of U (recall that f is surjective). By assumption the monoid $\mathcal{M}$ contains a U-shop g. Hence, $g \circ f$ would be a U'-shop with $U' = \{z, z_3, z_4, \ldots\}$ since $f(U') \subseteq U$ and $g(U) = D$. We get a contradiction as $|U'| < |U|$. $\qquad\square$

Lemma 34. *Let f be a U-shop in $\mathcal{M}$. There exists a permutation α of U such that: for any u in U,*

   (i) $f(u) \cap U = \{\alpha(u)\}$; *and,*

   (ii) $f^{-1}(u) \cap U = \{\alpha^{-1}(u)\}$.

Proof. It follows from Lemma 33 that for any u in U, $|f(u) \cap U| \leq 1$. Since f is a U-shop, every element in D has an antecedent in U under f and thus in particular for any u in U, $|f^{-1}(u) \cap U| \geq 1$. Note that if some element of U had no image in U then as U is finite, we would have an element of U with two distinct images in U. Hence, for any u in U, $|f(u) \cap U| = 1$ and the result follows. $\square$

The dual statements concerning X-shops hold.

Lemma 35. *Let f be a shop in $\mathcal{M}$. for any element z in D, $f^{-1}(z)$ contains at most one element of the set X, that is $|f^{-1}(z) \cap X| \leq 1$.*

Proof. By duality from Lemma 33. $\square$

Lemma 36. *Let f be an X-shop in $\mathcal{M}$. There exists a permutation β of X such that: for any x in X,*

   (i) $f(x) \cap X = \{\beta(x)\}$; *and,*

   (ii) $f^{-1}(x) \cap X = \{\beta^{-1}(x)\}$.

Proof. By duality from Lemma 34. $\square$

Lemma 37. *Let f be a shop in $\mathcal{M}$. If f is a U-X-shop then $f(X) \cap (U \setminus X) = \emptyset$.*

Proof. Assume for contradiction that for some $x_1 \in X$ and some $u_1 \in U \setminus X$, we have $u_1 \in f(x_1)$. Since f is an X-shop, every element is an antecedent under f of some element in X, in particular every element $x_2, x_3, \ldots \in X$ (different from x_1) has a unique image $x_2', x_3', \ldots \in X$ (see Lemma 36). Some element of X, say x_i does not occur in these images. Necessarily, x_1 reaches x_i. Note that x_i can not also belong to U as otherwise, x_i and u_1, two distinct elements of U, would be reached by x_1, contradicting Lemma 33. Thus, we must have that x_i belongs to $X \setminus U$. Let $U' := U$ and $X' := X \setminus \{x_i\} \cup \{u_1\}$. Note that $f^2 := f \circ f$, the second iterate of f, is a U'-X'-shop with $|U'| = |U|$, $|X'| = |X|$ and $|U' \cap X'| < |U \cap X|$. This contradicts our hypothesis on U and X. $\square$

Proposition 38 (**canonical U-X-shop**). *Let $\mathcal{M}$ be a DSM over a set D and U and X be minimal subsets of D such that: there is a U-shop in $\mathcal{M}$; there is an X-shop in $\mathcal{M}$ and $U \cup X$ is minimal. Then, there is a U-X-shop h in $\mathcal{M}$ that has the following properties:*

   (i) *for any y in $U \cap X$, $h(y) \cap (U \cup X) = \{y\}$;*

   (ii) *for any x in $X \setminus U$, $h(x) \cap (U \cup X) = \{x\}$;*

   (iii) *for any u in $U \setminus X$, $h(u) \cap (U \cup X) = \{u\} \cup X_u$, where $X_u \subseteq X \setminus U$; and,*

(iv) $h(U \setminus X) \cap X = \bigcup_{u \in U \setminus X} X_u = X \setminus U$.

Proof. By assumption, $\mathcal{M}$ contains a U-X-shop f. Let α and β be permutations of U and X, respectively, as in Lemmata 34 and 36. Let r be the least common multiple of the order of the permutations β and α. We set h to be the rth iterate of f and we now know that $h(z) \ni z$ for any element z and that h is a U-X-shop by 19. Let y in $U \cap X$, we know that $h(y) \ni y$. We can not have another element from $U \cup X$ in $h(y)$ by Lemmata 33 and 36. This proves (i). Let x in $X \setminus U$, we know that $h(x) \ni x$. We can not have an element from X distinct from x in $h(x)$ by Lemma 36 and we can not have an element from $U \setminus X$ in $h(x)$ by Lemma 37. This proves (ii). Let u in $U \setminus X$, we know that $h(u) \ni u$. We can not have an element from U distinct from u in $h(u)$ by Lemma 34. We may have however some elements from $X \setminus U$ in $h(u)$. Thus, there is a set $\emptyset \subseteq X_u \subseteq X \setminus U$ such that $h(u) \cap (U \cup X) = \{u\} \cup X_u$. This proves (iii). By construction h is a U-shop and every element must have an antecedent in U under h. Since by the first three points, elements from $X \setminus U$ can only be reached from elements in $U \setminus X$, the last point (iv) follows. $\square$

Theorem 39. *The U-X-core is unique up to isomorphism.*

Proof. Let h_1 be a U_1-X_1-shop with minimal $|U_1|$, $|X_1|$ and $|U_1 \cup X_1|$ and let h_2 be a U_2-X_2-shop with minimal $|U_2|$, $|X_2|$ and $|U_2 \cup X_2|$. Hence, $h_1 \circ h_2$ is a $h_1(X_2) \cap X_1$-shop with $|h_1(X_2)| \leq |X_1|$. By minimality of X_1, $|h_1(X_2)| = |X_1|$, and the restriction of h_1 to domain X_2 and codomain X_1 induces a surjective homomorphism from the substructure induced by X_2 to the substructure induced by X_1. Similarly h_2 induces a surjective homomorphism in the other direction. As we work with finite structures, h_1 induces an isomorphism i from the substructure induced by X_1 to the substructure induced by X_2. By duality, we also get that h_1 induces an isomorphism i' from the substructure induced by U_1 to the substructure induced by U_2. By construction, i and i' agree on $U_1 \cap X_1$ (necessarily to $U_2 \cap X_2$) and the result follows. $\square$

3.6 Positive Horn ($\{\exists, \forall, \wedge\}$-FO)

In primitive positive and positive Horn logic, one normally considers equalities to be permitted. From the perspective of computational complexity of CSP and QCSP, this distinction is unimportant as equalities may be propagated out by substitution. In the case of positive Horn and QCSP, though, equality does allow the distinction of a trivial case that can not be recognised without it. The sentence $\exists x \forall y\ x = y$ is true exactly on structures of size one. The structures $\mathcal{K}_1$ and $2\mathcal{K}_1$, containing empty relations over one element and two elements, respectively, are therefore distinguishable in $\{\exists, \forall, \wedge, =\}$-FO, but not in $\{\exists, \forall, \wedge\}$-FO. So many results from this section apply only for *non-trivial* structures of size ≥ 2.

For $\{\exists, \forall, \wedge\}$-FO, the correct concept to transfer winning strategies is that of *surjective homomorphism from a power*. Recall first that the *product* $\mathcal{A} \times \mathcal{B}$ of two

structures $\mathcal{A}$ and $\mathcal{B}$ has domain $\{(x,y) : x \in A, y \in B\}$ and for a relation symbol R, $R^{\mathcal{A} \times \mathcal{B}} := \{((a_1, b_1), \ldots, (a_r, b_r)) : (a_1, \ldots, a_r) \in R^{\mathcal{A}}, (b_1, \ldots, b_r) \in R^{\mathcal{B}}\}$; and, similarly for a constant symbol c, $c^{\mathcal{A} \times \mathcal{B}} := (c^{\mathcal{A}}, c^{\mathcal{B}})$. The *mth power* $\mathcal{A}^m$ of $\mathcal{A}$ is $\mathcal{A} \times \ldots \times \mathcal{A}$ (m times).

Lemma 40 (**strategy transfer for** $\{\exists, \forall, \wedge\}$-FO)**.** *Let $\mathcal{A}$ and $\mathcal{B}$ be two structures and $m \geq 1$ such that there is a surjective homomorphism from $\mathcal{A}^m$ to $\mathcal{B}$. Then, for every sentence φ in $\{\exists, \forall, \wedge\}$-FO, if $\mathcal{A} \models \varphi$ then $\mathcal{B} \models \varphi$.*

Proof. For $m = 1$, the proof is similar to Lemma 22. A projection from $\mathcal{A}^m$ to $\mathcal{A}$ is a surjective homomorphism. This means that for every sentence φ in $\{\exists, \forall, \wedge\}$-FO, if $\mathcal{A}^m \models \varphi$ then $\mathcal{A} \models \varphi$. For the converse, one can consider the "product strategy" which consists in projecting over each coordinate of $\mathcal{A}^m$ and applying the strategy for $\mathcal{A}$. This means that $\mathcal{A}$ and $\mathcal{A}^m$ are $\{\exists, \forall, \wedge\}$-FO-equivalent. For further details see [24, Lemma 1&2]. $\square$

Example 41. Consider an undirected bipartite graph with at least one edge $\mathcal{G}$ and $\mathcal{K}_2$ the graph that consists of a single edge. There is a surjective homomorphism from $\mathcal{G}$ to $\mathcal{K}_2$. Note also that $\mathcal{K}_2 \times \mathcal{K}_2 = \mathcal{K}_2 + \mathcal{K}_2$ (where $+$ stands for disjoint union) which we write as $2\mathcal{K}_2$. Thus, $\mathcal{K}_2^j = 2^{j-1}\mathcal{K}_2$ (as $\times$ distributes over $+$). Hence, if $\mathcal{G}$ has no isolated element and m edges there is a surjective homomorphism from $\mathcal{K}_2^{1+\log_2 m}$ to $\mathcal{G}$.

This examples provides a lower bound for m which we can improve.

Proposition 42 (**lower bound**)**.** *For any $m \geq 2$, there are structures $\mathcal{A}$ and $\mathcal{B}$ with $|A| = m$ and $|B| = m + 1$ such that there is only a surjective homomorphism from $\mathcal{A}^j$ to $\mathcal{B}$ provided that $j \geq |A|$.*
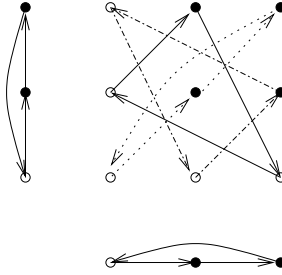
Figure 3.1: the power of oriented cycles is a sum of oriented cycles (the various arcs are all of the same type, they are drawn in different guise to highlight the different oriented cycles).

sketch. We consider a signature that consists of a binary symbol E together with a monadic predicate R. Consider for $\mathcal{A}$ an oriented cycle with m vertices, for which R holds for all but one. Consider for $\mathcal{B}$ an oriented cycle with m vertices, for which R does not hold, together with a self-loop on which R holds. The square of $\mathcal{A}$ will

consists of $|A| = m$ oriented cycles with m vertices: one cycle will be a copy of $\mathcal{A}$, all the other will be similar but with two vertices on which R does not hold (this is depicted on Figure 3.1 in the case $m = 3$: white vertices do not satisfy R while black ones do). It is only for $j = m$ that we will get as an induced substructure of $\mathcal{A}^j$ one copy of an oriented cycle on which R does not hold as in $\mathcal{B}$. □

There is also a *canonical* $\{\exists, \forall, \wedge\}$-FO-*sentence* which turns out to be in Π_2-form, that is with a quantifier prefix of the form $\forall^\star \exists^\star$. We consider temporarily structures with m constants $c_1, c_2, \ldots, c_m$; let $\mathbf{t}$ in A^m describe the position of these constants in a structure $\mathcal{A}$; and, write $\mathcal{A}_\mathbf{t}$ for the corresponding structure with constants. We consider the canonical conjunctive query of the structure with constants $\bigotimes_{\mathbf{t} \in A^m} \mathcal{A}_\mathbf{t}$, (where $\bigotimes$ denote the product), identifying the constants with some variables $\mathbf{w} = w_1, \ldots, w_m$ and using variables $\mathbf{v}$ for the other elements. We turn this quantifier-free formula into a sentence by adding the prefix $\forall \mathbf{w} \exists \mathbf{v}$. Keeping this in mind, we can also give the following direct definition, but it dilutes the intuition somewhat.

Definition 43 (**Canonical** $\{\exists, \forall, \wedge\}$-FO **sentence**). *Let $\mathcal{A}$ be a structure and $m > 0$. Let $\mathbf{r}$ be an enumeration of the elements of $\tilde{A} := \mathcal{A}^{|A|^m}$.*

$$\theta_{\mathcal{A},m}^{\{\exists,\forall,\wedge\}\text{-FO}} := \forall \mathbf{w} \exists \mathbf{v} \varphi_{\tilde{\mathcal{A}}(\mathbf{r})}(\mathbf{v}) \wedge \bigwedge_{\mathbf{t} \in A^m} w_1 = v_{\mathbf{t}, \mathbf{t}[1]} \ldots \wedge w_m = v_{\mathbf{t}, \mathbf{t}[m]}.$$

Observe that we may propagate the equalities out of $\theta_{\mathcal{A},m}^{\{\exists,\forall,\wedge\}\text{-FO}}$ to obtain an equivalent sentence: e.g. we remove $w_1 = v_{\mathbf{t}, \mathbf{t}[1]}$ and replace every occurrence of $v_{\mathbf{t}, \mathbf{t}[1]}$ by w_1. Observe also that $\mathcal{A} \models \theta_{\mathcal{A},m}^{\{\exists,\forall,\wedge\}\text{-FO}}$. Indeed, assume that $\mathbf{t} \in A^m$ is the assignment chosen for the universal variables $\mathbf{w}$. There is a natural projection from $\bigotimes_{\mathbf{t} \in A^m} \mathcal{A}_\mathbf{t}$ to $\mathcal{A}_\mathbf{t}$ which is a homomorphism. This homomorphism corresponds precisely to a winning strategy for the existential variables $\mathbf{v}$.

Theorem 44 (**Containment for** $\{\exists, \forall, \wedge\}$-FO [24]). *Let $\mathcal{A}$ and $\mathcal{B}$ be two non-trivial structures. The following are equivalent.*

  (i) *for every sentence φ in $\{\exists, \forall, \wedge\}$-FO, if $\mathcal{A} \models \varphi$ then $\mathcal{B} \models \varphi$.*

 (ii) *There exists a surjective homomorphism from $\mathcal{A}^r$ to $\mathcal{B}$, with $r \leq |A|^{|B|}$.*

(iii) $\mathcal{B} \models \theta_{\mathcal{A},|B|}^{\{\exists,\forall,\wedge\}\text{-FO}}$

  where $\theta_{\mathcal{A},|\mathcal{B}|}^{\{\exists,\forall,\wedge\}\text{-FO}}$ is a canonical sentence of $\{\exists, \forall, \wedge\}$-FO with quantifier prefix $\forall^{|\mathcal{B}|} \exists^\star$ that is defined in terms of $\mathcal{A}$ and modelled by $\mathcal{A}$ by construction.

sketch. (ii) implies (i) by Lemma 40. (i) implies (iii) since $\mathcal{A}$ models $\psi_{\mathcal{A},|\mathcal{B}|}$. (iii) implies (ii) by construction of $\psi_{\mathcal{A},|\mathcal{B}|}$. We may chose for the universal variables $\mathbf{w}$ an enumeration of $\mathcal{B}$. The winning strategy on $\mathcal{B}$ induces a surjective homomorphism from $\mathcal{A}^r$ (for further details see [24, Theorem 3] and comments on the following page). □

Following our approach for the other logics, we now define a minimal representative as follows.

Definition 45. *A* Q*-core* $\mathcal{B}$ *of a structure* $\mathcal{A}$ *is a minimal substructure of* $\mathcal{A}$ *such that for every sentence* φ *in* $\{\exists, \forall, \wedge\}$*-FO,* $\mathcal{A} \models \varphi$ *if and only if* $\mathcal{B} \models \varphi$.

(a) The graph $\mathcal{A}_2$. (b) The graph $\mathcal{A}_3$.

(c) The graph $\mathcal{A}_2 \times \mathcal{A}_2$. (d) A homomorphism to $\mathcal{A}_3$.
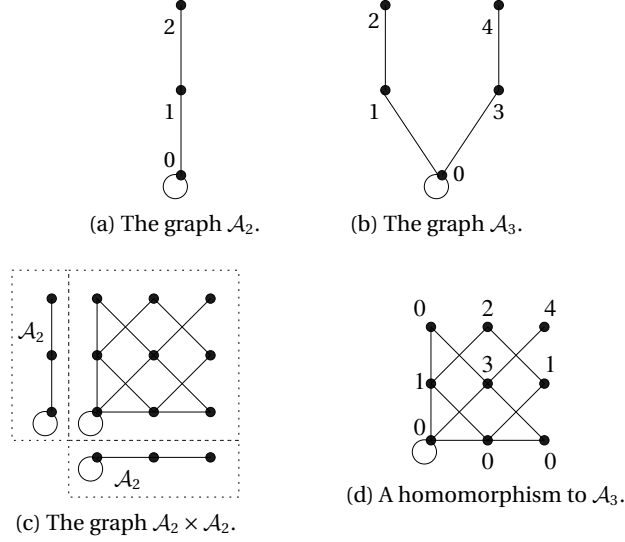
Figure 3.2: surjective homomorphism from a power.

Example 46. We consider the graph $\mathcal{A}_3$ depicted on Figure 3.2b and its subgraph $\mathcal{A}_2$ which consists of a loop attached to a path of length 2. The map $f(0) := 0$, $f(1) := 1$, $f(2) := 2$, $f(3) := 0$, $f(4) := 0$ is a surjective homomorphism from $\mathcal{A}_3$ to $\mathcal{A}_2$. The square of $\mathcal{A}_2$ is depicted on Figure 3.2c; and, a surjective homomorphism from it to $\mathcal{A}_3$ is depicted on Figure 3.2d. Thus $\mathcal{A}_3$ and $\mathcal{A}_2$ are equivalent w.r.t. $\{\exists, \forall, \wedge\}$-FO. One can also check that $\mathcal{A}_2$ is minimal and is therefore a Q-core of $\mathcal{A}_3$.

The behaviour of the Q-core differs from its cousins the core and the U-X-core.

Proposition 47. *The Q-core of a* 3*-element structure* $\mathcal{A}$ *is not always an induced substructure of* $\mathcal{A}$.

Proof. Consider the signature $\sigma := \langle E, R, G \rangle$ involving a binary relation E and two unary relations R and G. Let $\mathcal{A}$ and $\mathcal{B}$ be structures with domain $\{1, 2, 3\}$ with the following relations.

$$E^{\mathcal{A}} := \{(1,1),(2,3),(3,2)\} \quad R^{\mathcal{A}} := \{1,2\} \quad G^{\mathcal{A}} := \{1,3\}$$
$$E^{\mathcal{B}} := \{(1,1),(2,3),(3,2)\} \quad R^{\mathcal{B}} := \{1\} \quad G^{\mathcal{B}} := \{1\}$$

Since $\mathcal{B}$ is a substructure of $\mathcal{A}$, we have $\mathcal{B} \longrightarrow\!\!\!\!\rightarrow \mathcal{A}$. Conversely, the square of $\mathcal{A}^2$ contains an edge that has no vertex in the relation R and G, which ensures that $\mathcal{A}^2 \longrightarrow\!\!\!\!\rightarrow \mathcal{B}$ (see Figure 3.3). Observe also that no two-element structure $\mathcal{C}$, and *a fortiori* no two-element substructure of $\mathcal{A}$ agrees with them on $\{\exists, \forall, \wedge\}$-FO. Indeed,
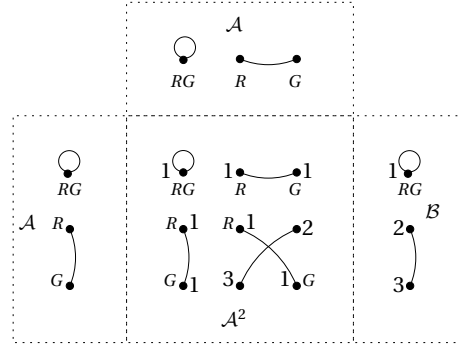
Figure 3.3: example of two distinct 3-element structures (signature, E binary and two unary predicates R and G) that are equivalent w.r.t. $\{\exists, \forall, \wedge\}$-FO.

if a structure $\mathcal{C}$ agrees on $\{\exists, \forall, \wedge\}$-FO with $\mathcal{B}$, it agrees also on $\{\exists, \wedge\}$-FO. Thus, the core of $\mathcal{B}$ is also the core of $\mathcal{C}$ and must appear as an induced substructure of $\mathcal{C}$. This core is the one-element substructure of $\mathcal{B}$ induced by 1. In order to have a surjective homomorphism from a power of $\mathcal{C}$ to $\mathcal{B}$, this power must contain a non-loop, and so does $\mathcal{C}$. This non-loop must in $\mathcal{C}$ be adjacent to another vertex (this is a $\{\exists, \forall, \wedge\}$-FO-expressible property that holds in $\mathcal{B}$ $\forall x \exists y \, E(x, y)$). The structure $\mathcal{C}$ would therefore be a two element structure satisfying

$$E^{\mathcal{C}} \subseteq \{(1,1),(1,2),(2,1)\} \quad R^{\mathcal{C}} \subseteq \{1\} \quad G^{\mathcal{C}} \subseteq \{1\}$$

A power of $\mathcal{C}$ would therefore be connected, which is not the case of $\mathcal{B}$, preventing the existence of any surjective homomorphism. $\qquad\square$

We do not know whether the Q-core of a structure is unique up to isomorphism. We also do not know whether it can be computed in a greedy fashion as the core can be (by finding increasingly smaller equivalent substructures). The Q-core behaves well in some cases: for Boolean structures, for unary structures (all relations are unary), and for partially reflexive forests. It is not impossible that one may compute the Qcore of a structure by computing incrementally smaller substructures that are equivalent via surjective homomomorphism from a square. The elusiveness of this concept of Q-core is discussed in some details in [72].

3.7 The case of $\{\exists, \vee\}$-FO

We do not really need to investigate coreness or equivalence for this fragment, which is rather silly, and of little relevance to our complexity classification program. We will see that the model checking problem of the larger fragment $\{\exists, \vee, =, \neq\}$-FO($\mathcal{D}$) is in L, for any structure $\mathcal{D}$. Its only interest here is that we can prove that the $\{\exists, \vee\}$-FO-core is not unique up to isomorphism.

Given a structure $\mathcal{A}$, for each relational symbol R of arity r and every equivalence relation ϵ over the index set of this relation $\{1, 2, \ldots, r\}$, we say that (R, ϵ) is *witnessed* in $\mathcal{A}$ if, and only if, there are some elements $a_1, a_2, \ldots, a_r$ of $\mathcal{A}$ that satisfy $R(a_1, a_2, \ldots, a_r)$ and agree with ϵ in the sense that $i \epsilon j$ implies $a_i = a_j$ for any indices $1 \le i, j \le r$.

Let $\bar{i}$ be the smallest index that is equivalent to i under ϵ. We associate naturally a sentence $\varphi_{R,\epsilon}$ of $\{\exists, \vee\}$-FO to R, ϵ as follows: this sentence has existential variables $x_1, x_2, \ldots, x_e$ where e is the number of equivalence classes of ϵ and a single atom $R(x_{\bar{1}}, x_{\bar{2}}, \ldots, x_{\bar{r}})$.

We can define a canonical sentence that witnesses $\{\exists, \vee\}$-FO-containment in $\mathcal{A}$, but it is not a sentence of $\{\exists, \vee\}$-FO, but a conjunction of sentences of $\{\exists, \vee\}$-FO.

$$\theta_{\mathcal{A}}^{\{\exists, \vee\}\text{-FO}} := \bigwedge_{\text{symbol } R} \bigwedge_{\mathcal{A} \text{ witnesses } (R, \epsilon)} \varphi_{R, \epsilon}.$$

Note that $\theta_{\mathcal{A}}^{\{\exists, \vee\}\text{-FO}}$ is modelled by $\mathcal{A}$ by construction.

Observe that if $\mathcal{A}$ witnesses (R, ϵ) then it witnesses (R, ϵ') for every ϵ' that is finer than ϵ. By extension, we will speak of coarsest witnessed (R, ϵ) when ϵ is a coarsest equivalence relation among those pairs (R, ϵ') that witnessed.

Proposition 48 (**Containment for** $\{\exists, \vee\}$**-FO**)**.** *Let $\mathcal{A}$ and $\mathcal{B}$ be two structures. The following are equivalent.*

(i) *for every sentence φ in $\{\exists, \vee\}$-FO, if $\mathcal{A} \models \varphi$ then $\mathcal{B} \models \varphi$.*

(ii) *Every (R, ϵ) that is witnessed in $\mathcal{A}$ is also witnessed in $\mathcal{B}$.*

(iii) *The coarsest (R, ϵ) that are witnessed in $\mathcal{A}$ are also witnessed in $\mathcal{B}$.*

(iv) *For every (R, ϵ) that is witnessed in $\mathcal{A}$, $\mathcal{B} \models \varphi_{R, \epsilon}$.*

(v) *$\mathcal{B} \models \theta_{\mathcal{A}}^{\{\exists, \vee\}\text{-FO}}$, where $\theta_{\mathcal{A}}^{\{\exists, \vee\}\text{-FO}}$ is a conjunction of sentences of $\{\exists, \vee\}$-FO modelled by $\mathcal{A}$ by construction.*

Proof. This is essentially trivial and follows directly from our definitions. $\square$

Corollary 49. *Let $\mathcal{A}$ and $\mathcal{B}$ be two structures. The following are equivalent.*

(i) *for every sentence φ in $\{\exists, \vee\}$-FO, if $\mathcal{A} \models \varphi$ then $\mathcal{B} \models \varphi$.*

(ii) *The coarsest (R, ϵ) that are witnessed in both structures coincide.*

Proposition 50. *The following holds for any signature that contains one relational symbol of arity 3 or more.*

(i) *The strong $\{\exists, \vee\}$-FO-core is not unique up to isomorphism.*

(ii) *The $\{\exists, \vee\}$-FO-core is not unique up to isomorphism.*

(iii) *The notions of $\{\exists, \vee\}$-FO-core and strong $\{\exists, \vee\}$-FO-core do not coincide in general.*

Proof. W.l.o.g. we assume a symbol R of arity 3 and consider structures with 2 elements 1 and 2: the structure $\mathcal{A}$ such that $R(1,1,2)$ and $R(1,2,2)$ hold and no other tuples; and, the structure $\mathcal{B}$ such that $R(1,1,2)$ and $R(2,1,1)$ hold and no other tuples.

These structures are not isomorphic but they are {∃, ∨}-FO-equivalent as they have the same canonical sentence:

$$\exists x, \exists y\, R(x,x,y) \wedge \exists x, \exists y\, R(x,y,y).$$

No strictly smaller (w.r.t. inclusion) structure is {∃, ∨}-FO-equivalent to $\mathcal{A}$ and $\mathcal{B}$ so both are strong {∃, ∨}-FO-core and claim (i) follows.

The disjoint union $\mathcal{C}$ of $\mathcal{A}$ and $\mathcal{B}$ has both $\mathcal{A}$ and $\mathcal{B}$ as {∃, ∨}-FO-cores (also as strong {∃, ∨}-FO-cores) and claim (ii) follows.

The four element structure $\mathcal{D}$ such that $R(1,1,2)$ and $R(3,4,4)$ holds is its own {∃, ∨}-FO-core (any proper substructure would not be {∃, ∨}-FO-equivalent). However, the structure $\mathcal{D}$ is not a strong {∃, ∨}-FO-core: its strong {∃, ∨}-FO-core have two elements and are $\mathcal{A}$ and $\mathcal{B}$ (recall that strong {∃, ∨}-FO-core are equivalent structures that are minimal both w.r.t. size and inclusion). This proves our last claim. □

4. Complexity classification for most fragments

In this chapter we will classify the complexity of the model checking problem for most fragments of FO, namely those from the first, second and third classes we introduced in Chapter 2. These fragments do not exhibit a very rich complexity classification except for two fragments, $\{\exists, \wedge, \neq\}$-FO and $\{\exists, \forall, \wedge, \neq\}$-FO, which exhibits non trivial classifications in the Boolean case, but whose classification is a simple corollary of Schaefer's classification for Boolean CSP and QCSP.

4.1 Boolean CSP and QCSP

We recall below some well known results concerning the complexity of Boolean CSP and QCSP which we will need later. For definitions and further details regarding the proof, the reader may consult the nice survey by Chen [20].

Theorem 51 ([97])**.** *Let $\mathcal{D}$ be a Boolean structure. Then CSP($\mathcal{D}$) (equivalently, $\{\exists, \wedge\}$-FO($\mathcal{D}$)) is in* P *if, and only if, all relations of $\mathcal{D}$ are simultaneously 0-valid, 1-valid, Horn, dual-Horn, bijunctive or affine, and otherwise it is* NP*-complete.*

A similar result holds when universal quantifiers are added to the mix. (it was sketched in the presence of constants [97], then proved in the absence of constants in [31] and [35]).

Theorem 52. *Let $\mathcal{D}$ be a Boolean structure. Then QCSP($\mathcal{D}$) (equivalently, $\{\exists, \forall, \wedge\}$-FO($\mathcal{D}$)) is in* P *if, and only if, all relations of $\mathcal{D}$ are simultaneously Horn, dual-Horn, bijunctive or affine, and otherwise it is* Pspace*-complete.*

Example 53. The canonical example of a relation that does not fall in any of the tractable cases is NAE $:= \{0, 1\}^3 \setminus \{(0, 0, 0), (1, 1, 1)\}$. Let $\mathcal{B}_{\mathrm{NAE}}$ be the Boolean structure with this relation. It follows from the above theorems that CSP($\mathcal{B}_{\mathrm{NAE}}$) is NP-complete and that QCSP($\mathcal{B}_{\mathrm{NAE}}$) is Pspace-complete.

Example 54. For larger domains, though the classification remains open, the canonical hard problem is induced by the relation $\neq$. Let $\mathcal{K}_n$ denote the clique of size n (we view an undirected graph as a structure with a single binary predicate E that is symmetric). For $n \geq 3$, CSP($\mathcal{K}_n$) is a reformulation of the n-colourability problem and is NP-complete. It is also known that for $n \geq 3$ QCSP($\mathcal{K}_n$) is Pspace-complete [10].

4.2 First Class

Recall that this class consists of the following fragments:

- $\{\exists, \vee\}$-FO,
- $\{\exists, \vee, =\}$-FO, and
- $\{\exists, \vee, \neq\}$-FO.

All these fragments have a trivial model checking problem.

Proposition 55. (i) *When $\mathcal{D}$ has a single element, the model checking problem for FO is in* L.

(ii) *The model checking problem $\{\exists, \vee, \neq, =\}$-FO is in* L.

Proof. (i) In the case where $|D| = 1$, every relation is either empty or contains all tuples (one tuple), and the quantifiers $\exists$ and $\forall$ are semantically equivalent. Hence, the problem translates to the Boolean Sentence Value Problem (under the substitution of 0 and 1 for the empty and non-empty relations, respectively), known to be in L [63].

(ii) We may assume by the previous point that $|D| > 1$. We only need to check if one of the atoms that occurs as a disjunct in the input sentence holds in $\mathcal{D}$. Since $|D| > 1$, a sentence with an atom like $x = y$ or $x \neq y$ is always true in $\mathcal{D}$. For sentences of $\{\exists, \vee\}$-FO, the atoms may have implicit equality as in $R(x, x, y)$ for a ternary predicate R: in any case, each atom may be checked in constant time since $\mathcal{D}$ is a fixed structure, resulting in overall logspace complexity. $\square$

4.3 Second Class

Recall that this class consists of the following fragments.

- $\{\exists, \wedge, \vee\}$-FO, $\{\exists, \wedge, \vee, =\}$-FO
- $\{\exists, \wedge, \vee, \neq\}$-FO
- $\{\exists, \forall, \wedge, \vee, \neg\}$-FO
- $\{\exists, \forall, \wedge, \vee, \neq\}$-FO

We will see that all fragments of the second class follow a natural dichotomy.

Corollary 56. *For any syntactic fragment $\mathscr{L}$ of FO in the second class, the model checking problem $\mathscr{L}(\mathcal{D})$ is trivial (in L) when the $\mathscr{L}$-core of $\mathcal{D}$ has one element and hard otherwise (NP-complete for existential fragments, Pspace-complete for fragments containing both quantifiers).*

Proof. This is a direct consequence of Propositions 57, 58, 59 and 63 and of the characterisation of cores for these fragments from § 3, namely Proposition 21, Proposition 16 and Corollary 18. $\square$

Complexity of $\{\exists, \forall, \wedge, \vee, \neq\}$**-FO and** $\{\exists, \forall, \wedge, \vee, \neg\}$**-FO**

We start with the largest fragment from the second class, which turns out to exhibit trivial dichotomies.

Proposition 57. *In full generality, the class of problems* $\{\exists, \forall, \wedge, \vee, \neq\}$*-FO*$(\mathcal{D})$ *exhibits dichotomy: if* $|D| = 1$ *then the problem is in* L*, otherwise it is* Pspace*-complete. Consequently, the fragment extended with* $=$ *follows the same dichotomy.*

Proof. When $|D| \geq 2$, Pspace-hardness may be proved using no extensional relation of D other than $\neq$. The formula $\varphi_{\mathcal{K}_{|D|}}(x, y) := (x \neq y)$ simulates the edge relation of the clique $\mathcal{K}_{|D|}$ and the problem $\{\exists, \forall, \wedge\}$-FO$(\mathcal{K}_n)$ better known as QCSP$(\mathcal{K}_n)$ is Pspace-complete for $n \geq 3$ [10]. For n $= 2$, we use a reduction from the problem QCSP$(\mathcal{B}_{\mathrm{NAE}})$ to prove that $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{K}_2)$ is Pspace-complete. Let φ be an input for QCSP$(\mathcal{B}_{\mathrm{NAE}})$. Let φ' be built from φ by substituting all instances of NAE(x, y, z) by $E(x, y) \vee E(y, z) \vee E(x, z)$. It is easy to see that $\mathcal{B}_{\mathrm{NAE}} \models \varphi$ iff $\mathcal{K}_2 \models \varphi'$, and the result follows.

Note that we have not used $=$ in our hardness proof; and, in the case $|D| = 1$, we may allow $=$ without affecting tractability (triviality). Thus, the fragment extended with $=$ follows the same delineation. $\square$

The case of $\{\exists, \forall, \wedge, \vee, \neg\}$-FO is similar but we use $\sim$ instead of $=$ for hardness.

Proposition 58. *In full generality, the class of problems* $\{\exists, \forall, \wedge, \vee, \neg\}$*-FO*$(\mathcal{D})$ *exhibits dichotomy: if all relations of* $\mathcal{D}$ *are* trivial *(either empty or contain all tuples) then the problem is in* L*, otherwise it is* Pspace*-complete.*

Proof. If all relations are trivial then $\sim$ has a single equivalence class. Thus, $\mathcal{D}/\sim$ has a single element. By $\{\exists, \forall, \wedge, \vee, \neg\}$-FO-equivalence of $\mathcal{D}$ and $\mathcal{D}/\sim$ (see Proposition 21), it suffices to check whether an input φ in $\{\exists, \forall, \wedge, \vee, \neg\}$-FO holds in $\mathcal{D}/\sim$. Since the latter has a single element, the problem is in L.

Otherwise, the equivalence relation $\sim$ has at least 2 equivalence classes since $\mathcal{D}$ is non trivial. We may now follow the same proof as in Proposition 57, using $\sim$ in lieu of $=$, and Pspace-hardness follows. $\square$

Complexity of $\{\exists, \wedge, \vee\}$**-FO and** $\{\exists, \wedge, \vee, =\}$**-FO**

Proposition 59. *In full generality, the class of problems* $\{\exists, \wedge, \vee\}$*-FO*$(\mathcal{D})$ *exhibits dichotomy: if the core of* $\mathcal{D}$ *has one element then the problem is in* L*, otherwise it is* NP*-complete. As a corollary, the class of problems* $\{\exists, \wedge, \vee, =\}$*-FO*$(\mathcal{D})$ *exhibits the same dichotomy.*

In preliminary work by Martin on this [76, 75], the proof of the above is combinatorial and appeals to Hell and Nešetřil's dichotomy theorem for undirected graphs [52]. We will give here an algebraic proof which uses a variant of the Galois connection $\mathsf{Inv} - \mathsf{End}$ due to Krasner [57] for $\{\exists, \wedge, \vee, =\}$-FO. The variant for $\{\exists, \wedge, \vee\}$-FO involves hyper-endomorphisms rather than endomorphisms because of the absence of equality. An hyper-endomorphisms of $\mathcal{B}$ is a function from B to

the power-set of B that is total and preserving (see Definition 3.3). Our purpose is to provide both a self-contained proof and a gentle introduction to the techniques we shall use for the fragment $\{\exists, \forall, \wedge, \vee\}$-FO.

For a set F of hyper-endomorphisms on the finite domain B, let $\mathsf{Inv}(F)$ be the set of relations on B of which each f in F is an hyper-endomorphism (when these relations are viewed as a structure over B). We say that $S \in \mathsf{Inv}(F)$ is invariant or *preserved* by (the hyper-endomorphisms in) F. Let $\mathsf{hE}(\mathcal{B})$ be the set of hyper-endomorphisms of $\mathcal{B}$. Let $\langle\mathcal{B}\rangle_{\{\exists,\wedge,\vee\}\text{-FO}}$ be the set of relations that may be defined on $\mathcal{B}$ in $\{\exists, \wedge, \vee\}$-FO.

Lemma 60. *Let* $\mathbf{r} := (r_1, \ldots, r_k)$ *be a k-tuple of elements of $\mathcal{B}$. There exists a formula* $\theta_{\mathbf{r}}^{\{\exists,\wedge,\vee\}\text{-FO}}(u_1, \ldots, u_k) \in \{\exists, \wedge, \vee\}$-FO *such that the following are equivalent.*

(i) $(\mathcal{B}, r_1', \ldots, r_k') \models \theta_{\mathbf{r}}^{\{\exists,\wedge,\vee\}\text{-FO}}(u_1, \ldots, u_k)$.

(ii) *There is a hyper-endomorphism from* $(\mathcal{B}, r_1, \ldots, r_k)$ *to* $(\mathcal{B}, r_1', \ldots, r_k')$.

Proof. let $\mathbf{s} := (b_1, \ldots, b_{|B|})$ an enumeration of the elements of $\mathcal{B}$ and $\varphi_{\mathcal{B}(\mathbf{r}, \mathbf{s})}(v_1, \ldots, v_{|B|})$ be the associated conjunction of positive facts. Set

$$\theta_{\mathbf{r}}^{\{\exists,\wedge,\vee\}\text{-FO}}(u_1, \ldots, u_k) := \exists v_1, \ldots, v_{|B|} \ \varphi_{\mathcal{B}(\mathbf{r}, \mathbf{s})}(v_1, \ldots, v_{|B|}).$$

The forward direction is clear as the witness $s_1', \ldots, s_{|B|}'$ for $v_1, \ldots, v_{|B|}$ provides an hyper-endomorphism f defined as $f(b_i) \ni s_i'$ and $f(r_i) \ni r_i'$.

For the backwards direction, one may build an endomorphism from $(\mathcal{B}, r_1, \ldots, r_k)$ to $(\mathcal{B}, r_1', \ldots, r_k')$ from the given hyper-endomorphism. The result follows from the implication from (i) to (iv) of Proposition 15. $\qquad\square$

Theorem 61. *For a finite structure $\mathcal{B}$ we have* $\langle\mathcal{B}\rangle_{\{\exists,\wedge,\vee\}\text{-FO}} = \mathsf{Inv}(\mathsf{hE}(\mathcal{B}))$.

Proof. Let $\varphi(\mathbf{v})$ be a formula of $\{\exists, \wedge, \vee\}$-FO with free variables $\mathbf{v}$. We denote also by $\varphi(\mathbf{v})$ the relation induced over $\mathcal{B}$.

1. $\varphi(\mathbf{v}) \in \langle\mathcal{B}\rangle_{\{\exists,\wedge,\vee\}\text{-FO}} \Rightarrow \varphi(\mathbf{v}) \in \mathsf{Inv}(\mathsf{hE}(\mathcal{B}))$. This is proved by induction on the complexity of $\varphi(\mathbf{v})$.

   (Base Cases.) When $\varphi(\mathbf{v}) := R(\mathbf{v})$, the variables $\mathbf{v}$ may appear multiply in R and in any order. Thus R is an instance of an extensional relation under substitution and permutation of positions. The result follows directly from the definition of hyper-endomorphisms.

   (Inductive Step.) There are three subcases. We progress through them in a workmanlike fashion. Take $f \in \mathsf{hE}(\mathcal{B})$.

   a) $\varphi(\mathbf{v}) := \psi(\mathbf{v}) \wedge \psi'(\mathbf{v})$. Let $\mathbf{v} := (v_1, \ldots, v_l)$. Suppose $\mathcal{B} \models \varphi(x_1, \ldots, x_l)$; then both $\mathcal{B} \models \psi(x_1, \ldots, x_l)$ and $\mathcal{B} \models \psi'(x_1, \ldots, x_l)$. By Inductive Hypothesis (IH), for any $y_1 \in f(x_1), \ldots, y_l \in f(x_l)$, both $\mathcal{B} \models \psi(y_1, \ldots, y_l)$ and $\mathcal{B} \models \psi'(y_1, \ldots, y_l)$, whence $\mathcal{B} \models \varphi(y_1, \ldots, y_l)$.

   b) $\varphi(\mathbf{v}) := \psi(\mathbf{v}) \vee \psi'(\mathbf{v})$. Let $\mathbf{v} := (v_1, \ldots, v_l)$. Suppose $\mathcal{B} \models \varphi(x_1, \ldots, x_l)$; then one of $\mathcal{B} \models \psi(x_1, \ldots, x_l)$ or $\mathcal{B} \models \psi'(x_1, \ldots, x_l)$; w.l.o.g. the former. By IH, for any $y_1 \in f(x_1), \ldots, y_l \in f(x_l)$, $\mathcal{B} \models \psi(y_1, \ldots, y_l)$, whence $\mathcal{B} \models \varphi(y_1, \ldots, y_l)$.

53

c) $\varphi(\mathbf{v}) := \exists w \ \psi(\mathbf{v}, w)$. Let $\mathbf{v} := (v_1, \ldots, v_l)$. Suppose $\mathcal{B} \models \exists w \ \psi(x_1, \ldots, x_l, w)$; then for some x', $\mathcal{B} \models \psi(x_1, \ldots, x_l, x')$. By IH, for any $y_1 \in f(x_1), \ldots, y_l \in f(x_l), y' \in f(x')$, $\mathcal{B} \models \psi(y_1, \ldots, y_l, y')$, whereupon $\mathcal{B} \models \exists w \ \psi(y_1, \ldots, y_l, w)$.

2. $S \in \mathsf{Inv}(\mathsf{hE}(\mathcal{B})) \Rightarrow S \in \langle \mathcal{B} \rangle_{\{\exists, \wedge, \vee\}\text{-FO}}$. Consider the k-ary relation $S \in \mathsf{Inv}(\mathsf{hE}(\mathcal{B}))$. Let $\mathbf{r}_1, \ldots, \mathbf{r}_m$ be the tuples of S. Set $\theta_S^{\{\exists, \wedge, \vee\}\text{-FO}}(u_1, \ldots, u_k)$ to be the following formula of $\{\exists, \wedge, \vee\}$-FO:

$$\theta_{\mathbf{r}_1}^{\{\exists, \wedge, \vee\}\text{-FO}}(u_1, \ldots, u_k) \vee \ldots \vee \theta_{\mathbf{r}_m}^{\{\exists, \wedge, \vee\}\text{-FO}}(u_1, \ldots, u_k).$$

For $\mathbf{r}_i := (r_{i1}, \ldots, r_{ik})$, note that $(\mathcal{B}, r_{i1}, \ldots, r_{ik}) \models \theta_{\mathbf{r}_i}^{\{\exists, \wedge, \vee\}\text{-FO}}(u_1, \ldots, u_k)$. That $\theta_S(u_1, \ldots, u_k) = S$ now follows from Part (ii) of Lemma 60, since $S \in \mathsf{Inv}(\mathsf{hE}(\mathcal{B}))$.

$\square$

Corollary 62. *Let $\mathcal{B}$ and $\mathcal{B}'$ be finite structures over the same domain B. If $\mathsf{hE}(\mathcal{B}) \subseteq \mathsf{hE}(\mathcal{B}')$ then $\{\exists, \wedge, \vee\}$-FO$(\mathcal{B}') \leq_L \{\exists, \wedge, \vee\}$-FO$(\mathcal{B})$.*

Proof. If $\mathsf{hE}(\mathcal{B}) \subseteq \mathsf{hE}(\mathcal{B}')$, then $\mathsf{Inv}(\mathsf{hE}(\mathcal{B}')) \subseteq \mathsf{Inv}(\mathsf{hE}(\mathcal{B}))$. From Theorem 61, it follows that $\langle \mathcal{B}' \rangle_{\{\exists, \wedge, \vee\}\text{-FO}} \subseteq \langle \mathcal{B} \rangle_{\{\exists, \wedge, \vee\}\text{-FO}}$. Recalling that $\mathcal{B}'$ contains only a finite number of extensional relations, we may therefore effect a Logspace reduction from $\{\exists, \wedge, \vee\}$-FO$(\mathcal{B}')$ to $\{\exists, \wedge, \vee\}$-FO$(\mathcal{B})$ by straightforward substitution of predicates. $\square$

of Proposition 59. By Proposition 16, we may assume w.l.o.g. that $\mathcal{D}$ is a core. This means that every hyper-endomorphism of $\mathcal{D}$ is in fact an automorphism – we identify hyper-endomorphisms whose range are singletons with automorphisms – and thus $\mathsf{hE}(\mathcal{D})$ is a subset of S_n where $n = |D|$. If D has one element, then the problem is trival. If D has two elements, then $\mathsf{hE}(\mathcal{D}) \subseteq \mathsf{hE}(\mathcal{B}_{\mathrm{NAE}}) = S_2$. By Lemma 62, it follows that $\{\exists, \wedge, \vee\}$-FO$(\mathcal{B}_{\mathrm{NAE}}) \leq_L \{\exists, \wedge, \vee\}$-FO$(\mathcal{D})$. Since the former is a generalisation of the NP-complete CSP$(\mathcal{B}_{\mathrm{NAE}})$, the latter is NP-complete. If $\mathcal{D}$ has $n \geq 2$ elements, we proceed similarly with $\mathcal{K}_n$. $\square$

Complexity of $\{\exists, \wedge, \vee, \neq\}$-FO

When $\neq$ is present, the proof is much simpler.

Proposition 63. *In full generality, the class of problems $\{\exists, \wedge, \vee, \neq\}$-FO$(\mathcal{D})$ exhibits dichotomy: if $|D| = 1$ then the problem is in L, otherwise it is NP-complete. Consequently, the fragment extended with $=$ follows the same dichotomy.*

Proof. The proof is similar to that of Proposition 57. Let $|D| = n$. The inequality symbol $\neq$ allows to simulate $\mathcal{K}_n$. When $n = 2$, using disjunction we may simulate $\mathcal{B}_{\mathrm{NAE}}$. NP-completeness follows by reduction from CSP$(\mathcal{K}_n)$ when $n \geq 3$ and from CSP$(\mathcal{B}_{\mathrm{NAE}})$ when $n = 2$. Note that equality is not used in our hardness proof and may trivially be allowed when $|D| = 1$. Thus, the classification is the same whether one allows $=$ or not. $\square$

4.4 Third Class

Recall the fragments from this class.

- $\{\exists, \wedge, \neq\}$-FO
- $\{\exists, \forall, \wedge, \neq\}$-FO
- ★ $\{\exists, \wedge\}$-FO, $\{\exists, \wedge, =\}$-FO
- ★ $\{\exists, \forall, \wedge\}$-FO, $\{\exists, \forall, \wedge, =\}$-FO

All the fragments in this class exhibit a non trivial complexity delineation. The complexity of the last two fragments remains open and correspond to long standing open questions. We will only classify the complexity of the first two fragments, which reduce to the Boolean CSP and the Boolean QCSP, respectively.

Proposition 64 (**Complexity of** $\{\exists, \wedge, \neq\}$-FO$(\mathcal{D})$)**.** *In full generality, the problem* $\{\exists, \wedge, \neq\}$-FO$(\mathcal{D})$ *is in* L *if* $|D| = 1$, *in* P *if* $|D| = 2$ *and* $\mathcal{D}$ *is bijunctive or affine, and* NP-*complete otherwise. The fragment extended with* $=$ *follows the same dichotomy.*

Proof. We classify first the fragment extended with $=$. When $|D| \geq 3$, we may use $\neq$ to simulate CSP$(\mathcal{K}_{|D|})$ which is NP-complete. When $|D| = 1$ the problem is trivially in L. We are left with the Boolean case. Let $\mathcal{D}_{\neq}$ denote the extension of $\mathcal{D}$ with $\neq$. Note that $\{\exists, \wedge, \neq\}$-FO$(\mathcal{D})$ coincides with $\{\exists, \wedge\}$-FO$(\mathcal{D}_{\neq})$ which is the Boolean CSP$(\mathcal{D}_{\neq})$. We apply Schaefer's theorem. The relation $\neq$ is neither Horn, nor dual-Horn, nor 0-valid nor 1-valid as it is not closed under any of the following Boolean operations: $\wedge$, $\vee$, c_0 or c_1 (the constant functions 0 and 1). The relation $\neq$ is both bijunctive and affine as it is closed under both the Boolean majority and minority operation (see Chen's survey for the definitions [20]). Consequently, $\{\exists, \wedge, \neq\}$-FO$(\mathcal{D})$ is in P if $\mathcal{D}$ is bijunctive or affine and NP-complete otherwise.

Note that we have not used $=$ in the hardness proof when $|D| \geq 3$. When $|D| = 2$, we appeal to Schaefer's theorem (Theorem 51), the proof of which relies on the Galois connection $\mathsf{Pol} - \mathsf{Inv}$ which assumes presence of $=$. However, the hardness proofs in Schaefer's theorem rely on logical reductions from $\{\exists, \wedge\}$-FO$(\mathcal{B}_{\mathrm{NAE}})$, which use definability of $\mathcal{B}_{\mathrm{NAE}}$ in $\{\exists, \wedge\}$-FO. Hence, our claim follows for the fragment $\{\exists, \wedge, \neq\}$-FO. $\qquad\square$

Proposition 65. *In full generality, the problem* $\{\exists, \forall, \wedge, \neq\}$-FO$(\mathcal{D})$ *is in* L *if* $|D| = 1$, *in* P *if* $|D| = 2$ *and* $\mathcal{D}$ *is bijunctive or affine, and* Pspace-*complete otherwise. The fragment extended with* $=$ *follows the same dichotomy.*

Proof. This is similar to Proposition 64. When $|D| \geq 3$, we may use $\neq$ to simulate QCSP$(\mathcal{K}_{|D|})$ which is Pspace-complete. In the Boolean case, we apply Theorem 52 to $\{\exists, \forall, \wedge\}$-FO$(\mathcal{D}_{\neq})$ and the result follows.

Again equality is not used to prove hardness and the result follows for the fragment without $=$. $\qquad\square$

The case of $\{\exists, \wedge\}$-FO and $\{\exists, \wedge, =\}$-FO almost coincide as equality may be propagated out by substitution, and every sentence of the latter is logically equivalent to a sentence of the former, with the exception of sentences using only $=$

as an extensional predicate like $\exists x \, x = x$ which are tautologies as we only ever consider structures with at least one element. In the case of $\{\exists, \forall, \wedge, =\}$-FO, some equalities like $\exists x \exists y \, x = y$ and $\forall x \exists y \, x = y$ may also be propagated out by substitution. However, equalities like $\exists x \forall y \, x = y$ and $\forall x \forall y \, x = y$ can not, but they hold only in structures with a single element. This technical issue does not really affect the complexity classification, and it would suffice to consider $\{\exists, \wedge\}$-FO and $\{\exists, \forall, \wedge\}$-FO. The complexity classification for these four fragments remain open and correspond to the dichotomy conjecture for CSP and the classification program of the QCSP. In practice, we like to pretend that equality is present as it provides a better behaved algebraic framework, without affecting complexity.

This leaves the fragment $\{\exists, \forall, \wedge, \vee\}$-FO from our fourth class, which we deal with in the next chapter.

5. Tetrachotomy for equality-free positive first-order logic

The following – left as a conjecture at the end of [71, 80] – is one of the main contribution of Part I. Recall first that a shop f over a set D is an *A-shop* if there is an element u in D such that $f(u) = D$; and, that f is an *E-shop* if there is an element x of D such that $f^{-1}(x) = D$.

Theorem 66 (**Tetrachotomy** [70])**.** *Let $\mathcal{D}$ be any structure.*

I. *If $\mathcal{D}$ is preserved by both an A-shop and an E-shop, then $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{D})$ is in L.*

II. *If $\mathcal{D}$ is preserved by an A-shop but is not preserved by any E-shop, then $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{D})$ is NP-complete.*

III. *If $\mathcal{D}$ is preserved by an E-shop but is not preserved by any A-shop, then $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{D})$ is co-NP-complete.*

IV. *If $\mathcal{D}$ is preserved neither by an A-shop nor by an E-shop, then $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{D})$ is Pspace-complete.*

Proof. The upper bounds (membership in L, NP and co-NP) for Cases I, II and III were known from [71], but we reprove them here as a corollary of Theorem 30.

Note that an A-shop is simply a U-X-shop with $U = \{u\}$, for some u in D, and $X \subseteq D$. We may therefore replace every universal quantifier by the constant u and relativise every existential quantifier to X by Theorem 30. This means that $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{D})$ is in NP in the presence of an A-she.

Note that an E-shop is simply a U-X-shop with $X = \{x\}$ for some x in D, and $U \subseteq D$. So Case III is dual to Case II and we finally turn to Case I.

With both an A-shop and an E-shop, we have a U-X-shop with $U = \{u\}$ and $X = \{x\}$ where u and x are in D. We may therefore replace every universal quantifier by the constant u and every existential quantifier by the constant x, by Theorem 30. We have reduced $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{D})$ to the Boolean sentence value problem, known to be in L [63].

Theorem 71 deals with the lower bounds. NP-hardness for Case II and co-NP-hardness for Case III are proved in § 5.3. Pspace-hardness for Case III is proved in § 5.3. □

5.1 The Galois Connection $\mathsf{Inv} - \mathsf{shE}$

The results of this section appeared in [71].

Let $\mathsf{shE}(\mathcal{B})$ be the set of surjective hyper-endomorphisms of $\mathcal{B}$. Let $\langle\mathcal{B}\rangle_{\{\exists,\forall,\wedge,\vee\}\text{-FO}}$ be the sets of relations that may be defined on $\mathcal{B}$ in $\{\exists,\forall,\wedge,\vee\}$-FO.

Lemma 67. *Let* $\mathbf{r} := (r_1,\dots,r_k)$ *be a* k*-tuple of elements of* $\mathcal{B}$*. There exists a formula* $\theta_{\mathbf{r}}^{\{\exists,\forall,\wedge,\vee\}\text{-FO}}(u_1,\dots,u_k) \in \{\exists,\forall,\wedge,\vee\}$*-FO such that the following are equivalent.*

(i) $(\mathcal{B}, r_1',\dots,r_k') \models \theta_{\mathbf{r}}^{\{\exists,\forall,\wedge,\vee\}\text{-FO}}(u_1,\dots,u_k)$.

(ii) *There is a surjective hyper-endomorphism from* $(\mathcal{B}, r_1,\dots,r_k)$ *to* $(\mathcal{B}, r_1',\dots,r_k')$.

Proof. Let $\mathbf{r} \in B^k$, $\mathbf{s} := (b_1,\dots,b_{|B|})$ be an enumeration of B and $\mathbf{t} \in B^{|B|}$. Recall that $\varphi_{\mathcal{B}(\mathbf{r},\mathbf{s})}(u_1,\dots,u_k,v_1,\dots,v_{|B|})$ is a conjunction of the positive facts of $(\mathbf{r},\mathbf{s})$, where the variables $(\mathbf{u},\mathbf{v})$ correspond to the elements $(\mathbf{r},\mathbf{s})$.

Similarly, $\varphi_{\mathcal{B}(\mathbf{r},\mathbf{s},\mathbf{t})}(u_1,\dots,u_k,v_1,\dots,v_{|B|},w_1,\dots,w_{|B|})$ is the conjunction of the positive facts of $(\mathbf{r},\mathbf{s},\mathbf{t})$, where the variables $(\mathbf{u},\mathbf{v},\mathbf{w})$ correspond to the elements $(\mathbf{r},\mathbf{s},\mathbf{t})$. Set $\theta_{\mathbf{r}}^{\{\exists,\forall,\wedge,\vee\}\text{-FO}}(u_1,\dots,u_k) :=$

$$\exists v_1,\dots,v_{|B|}\ \varphi_{\mathcal{B}(\mathbf{r},\mathbf{s})}(u_1,\dots,u_k,v_1,\dots,v_{|B|}) \wedge \forall w_1 \dots w_{|B|}$$
$$\bigvee_{\mathbf{t}\in B^{|B|}} \varphi_{\mathcal{B}(\mathbf{r},\mathbf{s},\mathbf{t})}(u_1,\dots,u_k,v_1,\dots,v_{|B|},w_1,\dots,w_{|B|}).$$

[Backwards.] Suppose f is a surjective hyper-endomorphism from $(\mathcal{B}, r_1,\dots,r_k)$ to $(\mathcal{B}', r_1',\dots,r_k')$, where $\mathcal{B}' := \mathcal{B}$ (we will wish to differentiate the two occurrences of $\mathcal{B}$). We aim to prove that $\mathcal{B}' \models \theta_{\mathbf{r}}^{\{\exists,\forall,\wedge,\vee\}\text{-FO}}(r_1',\dots,r_k')$. Choose arbitrary $s_1' \in f(b_1),\dots,s_{|B|}' \in f(b_{|B|})$ as witnesses for $v_1,\dots,v_{|B|}$. Let $\mathbf{t}' := (t_1',\dots,t_{|B|}') \in B'^{|B|}$ be any valuation of $w_1,\dots,w_{|B|}$ and take arbitrary $t_1,\dots,t_{|B|}$ s.t. $t_1' \in f(t_1),\ \dots,\ t_{|B|}' \in f(t_{|B|})$ (here we use surjectivity). Let $\mathbf{t} := (t_1,\dots,t_{|B|})$. It follows from the definition of a surjective hyper-endomorphism that

$$\mathcal{B}' \models \varphi_{\mathcal{B}(\mathbf{r},\mathbf{s})}(r_1',\dots,r_k',s_1',\dots,s_{|B|}') \wedge \varphi_{\mathcal{B}(\mathbf{r},\mathbf{s},\mathbf{t})}(r_1',\dots,r_k',s_1',\dots,s_{|B|}',t_1',\dots,t_{|B|}').$$

[Forwards.] Assume that $\mathcal{B}' \models \theta_{\mathbf{r}}^{\{\exists,\forall,\wedge,\vee\}\text{-FO}}(r_1',\dots,r_k')$, where $\mathcal{B}' := \mathcal{B}$. Let $b_1',\dots,b_{|B|}'$ be an enumeration of $B' := B$.[1] Choose some witness elements $s_1',\dots,s_{|B|}'$ for $v_1,\dots,v_{|B|}$ and a witness tuple $\mathbf{t} := (t_1,\dots,t_{|B|}) \in B^{|B|}$ s.t.

(†) $\mathcal{B}' \models \varphi_{\mathcal{B}(\mathbf{r},\mathbf{s})}(r_1',\dots,r_k',s_1',\dots,s_{|B|}') \wedge \varphi_{\mathcal{B}(\mathbf{r},\mathbf{s},\mathbf{t})}(r_1',\dots,r_k',s_1',\dots,s_{|B|}',b_1',\dots,b_{|B|}')$.

Consider the following partial hyper-functions from B to B'.

1. $f_{\mathbf{r}}$ given by $f_{\mathbf{r}}(r_i) := \{r_i'\}$, for $1 \le i \le k$.

2. $f_{\mathbf{s}}$ given by $f_{\mathbf{s}}(b_i) = \{s_i'\}$, for $1 \le i \le |B|$. (totality)

3. $f_{\mathbf{t}}$ given by $b_i' \in f_{\mathbf{t}}(b_j)$ iff $t_i = b_j$, for $1 \le i,j \le |B|$. (surjectivity)

---

[1] One may imagine $b_1,\dots,b_{|B|}$ and $b_1',\dots,b_{|B|}'$ to be the same enumeration, but this is not essential. In any case, we will wish to keep the dashes on the latter set to remind us they are in $\mathcal{B}'$ and not $\mathcal{B}$.

Let $f := f_{\mathbf{r}} \cup f_{\mathbf{s}} \cup f_{\mathbf{t}}$; f is a hyper-operation whose surjectivity is guaranteed by $f_{\mathbf{t}}$ (note that totality is guaranteed by $f_{\mathbf{s}}$). That f is a surjective hyper-endomorphism follows from the right-hand conjunct of (†). □

Theorem 68 (**Galois Connection** $\mathsf{Inv} - \mathsf{shE}$). *For a finite structure $\mathcal{B}$ we have* $\langle \mathcal{B} \rangle_{\{\exists, \forall, \wedge, \vee\}\text{-FO}} = \mathsf{Inv}(\mathsf{shE}(\mathcal{B}))$.

Proof.   1. $\varphi(\mathbf{v}) \in \langle \mathcal{B} \rangle_{\{\exists, \forall, \wedge, \vee\}\text{-FO}} \Rightarrow \varphi(\mathbf{v}) \in \mathsf{Inv}(\mathsf{shE}(\mathcal{B}))$. This is proved by induction on the complexity of $\varphi(\mathbf{v})$. We only have to deal with the case of universal quantification in the inductive step, the other cases having been dealt with in the proof of the $\mathsf{Inv} - \mathsf{hE}$ Galois Connection.

   (**Inductive Step** continued from proof of Theorem 61.)

   (d) $\varphi(\mathbf{v}) := \forall w \ \psi(\mathbf{v}, w)$. Let $\mathbf{v} := (v_1, \ldots, v_l)$. Suppose $\mathcal{B} \models \forall w \ \psi(x_1, \ldots, x_l, w)$; then for each x', $\mathcal{B} \models \psi(x_1, \ldots, x_l, x')$. By IH, for any $y_1 \in f(x_1), \ldots, y_l \in f(x_l)$, we have for all y' (remember f is surjective), $\mathcal{B} \models \psi(y_1, \ldots, y_l, y')$, whereupon $\mathcal{B} \models \forall w \ \psi(y_1, \ldots, y_l, w)$.

   2. $S \in \mathsf{Inv}(\mathsf{shE}(\mathcal{B})) \Rightarrow S \in \langle \mathcal{B} \rangle_{\{\exists, \forall, \wedge, \vee\}\text{-FO}}$. Consider the k-ary relation $S \in \mathsf{Inv}(\mathsf{shE}(\mathcal{B}))$. Let $\mathbf{r}_1, \ldots, \mathbf{r}_m$ be the tuples of S. Let $\theta_S^{\{\exists, \forall, \wedge, \vee\}\text{-FO}}(u_1, \ldots, u_k)$ be the following formula of $\{\exists, \forall, \wedge, \vee\}$-FO:

   $$\theta_{\mathbf{r}_1}^{\{\exists, \forall, \wedge, \vee\}\text{-FO}}(u_1, \ldots, u_k) \qquad \vee \qquad \ldots \qquad \vee \qquad \theta_{\mathbf{r}_m}^{\{\exists, \forall, \wedge, \vee\}\text{-FO}}(u_1, \ldots, u_k).$$

   For $\mathbf{r}_i := (r_{i1}, \ldots, r_{ik})$, note that $(\mathcal{B}, r_{i1}, \ldots, r_{ik}) \models \theta_{\mathbf{r}_i}^{\{\exists, \forall, \wedge, \vee\}\text{-FO}}(u_1, \ldots, u_k)$ (viewing the identity endomorphism as a surjective hyper endomorphism). That $\theta_S^{\{\exists, \forall, \wedge, \vee\}\text{-FO}}(u_1, \ldots, u_k) = S$ now follows from Part (ii) of Lemma 67, since $S \in \mathsf{Inv}(\mathsf{shE}(\mathcal{B}))$. □

□

Corollary 69. *Let $\mathcal{B}$ and $\mathcal{B}'$ be finite structures over the same domain B. If $\mathsf{shE}(\mathcal{B}) \subseteq \mathsf{shE}(\mathcal{B}')$ then $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{B}') \leq_{\mathrm{L}} \{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{B})$.*

Proof. If $\mathsf{shE}(\mathcal{B}) \subseteq \mathsf{shE}(\mathcal{B}')$, then $\mathsf{Inv}(\mathsf{shE}(\mathcal{B}')) \subseteq \mathsf{Inv}(\mathsf{shE}(\mathcal{B}))$. From Theorem 68, it follows that $\langle \mathcal{B}' \rangle_{\{\exists, \forall, \wedge, \vee\}\text{-FO}} \subseteq \langle \mathcal{B} \rangle_{\{\exists, \forall, \wedge, \vee\}\text{-FO}}$. Recalling that $\mathcal{B}'$ contains only a finite number of extensional relations, we may therefore effect a Logspace reduction from $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{B}')$ to $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{B})$ by straightforward substitution of predicates. □

Consequently, the complexity of $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{B})$ is characterised by $\mathsf{shE}(\mathcal{B})$.

5.2   The Boolean case

We consider the case $|B| = 2$, with the normalised domain $B := \{0, 1\}$ as a warm-up. It may easily be verified that there are five DSMs in this case, depicted as a lattice in Figure 7.2. The two elements of this lattice that represent the two subgroups of S_2 are drawn in the middle and bottom. We write $\frac{0}{1} \Big| \frac{01}{1}$ for the shop that sends 0 to $\{0, 1\}$ and 1 to $\{1\}$.
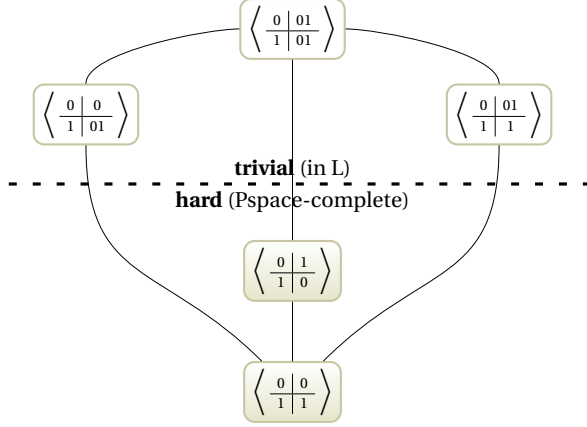
Figure 5.1: The boolean lattice of DSMs with their associated complexity.

Theorem 70 ([71]). *Let $\mathcal{B}$ be a boolean structure.*

I. *If either* $\frac{0\,|\,01}{1\,|\,1}$ *or* $\frac{0\,|\,0}{1\,|\,01}$ *is a surjective hyper-endomorphism of $\mathcal{B}$, then $\{\exists, \forall, \wedge, \vee\}$-FO($\mathcal{B}$) is in* L.

II. *Otherwise,* $\{\exists, \forall, \wedge, \vee\}$-FO($\mathcal{B}$) *is* Pspace*-complete.*

Proof. $\mathsf{shE}(\mathcal{B})$ must be one of the five DSMs depicted in Figure 7.2. If $\mathsf{shE}(\mathcal{B})$ contains $\frac{0\,|\,01}{1\,|\,1}$ then we may relativise every existential quantifier to 1 and every universal quantifier to 0 by Theorem 29 and evaluate in L the quantifier-free part. The case of $\frac{0\,|\,0}{1\,|\,01}$ is similar with the role of 0 and 1 swapped.

We prove that if $\mathsf{shE}(\mathcal{B}) = \langle \frac{0\,|\,1}{1\,|\,0} \rangle$ then $\{\exists, \forall, \wedge, \vee\}$-FO($\mathcal{B}$) is Pspace-complete. The structure $\mathcal{K}_2$ has DSM $\mathsf{shE}(\mathcal{B}) = \langle \frac{0\,|\,1}{1\,|\,0} \rangle$. It suffices therefore to prove that $\{\exists, \forall, \wedge, \vee\}$-FO($\mathcal{K}_2$) is Pspace-hard, which we did by reduction from QCSP($\mathcal{B}_{\text{NAE}}$) in the proof of Proposition 57.

If follows from Corollary 69 that when $\mathsf{shE}(\mathcal{B}) = \langle \frac{0\,|\,0}{1\,|\,1} \rangle$, $\{\exists, \forall, \wedge, \vee\}$-FO($\mathcal{B}$) is also Pspace-hard since $\langle \frac{0\,|\,0}{1\,|\,1} \rangle \subseteq \langle \frac{0\,|\,1}{1\,|\,0} \rangle$. $\qquad\square$

5.3 Proving Hardness

Our aim is to derive the following lower bounds.

Theorem 71 (**lower bounds**). *Let $\mathcal{D}$ be a structure.*

II. *If $\mathcal{D}$ is preserved by an A-shop but is not preserved by any E-shop, then $\{\exists, \forall, \wedge, \vee\}$-FO($\mathcal{D}$) is NP-hard.*

III. *If $\mathcal{D}$ is preserved by an E-shop but is not preserved by any A-shop, then $\{\exists, \forall, \wedge, \vee\}$-FO($\mathcal{D}$) is co-NP-hard.*

IV. *If $\mathcal{D}$ is preserved neither by an A-shop nor by an E-shop, then $\{\exists, \forall, \wedge, \vee\}$-FO($\mathcal{D}$) is Pspace-hard.*

It follows from Proposition 38 and Corollary 31 that the complexity of a structure $\mathcal{D}$ is the same as the complexity of its *U-X-core*. Hence in this Section, *we assume w.l.o.g. that $U \cup X = D$*. We will say in this case that the DSM $\mathcal{M}$ is *reduced*. In order to prove Theorem 71, we need to establish the following:

II. If U is of size one and X of size at least two then $\{\exists, \forall, \wedge, \vee\}$-FO($\mathcal{D}$) is NP-hard;

III. If X is of size one and U of size at least two then $\{\exists, \forall, \wedge, \vee\}$-FO($\mathcal{D}$) is co-NP-hard; and,

IV. If both U and X have at least two elements then $\{\exists, \forall, \wedge, \vee\}$-FO($\mathcal{D}$) is Pspace-hard.

In the following. we will describe a DSM $\mathcal{M}$ as being (NP, co-NP, Pspace)hard in the case that $\{\exists, \forall, \wedge, \vee\}$-FO($\mathcal{D}$) is hard for some $\mathcal{D} \in \mathsf{Inv}(\mathcal{M})$. In order to facilitate the hardness proof, we would like to show hardness of a monoid $\widehat{\mathcal{M}}$ with a very simple structure of which $\mathcal{M}$ is in fact a sub-DSM ($\widehat{\mathcal{M}}$ is the *completion* of $\mathcal{M}$). As in general $\widehat{\mathcal{M}}$ preserves fewer relations than $\mathcal{M}$, the hardness of $\mathcal{M}$ would follow. We would like the structure of $\widehat{\mathcal{M}}$ to be sufficiently simple for us to build canonically some gadgets for our hardness proof. Thus, we wish to better understand the form that elements of $\mathcal{M}$ may take. In order to do so, we first define the *canonical shop* of $\mathcal{M}$ to be the *U-X* shop h in $\mathcal{M}$, guaranteed by Proposition 38, with the property that $|h(z)|$ is maximal for each $z \in U \setminus X$. Note that this maximal h is unique, as given h_1 and h_2 of the form in Proposition 38, $h_1 \circ h_2$ is also of the required form, and further satisfies $|h_1 \circ h_2(z)| \geq |h_1(z)|, |h_2(z)|$, for all $z \in U \setminus X$.

Characterising reduced DSMs

Any *U-X*-shop in $\mathcal{M}$ will be shown to be in the following special form, reminiscent of the form of the canonical shop.

Definition 72. *We say that a shop f is in the* 3-permuted form *if there are a permutation ζ of $X \cap U$, a permutation χ of $X \setminus U$ and a permutation υ of $U \setminus X$ such that f satisfies:*

- *for any y in $U \cap X$, $f(y) = \{\zeta(y)\}$;*

- *for any x in $X \setminus U$, $f(x) = \{\chi(x)\}$; and,*

- *for any u in $U \setminus X$, $f(u) = \{\upsilon(u)\} \cup X_u$, where $X_u \subseteq X \setminus U$.*

Lemma 73. *If a shop f satisfies $f(X) \cap (U \setminus X) = \emptyset$ then f is in the* 3-permuted form.

Proof. The hypothesis forces an element of X to reach an element of X and Lemma 35 forces two elements of X to have different images. Since X is finite, there exists a permutation β of X such that for every x in X, $f(x) = \{\beta(x)\}$. Since

Lemma 33 forces in particular an element of U to have at most one element of U in its image and since U is finite, it follows that there exists a permutation α of U such that for every u in U, $f(u) \cap U = \{\alpha(U)\}$ and $f^{-1}(u) \cap U = \{\alpha^{-1}(U)\}$.

It follows that there exists a permutation ζ of $U \cap X$ such that for any y in $U \cap X$, $f(y) = \{\zeta(y)\}$.

The existence of a permutation χ of $X \setminus U$ such that β is the disjoint union of χ and ζ follows. Hence, for any x in $X \setminus U$, $f(x) = \{\chi(x)\}$.

Similarly, there must also be a permutation υ of $U \setminus X$ such that α is the disjoint union of υ and ζ. Hence, for any u in $U \setminus X$, $f(u) \cap U = \{\upsilon(u)\}$. Elements of $U \setminus X$ may however have some images in $X \setminus U$. So we get finally that for any u in $U \setminus X$, there is some $\emptyset \subseteq X_u \subseteq X \setminus U$ such that $f(u) = \{\upsilon(u)\} \cup X_u$. This proves that f is in the 3-permuted form and we are done. $\qquad\square$

Theorem 74. *Let $\mathcal{M}$ be a reduced DSM. Every shop in $\mathcal{M}$ is in the 3-permuted form. Moreover, every U-X-shop in $\mathcal{M}$ follows the additional requirement that the elements of $U \setminus X$ cover the set $X \setminus U$, more formally that*

$$f(U \setminus X) \cap X = \bigcup_{u \in U \setminus X} X_u = X \setminus U.$$

Proof. We can now deduce easily from Lemmata 37 and 73 that U-X-shops in $\mathcal{M}$ must take the 3-permuted form. It remains to prove that an arbitrary shop f in $\mathcal{M}$ is in the 3-permuted form. Let h be the canonical shop of $\mathcal{M}$. It follows from Lemma 19 that $f' := h \circ f \circ h$ is a U-X-shop. Hence, f' is in the 3-permuted form. Let z in X and u in $U \setminus X$. If $f(z) \ni u$ then $f'(z) \ni u$ and f' would not be in the 3-permuted form. It follows that $f(X) \cap (U \setminus X) = \emptyset$ and appealing to Lemma 73 that f is in the 3-permuted form. $\qquad\square$

We do not need the following result in order to prove our main result. But surprisingly in a reduced DSM, U and X are unique. This means that we may speak of *the canonical shop of $\mathcal{M}$* instead of some canonical U-X-shop. It also means that we can define the U-X-core of a structure $\mathcal{D}$ *without explicitly referring to U or X* as the minimal substructure of $\mathcal{D}$ which satisfy the same $\{\exists, \forall, \wedge, \vee\}$-FO sentences.

Theorem 75. *Let $\mathcal{D}$ be a structure that is both a U-X-core and a U'-X'-core then it follows that $U = U'$ and $X = X'$.*

Proof. We do a proof by contradiction. Let h and h' be the canonical U-X-shop and U'-X'-shop, respectively. Assume $U' \neq U$ and let x in $U' \setminus U$. Note that since $D = U \cup X$, our notation is consistent as x does belong to $X \setminus U$. Thus, there exists some u in $U \setminus X$ such that $h(u) \supseteq \{u, x\}$ (and necessarily $u \neq x$).

By Theorem 74, h has to be in the 3-permuted form w.r.t. U' and X', which means that h can send an element to at most one element of U'. Since x belongs to U', it follows that u belongs to $D \setminus U' = X' \setminus U'$. But the three permuted form prohibits an element of X' to reach an element of U'. A contradiction.

It does not follow yet that $X' = X$ as the pairs of sets may have shifting intersections. However, the dual argument to the above applies and yields $X = X'$. □

Corollary 76. *Let $\mathcal{D}$ be a finite structure. The U-X-core of $\mathcal{D}$ is unique up to isomorphism. It is a minimal induced substructure $\widetilde{\mathcal{D}}$ of $\mathcal{D}$, that satisfies the same $\{\exists, \forall, \wedge, \vee\}$-FO formulae with free-variables in $\widetilde{D}$. Moreover, once $\widetilde{D}$ is fixed, there are two uniquely determined subsets U and X such that $U \cup X = \widetilde{\mathcal{D}} \subset D$ which are minimal within D with respect to the following equivalent properties,*

- *$\mathcal{D}$ has $\forall U$-$\exists X$-relativisation w.r.t. $\{\exists, \forall, \wedge, \vee\}$-FO; or,*

- *$\mathcal{D}$ has a U-X-shop that may act as the identity over $U \cup X$.*

Proof. The last point follows from our definition of a U-X-core and from Proposition 38. It is equivalent to the $\forall U$-$\exists X$-*relativisation* property by Theorem 30. It follows that $\mathcal{D}$ and $\widetilde{\mathcal{D}}$ satisfy the same $\{\exists, \forall, \wedge, \vee\}$-FO formulae with free-variables in $\widetilde{D}$ (see Corollary 31). Conversely, if $\mathcal{D}$ and $\widetilde{\mathcal{D}}$ satisfy the same $\{\exists, \forall, \wedge, \vee\}$-FO formulae with free-variables in $\widetilde{D}$, then $\mathcal{D}$ has $\widetilde{D}$-$\widetilde{D}$-relativisation. The existence of a "$\widetilde{D}$-$\widetilde{D}$-shop" follows by Theorem 30. Enforcing the minimality criteria, we get some U-X-shop with some $U, X \subseteq \widetilde{D}$ (this is because, "we may proceed by retraction", as explained in the beginning of Subsection 3.5). Moreover, by minimality of $\widetilde{\mathcal{D}}$, we must have $U \cup X = \widetilde{D}$. We have a U-X-core as in our original definition in terms of a U-X-shop satisfying minimality criteria. It follows from Theorem 75 that U and X are unique (within $\widetilde{D}$). □

Recall that the $\{\exists, \forall, \wedge, \vee\}$-FO-core $\mathcal{D}'$ of $\mathcal{D}$ is the smallest (w.r.t. domain size) structure that is $\{\exists, \forall, \wedge, \vee\}$-FO-equivalent to $\mathcal{D}$.

Proposition 77. *The notion of a U-X-core and of a $\{\exists, \forall, \wedge, \vee\}$-FO-core coincide.*

Proof. Let $\mathcal{D}$ be a structure that is a U-X-core with (unique) subsets U and X. Let c be the canonical shop of $\mathcal{D}$.

Let $\mathcal{D}'$ be a $\{\exists, \forall, \wedge, \vee\}$-FO-core of $\mathcal{D}$, that is a smallest (w.r.t. domain size) structure that is $\{\exists, \forall, \wedge, \vee\}$-FO-equivalent to $\mathcal{D}$. Let U' and X' be subsets of D' witnessing that $\mathcal{D}'$ is a U'-X' core. Note that $U' \cup X' = D'$ by minimality of $\mathcal{D}'$ (and consequently, U' and X' are uniquely determined by Theorem 75). Let c' be the canonical shop of $\mathcal{D}'$.

By Proposition 21, since $\mathcal{D}$ and $\mathcal{D}'$ are $\{\exists, \forall, \wedge, \vee\}$-FO-equivalent, there exist two surjective hyper-morphisms g from $\mathcal{D}$ to $\mathcal{D}'$ and f from $\mathcal{D}'$ to $\mathcal{D}$.

Let U'' be a minimal subset of $(g)^{-1}(U')$ such that $g(U'') = U'$. Note that $f \circ c' \circ g$ is a U''-surjective shop of $\mathcal{D}$. By minimality of U, it follows that $|U| \le |U''| \le |U'|$. A similar argument over $\mathcal{D}'$ gives $|U'| \le |U|$, and consequently, $|U| = |U'|$. Moreover, since $c \circ (f \circ c' \circ g)$ is a U''-surjective X-total surjective hyperendomorphism of $\mathcal{D}$, By Theorem 75, it follows that $U = U''$.

This means that there is a bijection α' from U' to U such that, for any u' in U', $g^{-1}(u') = \{\alpha'(u')\}$.

By duality we obtain similarly that $|X|=|X'|$ and that there is a bijection β from X to X' such that, for any x in X, $g(x)=\{\beta(x)\}$.

Thus, g acts necessarily as a bijection from $U\cap X$ to $U'\cap X'$.

The map $\tilde{g}$ from D to D' defined for any u in U as $\tilde{g}(u):=\alpha'^{-1}(u)$ and $\tilde{g}(x):=\beta(x)$ is a homomorphism from $\mathcal{D}$ to $\mathcal{D}'$ that is both injective and surjective.

A symmetric argument yields a map $\tilde{f}$ that is a bijective homomorphism from $\mathcal{D}'$ to $\mathcal{D}$. Isomorphism of $\mathcal{D}'$ and $\mathcal{D}$ follows.

$\square$

Remark 78. To simplify the presentation, we defined the $\mathcal{L}$-core as a minimal structure w.r.t. domain size. Considering minimal structures w.r.t. inclusion, we would get the same notion for $\{\exists,\forall,\wedge,\vee\}$-FO. This is also the case for CSP, but it is not the case in general. For example, this is not the case for the logic $\{\exists,\forall,\wedge\}$-FO, which corresponds to QCSP [72].

Lemma 79. *Let $\mathcal{M}$ be a reduced DSM with associated sets U and X. There are only three cases possible.*

1. *$U\cap X\neq\emptyset$, $U\setminus X\neq\emptyset$ and $U\setminus X\neq\emptyset$.*

2. *$U=X$.*

3. *$U\cap X=\emptyset$.*

Proof. We prove that $U\subsetneq X$ is not possible. Otherwise, let x in $X\setminus U$ and h be the canonical shop. There exists some u in $U\subsetneq X$ such that $h(u)\ni x$ by U-surjectivity of h. Since u does not occur in the image of any other element than u under the canonical shop, this would mean that h is $X\setminus\{u\}$-total, contradicting the minimality of X.

By duality $X\subsetneq U$ is not possible either and the result follows. $\square$

The hard DSM above $\mathcal{M}$

Define the completion $\widehat{\mathcal{M}}$ of $\mathcal{M}$ to be the DSM that contains *all* shops in the 3-permuted form of $\mathcal{M}$. More precisely, the canonical shop of $\widehat{\mathcal{M}}$ is the shop $\hat{h}$ where every set X_u is the whole set $X\setminus U$, and, for every permutation ζ of $X\cap U$, χ of $X\setminus U$ and υ of $U\setminus X$, any shop in the 3-permuted form with these permutations is in $\widehat{\mathcal{M}}$. Note that by construction, $\mathcal{M}$ is a sub-DSM of $\widehat{\mathcal{M}}$. Note also that the minimality of U and X still holds in $\widehat{\mathcal{M}}$. We will establish hardness for $\widehat{\mathcal{M}}$, whereupon hardness of $\mathcal{M}$ follows from Theorem 68.

Cases II and III: NP-hardness and co-NP-hardness

We begin with Case II. We note first that $U=\{u\}$ and $|X|\geq 2$ implies $U\cap X=\emptyset$ by Lemma 79. The structure $\mathcal{K}_{|X|}\uplus\mathcal{K}_1$, the disjoint union of a clique of size $|X|$ with an isolated vertex u, has associated DSM $\widehat{\mathcal{M}}$. The problem $\{\exists,\wedge,\vee\}$-FO$(\mathcal{K}_{|X|}\uplus\mathcal{K}_1)$ is NP-hard, since the core of $\mathcal{K}_{|X|}\uplus\mathcal{K}_1$ is $\mathcal{K}_{|X|}$ by Proposition 59.

For Case III, we may assume similarly to above that $X=\{x\}$, $|U|\geq 2$ and $U\cap X=\emptyset$ by Lemma 79. We use the duality principle, which corresponds to taking the inverse of shops.

Proposition 80. *Let $\mathcal{D}$ be a structure. The set* $\mathsf{shE}(\overline{\mathcal{D}})$ *consists of exactly the inverses of the shops in* $\mathsf{shE}(\mathcal{D})$.

Proof. Follows directly from the definitions. $\qquad\qquad\qquad\qquad\qquad\square$

Observing that the inverse of a $\{x\}$-total U-surjective shop with $U \geq 2$ is a $\{U\}$-total $\{x\}$-surjective shop, it follows that $\{\exists, \forall, \wedge, \vee\}$-FO$(\overline{\mathcal{D}})$ is NP-hard and consequently $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{D})$ is co-NP-hard. Alternatively, we could use the structure $\overline{\mathcal{K}_{|U|} \uplus \mathcal{K}_1}$ which is $\{\forall, \vee, \wedge\}$-FO-equivalent to $\overline{\mathcal{K}_{|\mathcal{U}|}}$ and use the fact that $\{\forall, \vee, \wedge\}$-FO$(\overline{\mathcal{K}_{|\mathcal{U}|}})$ is co-NP-hard.

case IV: Pspace-hardness
We assume that $|U| \geq 2$ and $|X| \geq 2$ and consider the tree possible cases given by Lemma 79.

Case 1: when $U \cap X \neq \emptyset$, $U \setminus X \neq \emptyset$ and $X \setminus U \neq \emptyset$
Recall that if $\mathcal{M}$ is a sub-DSM of a hard DSM $\widehat{\mathcal{M}}$ then $\mathcal{M}$ is also hard (see Theorem 68).

We write $U \Delta X$ as an abbreviation for $(X \setminus U) \cup (U \setminus X)$. To build $\widehat{\mathcal{M}}$ from $\mathcal{M}$, we added all permutations, and chose for each set $X_u = X \setminus U$. We carry on with this completion process and consider the super-DSM $\mathcal{M}'$ which is generated by a single shop g' defined as follows:

- for every y in $X \cap U$, $g'(y) := X \Delta U$; and,

- for every z in $X \Delta U$, $g'(z) := X \cap U$, where $X \Delta U$ denotes $(X \setminus U) \cup (U \setminus X)$.

The complete bipartite graph $\mathcal{K}_{X \Delta U, X \cap U}$ has $\mathcal{M}'$ for DSM. Observing that there is a full surjective homomorphism from $\mathcal{K}_{X \Delta U, X \cap U}$ to $\mathcal{K}_2$, thus by Proposition 21 the two structures agree on all sentences of $\{\exists, \forall, \wedge, \vee, \neg\}$-FO and so also on all sentences of $\{\exists, \forall, \wedge, \vee\}$-FO. It suffices therefore to prove that $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{K}_2)$ is Pspace-hard, which we did by reduction from QCSP$(\mathcal{B}_{\mathrm{NAE}})$ in the proof of Theorem 70.

Case 2: when $U = X$
The clique $\mathcal{K}_{|U|}$ has DSM $\widehat{\mathcal{M}}$. The problem $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{K}_{|U|})$ is Pspace-complete by Theorem 70 in the Boolean case; and, beyond that, it is also Pspace-hard as a generalisation of the Pspace-complete QCSP$(\mathcal{K}_{|U|})$. The Pspace-completeness of $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{D})$ follows from Theorem 68.

Case 3: when $U \cap X = \emptyset$
We can no longer complete the monoid $\widehat{\mathcal{M}}$ into $\mathcal{M}'$, as we would end up with a trivial monoid. The remainder of this section is devoted to a generic hardness proof. Assume that $|U| = j \geq 2$ and $|X| = k \geq 2$ and w.l.o.g. let $U = \{1, 2, \ldots j\}$ and $X = \{j+1, j+2, \ldots j+k\}$. Recalling that the symmetric group is generated by a
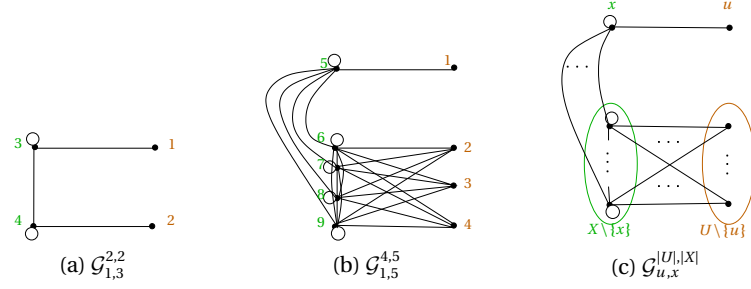
(a) $\mathcal{G}_{1,3}^{2,2}$

(b) $\mathcal{G}_{1,5}^{4,5}$

(c) $\mathcal{G}_{u,x}^{|U|,|X|}$

Figure 5.2: Main Gadget.

transposition and a cyclic permutation, let $\widehat{\mathcal{M}}$ be the DSM given by

$$
\Big\langle
\begin{array}{c|c}
1 & 2,j+1,\dots,j+k \\
\hline
2 & 1,j+1,\dots,j+k \\
\hline
3 & 3,j+1,\dots,j+k \\
\hline
\vdots & \vdots \\
\hline
j & j,j+1,\dots,j+k \\
\hline
j+1 & j+1 \\
\hline
j+2 & j+2 \\
\hline
j+3 & j+3 \\
\hline
\vdots & \vdots \\
\hline
j+k & j+k
\end{array}
,
\begin{array}{c|c}
1 & 2,j+1,\dots,j+k \\
\hline
2 & 3,j+1,\dots,j+k \\
\hline
3 & 4,j+1,\dots,j+k \\
\hline
\vdots & \vdots \\
\hline
j & 1,j+1,\dots,j+k \\
\hline
j+1 & j+1 \\
\hline
j+2 & j+2 \\
\hline
j+3 & j+3 \\
\hline
\vdots & \vdots \\
\hline
j+k & j+k
\end{array}
,
\begin{array}{c|c}
1 & 1,j+1,\dots,j+k \\
\hline
2 & 2,j+1,\dots,j+k \\
\hline
3 & 3,j+1,\dots,j+k \\
\hline
\vdots & \vdots \\
\hline
j & j,j+1,\dots,j+k \\
\hline
j+1 & j+2 \\
\hline
j+2 & j+1 \\
\hline
j+3 & j+3 \\
\hline
\vdots & \vdots \\
\hline
j+k & j+k
\end{array}
,
\begin{array}{c|c}
1 & 1,j+1,\dots,j+k \\
\hline
2 & 2,j+1,\dots,j+k \\
\hline
3 & 3,j+1,\dots,j+k \\
\hline
\vdots & \vdots \\
\hline
j & j,j+1,\dots,j+k \\
\hline
j+1 & j+2 \\
\hline
j+2 & j+3 \\
\hline
j+3 & j+4 \\
\hline
\vdots & \vdots \\
\hline
j+k & j+1
\end{array}
\Big\rangle .
$$

We will give a structure $\widehat{\mathcal{D}}$ such that $\mathsf{shE}(\widehat{\mathcal{D}}) = \widehat{\mathcal{M}}$. Firstly, though, given some fixed u in U and x in X, let $\mathcal{G}_{u,x}^{|U|,|X|}$ be the symmetric graph with self-loops with domain $D = U \cup X$ such that

- u and x are adjacent;

- The graph induced by X is a reflexive clique $\mathcal{K}_X^{\mathrm{ref}}$; and,

- $U \setminus \{u\}$ and $X \setminus \{x\}$ are related via a complete bipartite graph $\mathcal{K}_{|X \setminus \{x\}|,|U \setminus \{u\}|}$.

The structure $\mathcal{G}_{u,x}^{|U|,|X|}$ and the more specific $\mathcal{G}_{1,5}^{4,5}$ are drawn in Figure 5.2. Denote by $E_{u,x}^{|U|,|X|}$ the binary relation of $\mathcal{G}_{u,x}^{|U|,|X|}$ and let $\widehat{\mathcal{D}}$ be the structure with a single 4-ary relation $R^{\widehat{\mathcal{D}}}$ with domain $\widehat{D} = U \cup X$ specified as follows,

$$
R^{\widehat{\mathcal{D}}} \quad := \quad \bigcup_{u \in U} \left(\left(\bigcup_{x \in X} (u,x) \times E_{u,x}^{|U|,|X|} \right) \cup \left(\bigcup_{x_1,x_2,x_3 \in X} (x_1,x_2) \times E_{u,x_3}^{|U|,|X|} \right) \right).
$$

Essentially, when the first argument in a quadruple is from U, then the rest of the structure allows for the unique recovery of some $\mathcal{G}_{u,x}^{|U|,|X|}$; but if the first argument is from X then all possibilities from X for the remaining arguments are allowed. In particular, we note from the last big cup that (x_1,x_2,x_3,x_4) is a tuple of $R^{\widehat{\mathcal{D}}}$ for all quadruples x_1,x_2,x_3,x_4 in X.

Lemma 81. $\mathsf{shE}(\widehat{\mathcal{D}}) = \widehat{\mathcal{M}}$.

Proof. Recall that, according to Theorem 74 and our assumption on U, X and $\widehat{\mathcal{M}}$, a maximal (w.r.t. sub-shop inclusion) shop f' is of the following form,

- for any x in $X \setminus U = X$, $f(x) = \{\chi(x)\}$; and,

- for any u in $U \setminus X = U$, $f(u) = \{v(u)\} \cup X$.

where χ and v are permutations of X and U, respectively.

(Backwards; $\widehat{\mathcal{M}} \subseteq \mathsf{shE}(\widehat{\mathcal{D}})$.) It suffices to check that a maximal shop f' in $\widehat{\mathcal{M}}$ preserves $\widehat{\mathcal{D}}$. This holds by construction. We consider first tuples from $(x_1, x_2) \times E_{u,x_3}^{|U|,|X|}$.

- A tuple with elements from X only will map to a like tuple, which must occur, so we can ignore such tuples from now on.

- A tuple (x_1, x_2, u, x_3) maps either to $\big(\chi(x_1), \chi(x_2), v(u), \chi(x_3)\big)$ which appears in $\big(\chi(x_1), \chi(x_2)\big) \times E_{v(u),\chi(x_3)}^{|U|,|X|}$, or it maps to a tuple containing only elements from X.

- A tuple (x_1, x_2, x_3, u) maps either to $\big(\chi(x_1), \chi(x_2), \chi(x_3), v(u)\big)$, which appears in $\big(\chi(x_1), \chi(x_2)\big) \times E_{v(u),\chi(x_3)}^{|U|,|X|}$, or it maps to a tuple containing only elements from X.

We consider now tuples from $(u, x) \times E_{u,x}^{|U|,|X|}$.

- If the first coordinate u is mapped to $v(u)$, then the tuple is mapped to different tuples from $\big(v(u), \chi(x)\big) \times E_{v(u),\chi(x)}^{|U|,|X|}$, depending whether the second u is mapped to an element from X or to $v(u)$.

- Otherwise, the first coordinate u is mapped to an element x_1 from X, and some other element from u' in U occurs (or the tuple contains elements from X only) and a tuple is mapped to a tuple of the form $(x_1, \chi(x), v(u'), x_3)$ which appears in $\big(x_1, \chi(x)\big) \times E_{v(u'),x_3}^{|U|,|X|}$.

(Forwards; $\mathsf{shE}(\widehat{\mathcal{D}}) \subseteq \widehat{\mathcal{M}}$.) We proceed by contraposition, demonstrating that $R^{\widehat{\mathcal{D}}}$ is violated by any $f \notin \widehat{\mathcal{M}}$. We consider the different ways that f might not be in $\widehat{\mathcal{M}}$.

- If f is s.t. $u \in f(x)$ for $x \in X$ and $u \in U$ then we, e.g., take $(u, x, x, x) \in R^{\widehat{\mathcal{D}}}$ but $(z, u, u, u) \notin R^{\widehat{\mathcal{D}}}$ (for any $z \in f(u)$) and we are done. It follows that $f(X) = X$.

- Assume now that f is s.t. $\{x_1', x_2'\} \subseteq f(x)$ for $x_1' \neq x_2'$ and $x, x_1', x_2' \in X$. Let $u, u' \in U$ be s.t. $u' \in f(u)$. Take $(u, x, u, x) \in R^{\widehat{\mathcal{D}}}$; $(u', x_1', u', x_2') \notin R^{\widehat{\mathcal{D}}}$ and we are done. It follows that f is a permutation χ on X.

- Assume now that f is s.t. $\{u_1', u_2'\} \subseteq f(u)$ for $u_1' \neq u_2'$ and $u, u_1', u_2' \in U$. Let $x, x' \in X$ be s.t. $x' \in f(x)$. Take $(u, x, u, x) \in R^{\widehat{\mathcal{D}}}$; $(u_1', x', u_2', x') \notin R^{\widehat{\mathcal{D}}}$ and we are done. It follows that f restricted to U is a permutation v on U.

Hence, f is a sub-shop of a maximal shop f' from the DSM $\widehat{\mathcal{M}}$, and f belongs to $\widehat{\mathcal{M}}$ (recall that a DSM is closed under sub-shops). The result follows. $\qquad\square$

Proposition 82. $\{\exists, \forall, \wedge, \vee\}$-FO$(\widehat{\mathcal{D}})$ *is Pspace-complete.*

Proof. We begin with the observation that $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{G}_{u,x}^{|U|,|X|})$ is Pspace-complete (for each $u \in U$ and $x \in X$). This follows straightforwardly from the Pspace-completeness of $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{G}_{1,3}^{2,2})$, the simplest gadget which is depicted

on Figure 5.2a. These gadgets $\mathcal{G}_{u,x}^{|U|,|X|}$ agree on all equality-free sentences – even ones involving negation – by Proposition 21, as there is a full surjective homomorphism from $\mathcal{G}_{u,x}^{|U|,|X|}$ to $\mathcal{G}_{1,3}^{2,2}$.

We will prove that $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{G}_{1,3}^{2,2})$ is Pspace-hard, by reduction from the Pspace-complete problem QCSP$(\mathcal{B}_{\mathrm{NAE}})$. Recall that we may assume w.l.o.g. that universal variables are relativised to U and that existential variables are relativised to X, by Theorem 30. Let φ be an instance of QCSP$(\mathcal{B}_{\mathrm{NAE}})$. We reduce φ to a (relativised) instance ψ of $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{G}_{1,3}^{2,2})$. The reduction goes as follows:

- an existential variable $\exists x$ of φ is replaced by an existential variable $\exists v_x \in X$ in ψ;

- a universal variable $\forall u$ of φ is replaced by $\forall u \in U, \exists v_u \in X, \ E(u, v_u)$ in ψ; and,

- every clause $C_i := R(\alpha, \beta, \gamma)$ in φ is replaced by the following formula in ψ,

$$\forall c_i \in U, \ E(c_i, v_\alpha) \vee E(c_i, v_\beta) \vee E(c_i, v_\gamma).$$

The truth assignment is read from $\exists$ choices in X for the variables v: we arbitrarily see one value in X as true and the other as false. It is not relevant which one is which for the problem not-all-equal satisfiability, we only need to ensure that no three variables involved in a clause can get the same value. The $\forall c_i \in U$ acts as a conjunction, enforcing "one of $v_\alpha, v_\beta, v_\gamma$ is true" and "one of $v_\alpha, v_\beta, v_\gamma$ is false". This means that at least one in three has a different value.

Now, we can prove that $\{\exists, \forall, \wedge, \vee\}$-FO$(\widetilde{\mathcal{D}})$ is Pspace-complete by substituting $R(u_0, x_0, u, v)$ for each instance of $E(u, v)$ in the previous proof, and by quantifying the sentence so-produced with the prefix $\forall u_0 \in U, \exists x_0 \in X$, once u_0 and x_0 are chosen, play proceeds as above but in the copy $\mathcal{G}_{u_0,x_0}^{|U|,|X|}$, and the result follows. $\qquad\square$

5.4 The Complexity of the Meta-Problem

The $\{\exists, \forall, \wedge, \vee\}$-FO$(\sigma)$ meta-problem takes as input a finite σ-structure $\mathcal{D}$ and answers L, NP-complete, co-NP-complete or Pspace-complete, according to the complexity of $\{\exists, \forall, \wedge, \vee\}$-FO$(\mathcal{D})$. The principle result of this section is that this problem is NP-hard even for some fixed and finite signature σ_0.[2]

Note that one may determine if a given shop f is a surjective hyper-endomorphism of a structure $\mathcal{D}$ in, say, quadratic time in $|D|$ Since we are not interested here in distinguishing levels within P, we will henceforth consider such a test to be a basic operation. We begin with the most straightforward case.

Proposition 83. *On input $\mathcal{D}$, the question "is $\{\exists, \forall, \wedge, \vee\}$-FO-$(\mathcal{D})$ in L?" is in P.*

---

[2]For now, σ_0 consists of two binary predicates and three monadic predicates. The monadic predicates are for convenience, but it is not clear whether a single binary predicate suffices.

Proof. By Theorem 66, we need to check whether there is both an A-shop and an E-shop in $\mathsf{shE}(\mathcal{D})$. In this special case, it suffices to test for each u, x in D, if the following $\{u\}$-$\{x\}$-shop f preserves $\mathcal{D}$: $f(u) := D$ and $f^{-1}(x) := D$. $\qquad\square$

Proposition 84. *For some fixed and finite signature σ_0, on input of a σ-structure $\mathcal{D}$, the question "is $\{\exists, \forall, \wedge, \vee\}$-FO($\mathcal{D}$) in NP (respectively, NP-complete, in co-NP, co-NP-complete)?" is NP-complete.*

Proof. The four variants are each in NP. For the first, one guesses and verifies that $\mathcal{D}$ has an A-shop, for the second, one further checks that there is no $\{u\}$-$\{x\}$-shop (see the proof of Proposition 83). Similarly for the third, one guesses and verifies that $\mathcal{D}$ has an E-shop; and, for the fourth, one further checks that there is no $\{u\}$-$\{x\}$-shop. The result then follows from Theorem 66.

For NP-hardness we will address the first problem only. The same proof will work for the second (for the third and fourth, recall that a structure $\mathcal{D}$ has an A-she iff its complement $\overline{\mathcal{D}}$ has an E-she). We reduce from *graph 3-colourability*. Let $\mathcal{G}$ be an undirected graph with vertices $V := \{v_1, v_2, \ldots, v_s\}$. We will build a structure $\mathcal{S}_{\mathcal{G}}$ over the domain D which consists of the disjoint union of "three colours" $\{0, 1, 2\}$, u, and the "vertices" from V.

The key observation is that there is a structure $\mathcal{G}_V$ whose class of surjective hyper-endomorphisms $\mathsf{shE}(\mathcal{G}_V)$ is generated by the following A-shop:

$$f_V := \begin{array}{c|l} 0 & 0 \\ \hline 1 & 1 \\ \hline 2 & 2 \\ \hline u & 0, 1, 2, u, v_1, \ldots, v_s \\ \hline v_1 & 0, 1, 2 \\ \hline v_2 & 0, 1, 2 \\ \hline \vdots & \vdots \\ \hline v_s & 0, 1, 2 \end{array}$$

The existence of such a $\mathcal{G}_V$ is in fact guaranteed by the Galois connection, fully given in [77], but that may require relations of unbounded arity, and we wish to establish our result for a fixed signature. So we will appeal to Lemma 85, below, for a σ_V-structure $\mathcal{G}_V$ with the desired class of surjective hyper-endomorphisms, where the signature σ_V consists of one binary relation and three monadic predicates. The signature σ_0 will be σ_V together with a binary relational symbol E.

The *structure* $\mathcal{S}_{\mathcal{G}}$ is defined as in $\mathcal{G}_V$ for symbols in σ_V, and for the additional binary symbol E, as the edge relation of the instance $\mathcal{G}$ of 3-colourability together with a clique $\mathcal{K}_3$ for the colours $\{0, 1, 2\}$. By construction, the following holds.

- Any she g of $\mathcal{S}_{\mathcal{G}}$ will be a subshop of f_V.
- Restricting such a shop g to V provides a set of mutually consistent 3-colourings: i.e. we may pick arbitrarily a colour from $g(v_i)$ to get a 3-colouring $\tilde{g}$. If there is an edge between v_i and v_j in $\mathcal{G}$, then $E(v_i, v_j)$ holds in $\mathcal{S}_{\mathcal{G}}$. Since g is a shop, for any pair of colours c_i, c_j, where $c_i \in g(v_i)$ and $c_j \in g(v_j)$, we must have that $E(c_i, c_j)$ holds in $\mathcal{S}_{\mathcal{G}}$. The relation E is defined as $\mathcal{K}_3$ over the colours. Hence $c_i \neq c_j$ and we are done.

- Conversely, a 3-colouring $\tilde{g}$ induces a sub-shop g of f_V: set g as f_V over elements from $\{0, 1, 2, u\}$ and as $\tilde{g}$ over V. The detailed argument is similar to the above.

This proves that graph 3-colourability reduces to the meta-question "is $\{\exists, \forall, \wedge, \vee\}$-FO($\mathcal{D}$) in NP". $\qquad\square$

Note that it follows from the given proof that the meta-problem itself is NP-hard. To see this, we take the structure $\mathcal{S}_G$ from the proof of Proposition 84 and ask which of the four classes L, NP-complete, co-NP-complete or Pspace-complete the corresponding problem belongs to. If the answer is NP-complete then $\mathcal{G}$ was 3-colourable; otherwise the answer is Pspace-complete and $\mathcal{G}$ was not 3-colourable.

Lemma 85. *Let σ_V be a signature involving one binary relations E' and three monadic predicates* Zero, One *and* Two. *There is a σ_V-structure $\mathcal{G}_V$ such that* $\mathsf{sh}\,\mathsf{E}(\mathcal{G}_V) = \langle f_V \rangle$.

Proof. We begin with the graph $\mathcal{G}'$ on signature $\langle E' \rangle$, depicted on Figure 5.3a. Note that

$$\mathsf{sh}\,\mathsf{E}(\mathcal{G}) := \langle\; \begin{array}{c|c} c & c \\ \hline u & c, u, v \\ \hline v & c \end{array} \;\rangle.$$

We now replace c by $\{0, 1, 2\}$ and v by V to obtain a graph $\mathcal{G}''$. Formally, this graph

(a) $\mathcal{G}'$ (b) $\mathcal{G}''$
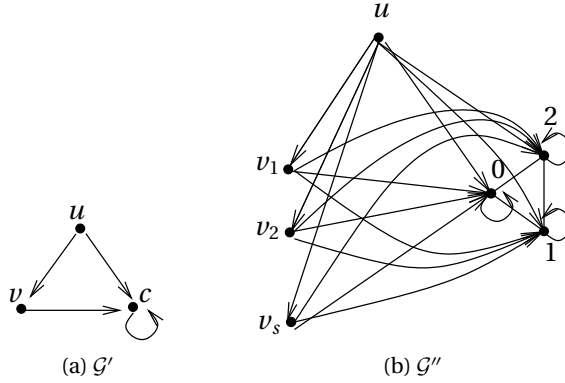
Figure 5.3: Building a structure with $\mathsf{sh}\,\mathsf{E}(\mathcal{G}_V) = \langle f_V \rangle$.

is the unique graph with domain $\{0, 1, 2, u\} \cup V$ such that the mapping which maps $\{0, 1, 2\}$ to c, fixes u and maps V to v, is a strong surjective homomorphism. By construction,

$$\mathsf{sh}\,\mathsf{E}(\mathcal{G}'') := \langle\; \begin{array}{c|c} 0 & 0, 1, 2 \\ \hline 1 & 0, 1, 2 \\ \hline 2 & 0, 1, 2 \\ \hline u & 0, 1, 2, u, v_1, \ldots, v_s \\ \hline v_1 & 0, 1, 2 \\ \hline \vdots & \vdots \\ \hline v_s & 0, 1, 2 \end{array} \;\rangle.$$

We now build $\mathcal{G}_V$ as the structure with binary relation E' which is the edge relation from $\mathcal{G}''$ and by setting the unary predicates as follows: Zero holds only over 0, One holds only over 1 and Two holds only over 2. This effectively fixes surjective hyper-endomorphisms to act as the identity over the colours $\{0, 1, 2\}$ as required. $\qquad \square$

6. Conclusion

We have classified the complexity of the model checking problem for all fragments of FO but those corresponding to the CSP and the QCSP. Our results are summarised on Figure 6.

6.1 The CSP dichotomy conjecture

For the CSP, the dichotomy conjecture has been proved in the Boolean case by Schaefer (see Theorem 51) and in the case of undirected graphs.

Theorem 86 ([52]). *Let $\mathcal{G}$ be an undirected graph. If $\mathcal{G}$ is bipartite then CSP($\mathcal{G}$) is in* L*, otherwise CSP($\mathcal{G}$) is* NP*-complete.*[1]

For CSP in general, it would suffice to settle the dichotomy conjecture for (certain) directed graphs [44]. The dichotomy conjecture has been settled for smooth digraphs (graphs with no sources and no sinks) [2]. According to the algebraic reformulation of the dichotomy conjecture, it would suffice to prove that every structure that has a *Sigger's term* has a tractable CSP (see [15, 18] for recent surveys on the algebraic approach to the dichotomy conjecture).

6.2 A QCSP tetrachotomy?

For the QCSP, much less is known. We have already seen that a dichotomy between P and Pspace-complete holds in the Boolean case (Theorem 52). However, the complexity is not even known for undirected graphs. It is fully classified for graphs with at most one cycle.

Theorem 87 ([79]). *Let $\mathcal{G}$ be an undirected graph.*

- *If $\mathcal{G}$ is bipartite then QCSP($\mathcal{G}$) is in* L*;*

- *if $\mathcal{G}$ is not bipartite and not connected then QCSP($\mathcal{G}$) is* NP*-complete; and,*

- *if $\mathcal{G}$ not bipartite, connected and contains at most one cycle then QCSP($\mathcal{G}$) is* Pspace*-complete.*

---

[1]In the bipartite case, assuming that the graph $\mathcal{G}$ has at least one edge, then the core of $\mathcal{G}$ is $\mathcal{K}_2$. The problem CSP($\mathcal{K}_2$) is 2-colourability which is in the complexity class *symmetric logspace* now known to be equal to L [91].

Figure 6.1: Classification of the complexity of the model-checking problem

The algebraic approach to QCSP uses *surjective polymorphisms* and has led to a trichotomy in the case where all graphs of permutations are available. Recall first the definition of some special surjective operations. A k-ary *near-unanimity operation f* satisfies

$$f(x_1,\ldots,x_k) = \begin{cases} x & \text{if } \{x_1,\ldots,x_k\} = \{x\}; \text{ and,} \\ x & \text{if all but one of } x_1,\ldots,x_k \text{ is equal to } x. \end{cases}$$

When $k = 3$, we speak of a *majority operation*. The k-ary *near projection operation* is defined as

$$l_k(x_1,\ldots,x_k) = \begin{cases} x_1 & \text{when } |\{x_1,\ldots,x_k\}| = k; \text{ and,} \\ x_k & \text{otherwise.} \end{cases}$$

The *ternary switching operation* is defined as

$$s(x,y,z) = \begin{cases} x & \text{if } y = z, \\ y & \text{if } x = z, \\ z & \text{otherwise.} \end{cases}$$

The *dual discriminator operation* is defined as

$$d(x,y,z) = \begin{cases} y & \text{if } y = z; \text{ and,} \\ x & \text{otherwise.} \end{cases}$$

When $f(x,y,z) = x - y + z$ w.r.t. some Abelian group structure, we say that f is an *affine operation*.

Theorem 88 ([10]). *Let $\mathcal{D}$ be a structure such that there is an extensional binary symbol for each graph of a permutation of D. Then the complexity of QCSP($\mathcal{D}$) follows the following trichotomy.*

- *If $\mathcal{D}$ has a surjective polymorphism which is the dual discriminator, the switching operation or an affine operation then QCSP($\mathcal{D}$) is in P.*

- *Else, if $l_{|D|}$ is a surjective polymorphism of $\mathcal{D}$ then QCSP($\mathcal{D}$) is NP-complete.*

- *Otherwise, QCSP($\mathcal{D}$) is Pspace-complete.*

In general, it is known that if a structure $\mathcal{D}$ is preserved by a *near-unanimity operation* then QCSP($\mathcal{D}$) is in P, because it implies a property of *collapsibility*. This property means that an instance holds if, and only, if all sentences induced by keeping only a bounded number of universal quantifiers – the so-called *collapsings* – hold [22].

For undirected partially reflexive graphs (*i.e.* with possible self-loops), we have the following partial classification (reformulated algebraically).

Theorem 89 ([78]). *Let $\mathcal{T}$ be a partially reflexive forest.*

- *If $\mathcal{T}$ is $\{\exists, \forall, \wedge\}$-FO-equivalent to a structure that is preserved by a majority operation then QCSP($\mathcal{T}$) is in* P*; and,*

- *otherwise, QCSP($\mathcal{T}$) is* NP-*hard.*

In the case of structures with all constants, Hubie Chen has ventured some conjecture regarding the NP/Pspace-hard border: he suggests that the *polynomially generated power property (PGP)* – a property which generalises collapsibility – explains a drop in complexity to NP. He also pointed out that the graphs involved in the first two cases of Theorem 87, namely bipartite graphs and disconnected graphs, are collapsible. He does so by exhibiting certain surjective polymorphisms for these graphs, known to imply collapsibility (see [23] for details).

6.3 Some questions

Currently the notion which should lead to tractability or at least a drop in complexity to NP of QCSP in the presence of constants (which could be construed as tractability since this problem is Pspace-complete in general) and seems the most amenable to analysis is that of *collapsibility*. Let us say that a structure $\mathcal{A}$ is n-collapsible to a set of elements C if the following holds: for any formula φ in $\{\exists, \forall, \wedge\}$-FO $\mathcal{A} \models \varphi$ iff for any c in C, for any formula ψ induced by replacing in each universal block of φ all but n variables by the constant c, we have $\mathcal{A} \models \psi$. A structure $\mathcal{A}$ is collapsible if it is n-collapsible for some n and some $C \subseteq A$.

Question 90. *Is collapsibility of a structure decidable?*

We know that 0-collapsibility (to a singleton) is characterised by having an A-shop. One can also wonder if there is a special kind of surjective polymorphism which characterises collapsibility. The surjective polymorphisms exhibited by Chen to show n-collapsibility w.r.t. some constant c of bipartite and disconnected graphs is an $n+1$-ary surjective idempotent polymorphisms such that any n-ary polymorphism induced by setting one coordinate to c is surjective.

Question 91. *Can collapsibility be characterised algebraically?*

Another important issue is to understand the fundamental nature of Q-cores.

Question 92. *What kind of properties can one infer on $\mathcal{A}$ when a Q-core of $\mathcal{A}$ satisfies a certain tractability condition?*

The Q-core is a combinatorial object worthy of independent studies, and we have left several questions unanswered.

Questions 93. *Is the Q-core unique up to isomorphism? Is the strong Q-core unique up to isomorphism? If they are, do they coincide, i.e. is the strong $\{\exists, \forall, \wedge\}$-FO-core of $\mathcal{A}$ a substructure of $\mathcal{A}$?*

There are also issues regarding the method in which it can be computed. The U-X-core may be computed relatively efficiently (the decision problem is DP-complete[2] [72]) and could be used as a preprocessing step before search in a QCSP solver augmented with disjunction. For the Q-core, the theoretical bounds are quite large. I have supervised Shwetha Raghuraman in summer 2012, who has implemented a tool to compute various $\mathscr{L}$-cores using Cplex. The aim was to run some experiments to check whether the Q-core or the strong Q-core were unique up to isomorphism, and get a better idea of the genuine bound on the power one needs to take to check for equivalence. One can not check anything non trivial (*i.e.* that we did not know already from inspection of small graphs) regarding these bounds as otherwise one would exceed the size of integers (32 or 64 bytes). This means that we had to stick to homomorphisms from a power 2 or 3. However, we iterated the reduction w.r.t. such small powers until a fixed point was reached. We have not really checked for uniqueness but in all examples we have either obtained the unique Q-core that we had found "by hand"; or, using a power 2 or 3 provided us with isomorphic results.

So let us define *an iterated power r Q-core $\mathcal{B}$* of a structure $\mathcal{A}$ has a smallest substructure of $\mathcal{A}$ such that there are sequences of structures $\mathcal{C}_0 := \mathcal{A}, \mathcal{C}_1, \ldots, \mathcal{C}_l := \mathcal{B}$ such that for every $0 \leq i < l$, there are surjective homomorphisms p_i from $(\mathcal{C}_i)^r$ to $\mathcal{C}_{i+1}$ and q_i from $(\mathcal{C}_{i+1})^r$ to $\mathcal{C}_i$.

Questions 94. *What bound in the exponent is sufficient when computing the Q-core? Is iterated square enough, i.e. does the iterated power 2 Q-core coincide with the Q-core?*

A positive answer would mean that deciding whether a structure is a Q-core is in DP.

---

[2]A problem in DP is the union of a problem in NP and a problem in co-NP.

PART II

DESCRIPTIVE COMPLEXITY OF THE CONSTRAINT SATISFACTION PROBLEM
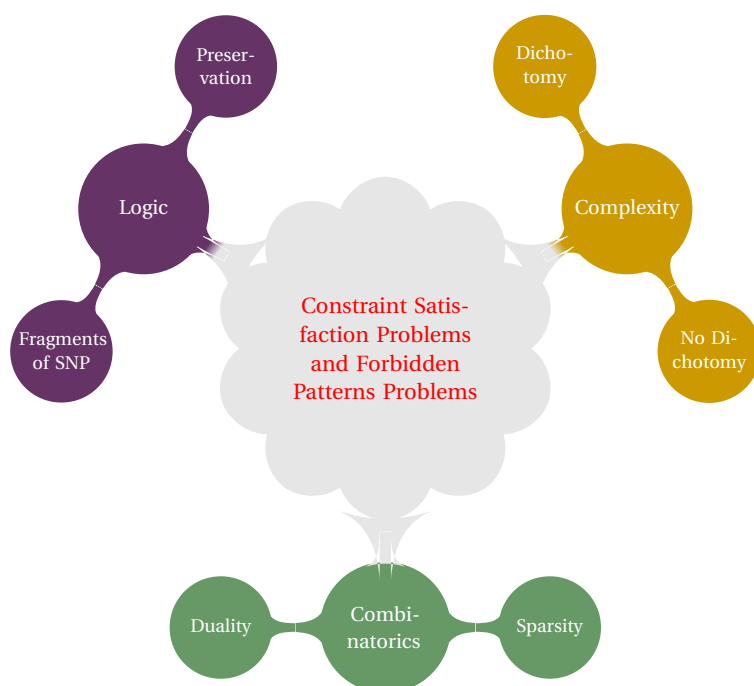
Feder and Vardi showed that the logic monotone monadic SNP without $\neq$ is intimately linked with the class of non-uniform Constraint Satisfaction Problems. This logic exhibits a dichotomy if and only if the dichotomy conjecture holds. This logic captures problems that can be expressed as finite union of forbidden patterns problems. These forbidden patterns problems are in general (well-behaved) infinite domain Constraint Satisfaction Problems. One can decide when they have a finite domain and, provided that they are given in a normal form, one can decide their containment. In the special case when these problems are first-order expressible, related questions of homomorphism dualities have been studied in Combinatorics. Such first-order problems arise when one consider first-order expressible problems that are preserved by inverse homomorphisms. These results from Combinatorics and logic can be lifted to obtain interesting results for problems that are not first-order expressible. In particular, when restricted to sufficiently sparse structures, forbidden patterns problems becomes finite domain constraint satisfaction problems.

7. Feder and Vardi's logic

Descriptive complexity theory seeks to classify problems, *i.e.*, classes of finite structures, as to whether they can be defined using formulae of some specific logic, in relation to their computational complexity. One of the seminal results in descriptive complexity is Fagin's theorem [42] which states that a problem can be defined in existential second-order logic (ESO) if, and only if, it is in the complexity class NP(throughout we equate a logic with the class of problems definable by the sentences of that logic). In their influential paper where they conjectured the dichotomy of CSP [44], Feder and Vardi also introduced the logic MMSNP, a fragment of ESO which is intimately linked to CSP. This logic is a good candidate for the role of the largest syntactic fragment of ESO to exhibit a dichotomy (between P and NP-complete). Indeed, Feder and Vardi proved that

- Any larger syntactic fragment can not exhibit a dichotomy; and,

- MMSNP exhibits a dichotomy if, and only if, the dichotomy conjecture for CSP holds.

The logic MMSNP does not capture CSP as such: every problem in CSP can be defined in MMSNP but there are problems in MMSNP which are not in CSP [44, 73]. In previous work with Iain Stewart [64, 74], we provided an effective method to decide given a sentence of MMSNP whether it defines a problem in CSP or not. It turns out that these problems in MMSNP that are not in CSP are actually *constraint satisfaction problems with an infinite domain*, whose templates have nice model theoretic properties, introduced by Bodirsky (see his habilitation thesis for recent developments [6]). So our previous result provides in fact a decision procedure that can tell whether a sentence of MMSNP defines a finite or an infinite CSP problem. In contrast, when the input of a problem definable in MMSNP is restricted to be of bounded degree, or from a proper minor closed class or more generally of bounded expansion, the restricted problem becomes a restricted *finite* CSP [68]. It is important to note that though there are infinite CSP *à la* Bodirsky which are not definable in MMSNP, this logic defines a large infinite class of natural infinite CSP which are worth studying in their own rights. For example, the complexity of problems in MMSNP have recently been investigated in the special case of monochromatic and loopless forbidden patterns [7].

In this chapter, we introduce MMSNP in some details. In Chapter 9, we shall see that the question of problems in MMSNP that are CSP has been investigated in Combinatorics under the name of (homomorphism) dualities.

7.1 Preliminaries

Existential Second Order Logic.   Fagin's theorem [42] equates definability in ESO with membership in the complexity class NP. For example, the class of 3-colourable graphs can be defined using a sentence of the following form.

$$\Phi_1 := \exists R, G, B, \text{ three sets partitioning the vertices}$$
$$\forall x, y, \neg\big(E(x,y) \wedge R(x) \wedge R(y)\big) \wedge \neg\big(E(x,y) \wedge G(x) \wedge G(y)\big)$$
$$\wedge \neg\big(E(x,y) \wedge B(x) \wedge B(y)\big)$$

A graph is represented as a relational structure whose domain consists of vertices equipped with a single binary predicate E representing the edge relation. The above sentence has two kinds of quantifiers: second-order predicates (always upper-case) which are interpreted as relations, like R which is interpreted as a set of vertices, and first-order variables (always lower case), like x, which is interpreted as a vertex. The three second order predicates R, G and B stand for three colours, say red, green and blue and the sentence asserts that the vertices may be coloured with these three colours in such a way that for every edge in the graph, the extremities have different colours.

   In this part, we shall mostly need second-order predicates that are sets, the so-called *monadic* predicates, and we shall only allow them to be existentially quantified as in the above example. Note that finitely many sets of vertices correspond essentially to a partition of the vertices in distinct *colours*. In combinatorial terms this means that in order to check a property we have to guess some colours for each vertex before verifying some first-order property over the coloured graph. Let us clarify this with another example.

$$\Phi_2 := \exists M, N \; \forall x, y, \; \neg(\neg M(x) \wedge \neg N(x))$$
$$\wedge \neg\big(E(x,y) \wedge M(x) \wedge N(x) \wedge M(y) \wedge N(y)\big)$$
$$\wedge \neg\big(E(x,y) \wedge \neg M(x) \wedge N(x) \wedge \neg M(y) \wedge N(y)\big)$$
$$\wedge \neg\big(E(x,y) \wedge M(x) \wedge \neg N(x) \wedge M(y) \wedge \neg N(y)\big)$$

There are two monadic predicates M and N in Φ_2 and for a given vertex x there are four cases to consider: x is in both M and N ($M(x) \wedge N(x)$ holds), x is in M but not in N ($M(x) \wedge \neg N(x)$ holds) etc. So the above sentence disallows one of the colour (with the conjunct $\neg(\neg M(x) \wedge \neg N(x))$) and states for the three other colours that an edge can not have both extremities of the same colour. In other words, this sentence defines also the fact that a graph is 3-colourable.

Monotone Monadic Strict NP *without inequalities.*   The two sentences Φ_1 and Φ_2 have a particular syntactic form:

 $\exists$ monadic predicates, $\forall$ variables ranging over vertices,   a first-order sentence.

Such sentences form the fragment SNP of ESO. It turns out that many combinatorial problems are definable in SNP, in particular every problem in CSP can be

defined by a SNP sentence. For example, in the case of 3-colourability, we may use the sentence Φ_2. Let us explain in a bit more detail how we may build this sentence in a systematic fashion. Recall first that for a CSP with template $\mathcal{T}$, a structure $\mathcal{A}$ is a yes-instance if, and only if, there exists a *homomorphism* from $\mathcal{A}$ to $\mathcal{T}$. That is, a mapping h from the domain of $\mathcal{A}$ to that of $\mathcal{T}$ such that every arc in $\mathcal{A}$ is mapped to an arc in $\mathcal{T}$ (assuming we deal with digraphs for now for the sake of simplicity). The 3-colourability problem, recast as a digraph problem, has as template $\mathcal{T}$ the digraph with 3 vertices and all possible arcs that are not self-loops. Viewing the 3 elements of $\mathcal{T}$ as colours, we have readily explained how to use 2 monadic predicates M and N and one forbidden combination of them $\neg(\neg M(x) \wedge \neg N(x))$ to encode three colours. In order to enforce a homomorphism, we now encode the non-arcs of $\mathcal{T}$ by adding negated conjuncts, one for each non-arc. For example, if $M(x) \wedge N(x)$ stands for the colour corresponding to the first vertex of $\mathcal{T}$ and since there is no self-loop around this vertex, we add the following negated conjunct to the sentence:

$$\neg\big(E(x,y) \wedge M(x) \wedge N(x) \wedge M(y) \wedge N(y)\big).$$

Doing this with every non-arc, we obtain the sentence Φ_2 given above. It is important to note that the sentence we build this way uses only *monadic* predicates. Furthermore, the first-order part is a conjunction of negated conjuncts; and, in every negated conjunct atoms from the input (the edge symbol E in our examples) appears always positively. This means that the sentence is *monotone*. Finally, we never use the symbol $\neq$. We have therefore built a sentence of SNP that is *monadic, monotone* and *without inequality*. The sentences of SNP satisfying these three restrictions form the logic MMSNP introduced by Feder and Vardi. As we may build such a sentence for every template, we now know that

$$\text{CSP} \subseteq \text{MMSNP}.$$

Some sentences of MMSNP give rise to problems that are not in CSP and are in fact constraint satisfaction problems with an infinite template. For example,

$$\Psi_1 := \forall x, \forall y, \forall z, \ \neg\big(E(x,y) \wedge E(y,z) \wedge E(z,x)\big)$$

expresses that there are no oriented 3-cycles in a digraph (and also no self-loop as the variables may be equal). It is not difficult to see that this problem is not in CSP. Assume for contradiction that there exists a template $\mathcal{T}$ with n elements for this problem. We may build a yes-instance $\mathcal{A}$ for ψ_1 as follows: take $n+1$ vertices and add between any pair of distinct vertices a directed path of length 3. By assumption, there exists a homomorphism from $\mathcal{A}$ to $\mathcal{T}$. This homomorphism must identify two distinct elements joined by a directed 3-path. Hence, $\mathcal{T}$ contains a loop or an oriented 3-cycle and is a no-instance which is absurd as the template is always a yes-instance.

The problem defined by Ψ_1 is in fact a CSP with an infinite template. It is not difficult to construct an infinite template for this problem: simply take as a tem-

plate the disjoint union of its yes-instances[1]. This infinite template is not particularly interesting, however, we may also construct for this problem an infinite template that has a nice model theoretic property called *ω-categoricacity*. From now on, by infinite CSP, we mean a problem with such a nice template. This property means in particular that the Galois-connection used in the finite case can be successfully adapted and that some logico-algorithmic results such as those involving Datalog still hold. We will refrain from going into more details and refer to Bodirsky's habilitation thesis [6] on his pioneering work on infinite CSP.

7.2 Dichotomy and descriptive complexity

We say that two classes of problem $\mathscr{P}$ and $\mathscr{P}'$ are *computationally equivalent* w.r.t. some notion of reduction if, and only if,

- for any problem Ω in $\mathscr{P}$ there exists a problem Ω' in $\mathscr{P}'$ such that Ω reduces to Ω' and Ω' reduces to Ω; and,

- conversely for any problem Ω' in $\mathscr{P}'$ there exists a problem Ω in $\mathscr{P}$ such that Ω' reduces to Ω and Ω reduces to Ω'.

In many cases, we will consider $\mathscr{P}' \subseteq \mathscr{P}$ and will only need to consider the first point.

Theorem 95 (Feder and Vardi [44])**.**

- NP *is computationally equivalent (w.r.t. polynomial-time reduction) to the logic* monotone monadic SNP with $\neq$.

- NP *is computationally equivalent (w.r.t. polynomial-time reduction) to the logic* monadic SNP without $\neq$.

- NP *is computationally equivalent (w.r.t. polynomial-time reduction) to the logic* monotone SNP with $\neq$.

proof (sketch). Feder and Vardi use the fact that monadic Datalog with $\neq$ but without negation can verify a polynomial-time encoding of a Turing machine computation. Rejection of this Datalog program is readily seen as a sentence of monotone SNP with $\neq$. Almost all existential predicates are monadic but for the need to describe the movement of the head of the Turing machine. Feder and Vardi circumvent this issue by assuming that the Turing machine is *oblivious* (a polynomial-time Turing machine can be simulated by an oblivious one that runs also in polynomial-time [94]). This means that the head movement is independent of the input and means that a suitable sentence of monotone monadic SNP with $\neq$ expresses the original problem.

Next, Feder and Vardi show that every sentence of monotone monadic SNP with $\neq$ is computationally equivalent to a sentence of monadic SNP without $\neq$. They do this by adding a binary predicate to the input that will play the role of $\neq$.

---

[1]This is true in general for any monotone problem that is closed under disjoint union.

This binary predicate is required to be an equivalence relation with the property that if an input or existential monadic relation holds on some elements, then it also holds when an element in an argument position is replaced by an element related to it under this binary predicate (this is precisely the equivalence relation $\sim$ from § 3.4).

Finally, they hint that every sentence of monotone monadic SNP with $\neq$ is computationally equivalent to a sentence of monotone SNP without $\neq$. The construction is somewhat more involved. The basic idea is that if this input structure were linearly ordered, say with a binary successor succ, then one could define equality as an existential predicate that is required to be an equivalence relation (this is not monotone but is irrelevant as the predicate is existential) such that it is not the case that $\text{succ}^\star(x, y) \wedge \text{eq}(x, y)$ where $\text{succ}^\star$ is the transitive closure of succ (this can be enforced in a monotone fashion w.r.t. the input predicate succ). However, it is not possible to check in a monotone fashion that a binary relation is a genuine successor. The idea is to settle for a "pseudo-successor", i.e. a layered directed acyclic graph (in each connected component of the directed acyclic graph, every directed path from a vertex to another has the same length). Adding a source vertex (with an input monadic predicate, intuitively some special element on the pseudo-successor) one can walk along the pseudo-successor and force vertices on the same layer to be equivalent under eq. Additional tricks mean that the formula is actually behaving as the original formula would on the linearly ordered structure obtained by taking the homomorphic image which identifies every element on the same layer. □

Together with Ladner's theorem [60], which implies that there are intermediate problems in NP that are neither in P nor NP-complete (assuming $P \subsetneq NP$) we have the following.

Corollary 96. *Assuming that* $P \subsetneq NP$, *the logics* monotone monadic SNP with $\neq$, monadic SNP without $\neq$ *and* monotone SNP with $\neq$ *do not have a dichotomy between* P *and* NP*-complete.*

Unlike the above three larger syntactic fragments of SNP, the logic MMSNP follows a dichotomy (assuming the dichotomy conjecture holds).

Theorem 97 (Feder and Vardi [44], Kun [58])**.** *The logic MMSNP is computationally equivalent (w.r.t. polynomial time reduction) to CSP.*

proof (sketch). Since CSP $\subseteq$ MMSNP, we only need to represent a problem in MMSNP by one in CSP. Each kind of conjunction of input which appears in a negated conjunct of the MMSNP sentence will give rise to a relational symbol in the corresponding CSP and the reduction from MMSNP to the corresponding CSP is straightforward (*e.g.* triangles will be encoded via a ternary predicate). The difficulty is to come up with a suitable reduction in the other direction. Feder and Vardi's proof uses a normal form for MMSNP, which ensures in particular that (the structure associated with) every negated conjunct is biconnected. This means that

no dichotomy

ESO=NP

| monadic SNP without $\neq$ | monotone monadic SNP with $\neq$ | monotone SNP without $\neq$ |

monotone monadic
SNP
without $\neq$
(MMSNP)

CSP

dichotomy conjecture
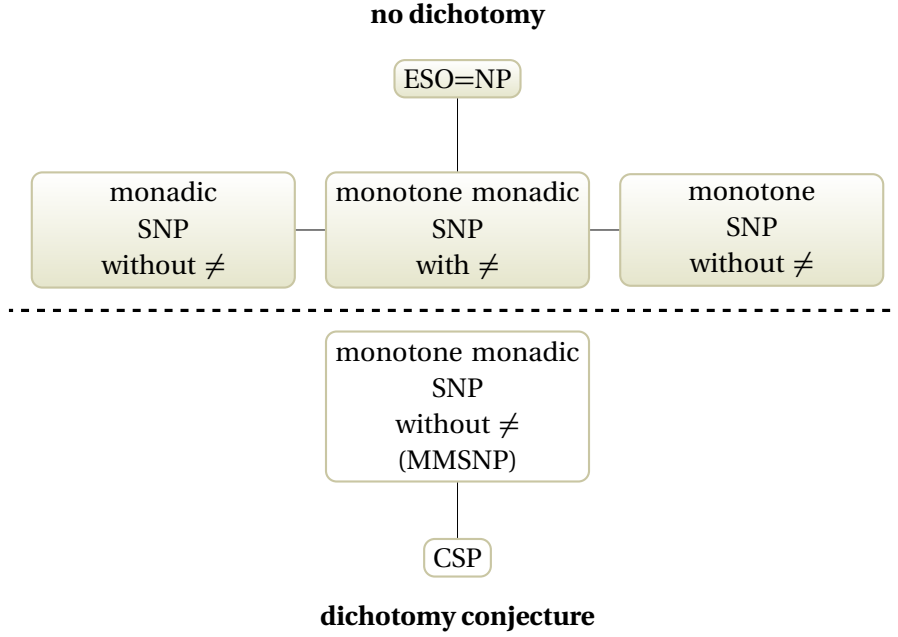
Figure 7.1: Logic and the dichotomy conjecture. The classes of problems are drawn according to inclusion from top to bottom. An edge between two classes indicates computational equivalence (other edges may be deduced by transitivity).

when reducing from the CSP to the MMSNP, if one does the converse translation (*e.g.* a ternary predicate induces the three edges of a triangle), provided that the structure has no short cycles (in our triangle example, no cycle involving three ternary tuples, for a suitable notion of cycle) then one has a correct reduction. A problem may arise if there are short cycles. Recall that the *girth* is the size of the shortest cycle. Feder and Vardi appealed to a generalisation to structures of a randomised reduction due to Erdös for graphs, which for a fixed bound on the template size and a fixed girth, will compute for every structure, a structure of high girth that is homomorphic to a template iff it was the case for the original structure (see Lemma 119 on page 98 for details). Thus, Feder and Vardi's original reduction was polynomial time but randomised. Kun has proposed a deterministic form of this construction using graph expanders [58]. □

7.3 Combinatorial view of MMSNP.

Recall the formula Ψ_1 of MMSNP: $\Psi_1 := \forall x, \forall y, \forall z, \neg\big(E(x,y) \wedge E(y,z) \wedge E(z,x)\big)$. Note that the negated conjunct $\neg\big(E(x,y) \wedge E(y,z) \wedge E(z,x)\big)$ n Ψ_1 essentially forbids the occurrence of an oriented 3-cycle. However, since the variables x, y and z may take the same value, this means in fact that we forbid the existence of a

homomorphism from the oriented 3-cycle to the instance. Hence, the problem defined by Ψ_1 can be seen as a dual problem to a CSP. Whereas in the case of CSP we ask whether there is a homomorphism from the instance $\mathcal{A}$ to the template $\mathcal{T}$, we will ask here whether there is no homomorphism from an obstruction $\mathcal{F}$ to the instance $\mathcal{A}$. In the case of more than one obstruction, we have essentially the fragment of MMSNP that has no monadic predicate (sentences of MMSNP that are also first-order). In general, such a problem is known to be an infinite CSP with an ω-categorical template, provided that the obstructions are connected (it is a corollary of [25] as pointed out in [8]).

Another example of a problem that is in MMSNP but not in CSP is:

$$\Psi_2 := \exists M, \, \forall x, y, z, \, \neg\big(E(x,y) \wedge E(y,z) \wedge E(z,x) \wedge M(x) \wedge M(y) \wedge M(z)\big)$$
$$\wedge \neg\big(E(x,y) \wedge E(y,z) \wedge E(z,x) \wedge \neg M(x) \wedge \neg M(y) \wedge \neg M(z)\big).$$

We have a single monadic predicate which encodes two colours, say white and black. The two negated conjuncts forbid two vertex-coloured structures, namely a white oriented 3-cycle $\mathcal{F}_1'$ and a black oriented 3-cycle $\mathcal{F}_2'$. Thus, the problem defined by Ψ_2 accepts an instance $\mathcal{A}$ whenever its vertices can be coloured in white and black into a structure $\mathcal{A}'$ such that there is neither a homomorphism from $\mathcal{F}_1'$ to $\mathcal{A}'$ nor a homomorphism from $\mathcal{F}_1'$ to $\mathcal{A}'$.

Forbidden patterns problems.
In general a *forbidden patterns problem* Ω is given by a finite set of coloured structures, which we call a *representation*. We insist that each such structure is *connected* and *contains at least one tuple*. The connectivity assumption means that just like a CSP our problems will be close under disjoint union.

It makes sense to formalise the (vertex-)colouring of a structure by a homomorphism into some structure describing the colours. So Ω is given by a structure $\mathcal{T}$ representing the colours and a set $\mathscr{F}'$ of $\mathcal{T}$-coloured structures, the so-called *forbidden patterns*. We call the pair $(\mathscr{F}', \mathcal{T})$ a *representation* of the problem Ω.

A $\mathcal{T}$-coloured structure is a pair $(\mathcal{F}, f)$ where f is a homomorphism from $\mathcal{F}$ to $\mathcal{T}$ which describes the colouring. The notion of structure homomorphism generalises naturally to coloured structures: given two $\mathcal{T}$-coloured structures $(\mathcal{F}, f)$ and $(\mathcal{G}, g)$, a homomorphism h from $(\mathcal{F}, f)$ to $(\mathcal{G}, g)$ is simply a homomorphism from $\mathcal{F}$ to $\mathcal{G}$ that preserves the colours, that is such that $f = g \circ h$. An instance $\mathcal{A}$ of the problem Ω represented by $(\mathscr{F}', \mathcal{T})$ is a *yes-instance* if, and only if, there exists a homomorphism h from $\mathcal{A}$ to $\mathcal{T}$ such that there is no homomorphism from any forbidden pattern $(\mathcal{F}, f)$ in $\mathscr{F}'$ to $(\mathcal{A}, h)$. When h is not a homomorphism or when there is a homomorphism from some forbidden pattern $(\mathcal{F}, f)$ in $\mathscr{F}'$ to $(\mathcal{A}, h)$, we say that $(\mathcal{A}, h)$ is *not valid* (w.r.t. $(\mathscr{F}', \mathcal{T})$); and, otherwise that it is *valid* (w.r.t. $(\mathscr{F}', \mathcal{T})$). We denote by FPP$(\mathscr{F}', \mathcal{T})$ the forbidden patterns problem represented by $\mathscr{F}', \mathcal{T}$ and by FPP the class of forbidden patterns problems.

Forbidden patterns problems are known to be infinite CSP in the sense of Bodirsky. A countable structure $\mathcal{A}$ is ω-categorical if any structure countable $\mathcal{B}$ that satisfies the same first-order sentences as $\mathcal{A}$ is isomorphic to $\mathcal{A}$.

Theorem 98 ([8])**.** *Let Ω be a forbidden patterns problem. There exists an ω-categorical structure Γ such that $\Omega = CSP(\Gamma)$.*

proof (sketch). This result is a direct consequence of the same result for forbidden patterns problems that can be defined without additional existential monadic predicates (as our example Ψ_1) [25]. We sketch the proof from [8] for completeness.

Given a forbidden patterns problem given by a sentence Ψ of MMSNP, the idea is to consider the signature extended with two symbols for each existential monadic predicate M of Ψ, one for the positive occurrence M^+ and one for the negative occurrence M^- (this is because we sill shortly pretend that these monadic predicates are part of the input and will need the induced formula to be monotone). Then we consider the first-order MMSNP sentence ψ' induced naturally over this extended signature by replacing every positive occurrence of M by M^+ and every negative occurrence $\neg M$ by M^- in Ψ and dropping the existential second-order prefix of Ψ.

For example, if we apply this construction to Ψ_2 we would get the following first-order sentence.

$$\psi'_2 := \forall x, y, z, \neg\big(E(x,y) \wedge E(y,z) \wedge E(z,x) \wedge M^+(x) \wedge M^+(y) \wedge M^+(z)\big)$$
$$\wedge \neg\big(E(x,y) \wedge E(y,z) \wedge E(z,x) \wedge M^-(x) \wedge M^-(y) \wedge M^-(z)\big).$$

We consider the ω-categorical structure Γ' such that $\psi' = CSP(\Gamma')$ from [25]. Taking first the induced substructure Γ'_c of Γ' that contains only those vertices x for which for every pair of new monadic symbols, either $M^+(x) \wedge \neg M^-(x)$ holds or $\neg M^+(x) \wedge M^-(x)$ holds; and then the reduct to the original symbols, *i.e.* forgetting the new monadic symbols, we obtain a structure Γ such that $\Psi = CSP(\Gamma)$, by construction. Indeed, a homomorphism from a structure $\mathcal{A}$ to Γ induces witnesses for the monadic predicates by looking up the values of the additional monadic predicates in the extension Γ'_c of Γ. Conversely, suitable witnesses to the fact that $\mathcal{A} \models \Psi$ allow to extend $\mathcal{A}$ to $\mathcal{A}'$ such that $\mathcal{A}' \models \psi'$; so there exists a homomorphism from $\mathcal{A}'$ to Γ'_c; and, by taking reducts, a homomorphism from $\mathcal{A}$ to Γ.

The structure Γ is ω-categorical as it can be constructed via first-order interpretation from Γ', a transformation known to preserve ω-categoricity. $\square$

Since every sentence in MMSNP captures a finite union of forbidden patterns problems [74][2], we obtain the following corollary.

Corollary 99. *For every sentence Φ of MMSNP, there exists a finite set of forbidden patterns problems $\Omega_1, \Omega_2, \ldots, \Omega_n$ such that $\Phi := \bigcup_{1 \leq i \leq n} \Omega_i$ (viewing a decision problem as the set of its yes-instances). Moreover, there are (not necessarily finite) ω-categorical structures $\Gamma_1, \Gamma_2, \ldots, \Gamma_n$ such that for every $1 \leq i \leq n$, $\Omega_i := CSP(\Gamma_i)$. Consequently, $\Phi := \bigcup_{1 \leq i \leq n} CSP(\Gamma_i)$.*

---

[2]Note that this is only a consequence of our definition which insists that forbidden patterns are connected. Allowing for disconnected obstructions would allow to equate the two.

7.4 When are forbidden patterns problems constraint satisfaction problems?

In [64, 74], we refined the normal form for MMSNP that Feder and Vardi used in the proof of Theorem 97. We shall recall shortly what conditions our normal form entails and hint at how it can be enforced on an example. Let us first introduce some vocabulary. We say that a coloured structure is *weakly valid* w.r.t. a representation of a forbidden patterns problem if there is no *injective* homomorphism from a forbidden pattern into this coloured structure. A forbidden pattern that consists of a coloured structure with a single tuple that mentions each element exactly once[3] is said to be *conform*. When a forbidden pattern is conform, we may drop it from the list of forbidden patterns and enforce its constraint by amending the structure $\mathcal{T}$ accordingly (by removing the corresponding tuple from $\mathcal{T}$).

Combinatorially, a problem in CSP can be seen as a homomorphism problem represented by a finite structure $\mathcal{T}$, the so-called template. It is well known that the containment of CSP corresponds exactly to the existence of a homomorphism from one template to another. More precisely, the CSP with template $\mathcal{T}_1$ is contained in the CSP with template $\mathcal{T}_2$ if, and only if, there is a homomorphism from $\mathcal{T}_1$ to $\mathcal{T}_2$. Therefore the category of relational structures and homomorphisms crops up naturally in the study of CSP [53].

The key ingredient of our refinement of Feder and Vardi's normal form of MMSNP is to take into account the fact that some colours might actually be redundant in the representation of the problem. To formalise this, we introduced the notion of a *recolouring* from (the representation of) a forbidden patterns problem Ω_1 to (the representation of) another forbidden patterns problem Ω_2, which is simply a particular homomorphism which states how the colours of a problem Ω_1 can be transformed into the colours of a problem Ω_2. Such a recolouring implies that Ω_1 is contained in Ω_2 (see Proposition 115 on page 96). The converse does not hold in general but we shall see later that it does hold when the problems are given in our normal form (see Theorem 114), which suggests that representations of forbidden patterns problems given in a normal form and recolourings provide us with the right category in the context of MMSNP.

Let us recall the formal definition before looking at an example.

Definition 100 (recolouring [64, 74])**.** *Let $(\mathscr{F}_1', \mathcal{T}_1)$ and $(\mathscr{F}_2', \mathcal{T}_2)$ be two representations. A recolouring from $(\mathscr{F}_1', \mathcal{T}_1)$ to $(\mathscr{F}_2', \mathcal{T}_2)$ is a homomorphism from $\mathcal{T}_1$ to $\mathcal{T}_2$ such that for every $(\mathcal{F}_2, f_2)$ in $\mathscr{F}_2'$, any of its inverse image $(\mathcal{F}_2, f_1)$ under r is not valid w.r.t. $(\mathscr{F}_1', \mathcal{T}_1)$. By inverse image, we mean that $(\mathcal{F}_2, f_1)$ is a $\mathcal{T}_1$-coloured structure such that $f_2 = r \circ f_1$.*

Example 101. We consider the two forbidden patterns given on Figure 7.2 (note how the colours of the vertices of a forbidden pattern are simply given by labelling a vertex with its colour). The problem represented by $\mathcal{T}_2$ and $\mathscr{F}_2'$ is a variant of the problem defined by Ψ_2 in which triangles have arcs in both directions. Let r be

---

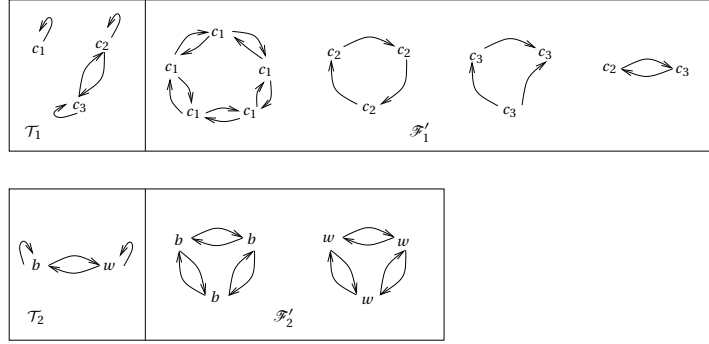[3]Self-loops and their generalisation like $R(x, x, y)$ are not conform.

Figure 7.2: two forbidden patterns problems.

the mapping from the colours of the first problem, namely $\{c_1, c_2, c_3\}$ to those of the second problem, namely $\{b, w\}$, that maps c_1, c_2 and c_3 to b. Note that r is indeed a homomorphism from $\mathcal{T}_1$ to $\mathcal{T}_2$. The only forbidden pattern of the second problem whose colours are in the image of r is the black triangle (the first forbidden pattern of $\mathscr{F}_2'$ listed on the figure). We need to show that every triangle whose vertices is coloured via r^{-1} is invalidated by the first problem. This can happen in two ways: the colouring may not be a homomorphism to $\mathcal{T}_1$, or some forbidden pattern in $\mathscr{F}_1'$ invalidates it. If the colours of the three vertices of the triangle are replaced by c_1, then the 5-cycle (the first forbidden pattern of $\mathscr{F}_1'$ listed on the figure) invalidates this choice of colours. Similarly, if the vertices are all coloured by c_2 only or c_3 only then the two next forbidden patterns on the figure invalidate these choices. If the colours of the three vertices of the triangle are replaced by c_1 and other colours then the colouring is not a homomorphism to $\mathcal{T}_1$. If the colours are replaced by c_2 and c_3 but not c_1 then the last forbidden pattern listed on the figure invalidates this choice. This shows that r is a recolouring from the first problem to the second problem.

Remark 102. Note that the composition of two recolourings is a recolouring and that we have an identity recolouring. So we have a category of representations and recolourings. The trivial representation with no forbidden patterns of the form $(\emptyset, \mathcal{T})$ corresponds naturally to the structure $\mathcal{T}$; and, recolourings are simply homomorphisms for such trivial representations. So the category of structure and homomorphisms embeds in that of representations and recolourings.

If every recolouring from $(\mathcal{T}, \mathscr{F}')$ to itself (i.e. some particular endomorphism of $\mathcal{T}$) is an automorphism of $\mathcal{T}$ then we say that the representation $(\mathcal{T}, \mathscr{F}')$ is a *core*. We are now ready to define our normal form.

Definition 103 (Normal Form [74]). *A representation $(\mathcal{T}, \mathscr{F}')$ of a forbidden patterns problem Ω is said to be* in the normal form *if, and only if it satisfies the following six conditions.*

($\mathfrak{p}_1$) *An instance is valid if, and only if, it is weakly valid.*

($\mathfrak{p}_2$) *Every pattern of $\mathscr{F}'$ is a core (as a coloured structure).*

($\mathfrak{p}_3$) *It is not the case that $(\mathcal{F}_1, f_1)$ is a substructure of $(\mathcal{F}_2, f_2)$, for any distinct patterns $(\mathcal{F}_1, f_1)$ and $(\mathcal{F}_2, f_2)$ in $\mathscr{F}'$.*

($\mathfrak{p}_4$) *No pattern of $\mathscr{F}'$ is conform.*

($\mathfrak{p}_5$) *Every forbidden pattern is biconnected.*

($\mathfrak{p}_6$) *The representation $(\mathcal{T}, \mathscr{F}')$ is a core.*

Example 104. Let Ω_4 be the problem given on the top of Figure 7.3. We shall discuss briefly how its normal form is computed without explaining why the obtained problem is equivalent, for further details please refer to [74].

First we enforce $\mathfrak{p}_1$ to $\mathfrak{p}_3$ simply by taking the homomorphic image of the forbid-
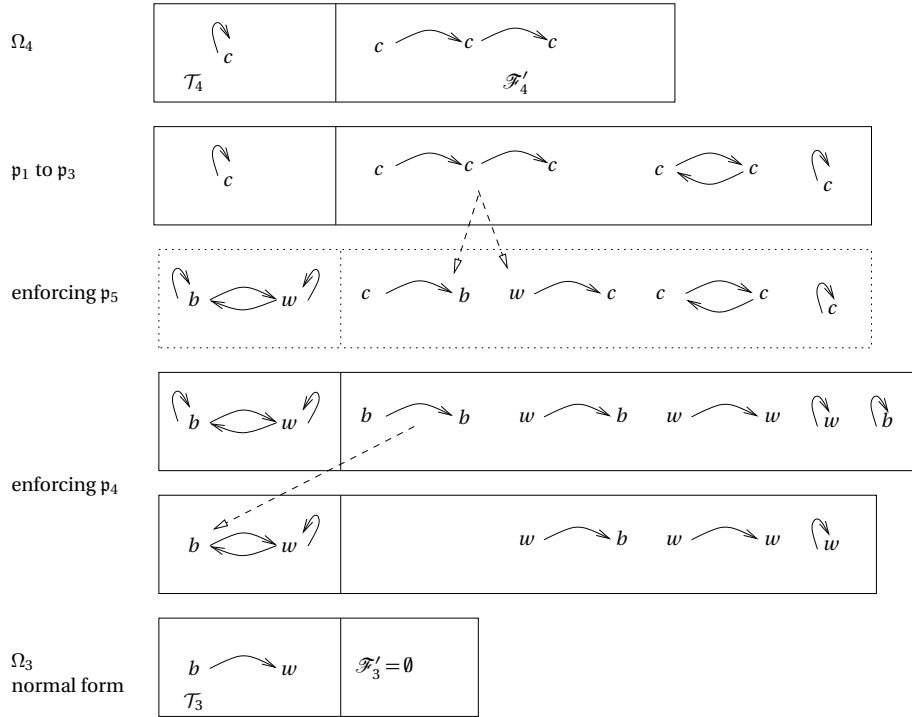
Figure 7.3: Computing the normal form.

den pattern, keeping only the minimal ones with respect to injective homomorphisms. Note that $\mathfrak{p}_4$ holds also in the representation of the problem we obtain this way which is given in the second row on the figure.

Next, we enforce $\mathfrak{p}_5$ by splitting the path of length two along its articulation point and copying its colour c into two new colours b and w, one for the substructure to the left of this articulation point, one for the substructure to the right of this articulation point. Replacing elsewhere the colour c by w and b in all possible ways and simplifying again by keeping the minimal patterns to enforce $\mathfrak{p}_3$, we obtain

the representation which is given in the fourth row of the figure. Note that it no longer satisfies $\mathfrak{p}_4$.

We enforce progressively $\mathfrak{p}_4$ by removing the conform forbidden patterns and removing the corresponding tuple in the structure describing the colours. We also remove any forbidden pattern that is no longer a coloured structure. We finally obtain this way the problem Ω_3 given in the last row on the figure.

One can compute effectively the normal form of a forbidden patterns problems and decide whether it is a finite CSP or not. If not, we now know that it is an infinite CSP *à la* Bodirsky.

Theorem 105 ([64, 74]). *Let Ω be a forbidden patterns problems given by some representation. Its normal representation $(\mathscr{F}', \mathcal{T})$ can be effectively computed. The problem Ω is a CSP with finite template $\mathcal{T}$ if $\mathscr{F}' = \emptyset$; and, a CSP with a countably infinite ω-categorical template, otherwise.*

proof (sketch). Regarding the computation of the normal form, the basic ideas are given in the above example. In general, there are issues regarding the order in which we impose some transformations (we might lose a previously enforced property) and termination (when splitting a non biconnected forbidden pattern, we might increase the sum of the sizes of the forbidden patterns, and need to perform this operation "in batch" to ensure termination). For more details regarding the computation of the normal form, and the fact that it represents indeed the same problem, please refer to [74].

When $\mathscr{F}' = \emptyset$, it follows directly from the definition and the fact that the normal representation represents the same problem that Ω is a CSP with finite template $\mathcal{T}$.

When $\mathscr{F}' \neq \emptyset$, our observation that we have an ω-categorical template follows from Theorem 98. It remains to prove that the problem is not a CSP with a finite template. The construction is similar in principle to our argument that Ψ_1 is not in CSP given at the beginning of § 7.3, though because we have colours the argument becomes a bit more involved.

We claim that the structure $\mathcal{T}$ is a no-instance of the problem. Otherwise, there would be an endomorphism h of $\mathcal{T}$ such that $(\mathcal{T}, h)$ is valid. It can be proved that h would be a recolouring to some "induced sub-representation" of $(\mathscr{F}', \mathcal{T})$, which would contradict the fact that $(\mathscr{F}', \mathcal{T})$ is a core by property $\mathfrak{p}_6$ (the argument uses properties $\mathfrak{p}_2, \mathfrak{p}_3$ and $\mathfrak{p}_4$).

Picking an endomorphism h of $\mathcal{T}$, say the identity, we obtain a coloured structure $(\mathcal{T}, h)$ that is not weakly valid (using $\mathfrak{p}_1$) and since h is a homomorphism, the non validity comes from occurrence(s) of biconnected pattern(s) (using $\mathfrak{p}_5$). We can "open up" the structure $(\mathcal{T}, h)$ along a cycle in such an occurrence. Hence, by taking suitable inverse homomorphic images of the coloured structure $(\mathcal{T}, h)$, we obtain eventually a structure that is valid w.r.t. the normal representation $(\mathscr{F}', \mathcal{T})$, which we call the *gadget*.

Given an integer n and plugging together multiple copies of this gadget in a suitable way, we obtain a coloured structure that is also valid (here we rely again on $\mathfrak{p}_5$ and use the randomised construction used in the proof of Theorem 97 to make sure that no obstruction can arise between multiple copies of the gadget). The underlying structure $\mathcal{S}$ of this coloured structure is therefore a yes-instance of the problem Ω. By construction, any homomorphic image of $\mathcal{S}$ with less than n elements is necessarily a no-instance of the problem Ω, because it must contain a copy of $\mathcal{T}$ or one of its homomorphic image (recall that $\mathcal{T}$ is a no-instance). This proves that Ω can not be a CSP with a template of size less than n. □

Remark 106. As a corollary, we can decide effectively whether a sentence of MMSNP captures a finite union of finite CSPs or not. We first turn the sentence into a finite union of forbidden patterns problems (this is effective). Next, we compute their normal forms. Finally, we check for recolourings between these normal representations as recolouring captures containment for problems in the normal form (see Theorem 114 on page 96). In fact, a weaker form of this result where we only check containment to a CSP suffices (see Proposition 118 on page 97), as we only care whether one forbidden patterns problem that is not a CSP survives this simplification process. The above proof can be extended by amending the gadget (see [74] for details).

7.5 Preservation

Preservation theorems relate syntax and semantic and can be stated in the following form. For a sentence φ in some logic (typically first-order logic), the following assertions are equivalent.

(i) The sentence φ is logically equivalent to a sentence ψ of a restricted syntactic form (*e.g.* existential, existential positive, ...).

(ii) The class of models of φ satisfy certain properties (*e.g.* closure under extension, closure under homomorphic images,...).

Many preservation theorems fail in the finite, *i.e.* when we replace models by finite models above (the implication from (i) to (ii) holds but not the converse).[4]

Rossman has proved that the homomorphism preservation theorem holds in the finite, a theorem which can be restated as follows in the context of MMSNP. Recall first that a first-order sentence φ is *closed under inverse homomorphism in the finite* if and only if, for any finite structures $\mathcal{A}$ and $\mathcal{B}$, if there exists a homomorphism from $\mathcal{A}$ to $\mathcal{B}$ and $\mathcal{B} \models \varphi$ then $\mathcal{A} \models \varphi$.

Theorem 107 (Finite HPT theorem[93]). *Let φ be a first-order sentence. The following are equivalent.*

(i) *φ is logically equivalent in the finite to a sentence ψ that is both MMSNP and first-order (ψ has no existential monadic predicates).*

---

[4]For further details on the failure of preservation theorems in the finite, see *e.g.* [92].

(ii) φ *is closed under inverse homomorphism in the finite.*

We say that φ is closed under *disjoint union in the finite* if and only if, for any finite structures $\mathcal{A}$ and $\mathcal{B}$, if $\mathcal{A} \models \varphi$ and $\mathcal{B} \models \varphi$ then $\mathcal{A} + \mathcal{B} \models \varphi$, where $\mathcal{A} + \mathcal{B}$ denotes the disjoint union of $\mathcal{A}$ and $\mathcal{B}$. If we insist further that φ is closed under *disjoint union,* we get a *one colour forbidden patterns problem* (or without colours in the sense that we no longer need coloured obstructions and that obstructions suffices).

Corollary 108. *Let φ be a first-order sentence. The following are equivalent.*

(i) φ *defines a one colour forbidden patterns problem, i.e. there exists a finite set of connected obstructions $\mathscr{F}$ such that* [5] $\varphi = \bigcap_{\mathcal{F} \in \mathscr{F}} \{\mathcal{F} \nrightarrow .\}$

(ii) φ *is closed under homomorphism and disjoint union.*

Proof. The implication (i) to (ii) holds trivially. We prove the converse. Applying Rossman's theorem we obtain a first-order MMSNP sentence ψ. Each negated conjunct of ψ correspond to a (not-necessarily connected) structure $\mathcal{F}$ and there is a finite set of structures $\mathscr{F}$ such that

$$\varphi = \bigcap_{\mathcal{F} \in \mathscr{F}} \{\mathcal{F} \nrightarrow .\}$$

Note that for any set of finite structures $\mathscr{F}$,

$$\bigcap_{\mathcal{F} \in \mathscr{F} \cup \{\mathcal{F}_1 + \mathcal{F}_2\}} \{\mathcal{F} \nrightarrow .\} = \Big(\bigcap_{\mathcal{F} \in \mathscr{F} \cup \{\mathcal{F}_1\}} \{\mathcal{F} \nrightarrow .\} \Big) \cup \Big(\bigcap_{\mathcal{F} \in \mathscr{F} \cup \{\mathcal{F}_2\}} \{\mathcal{F} \nrightarrow .\} \Big)$$

So φ is the disjoint union of one colour forbidden patterns problems. If there is a single forbidden patterns problem then we are done. Otherwise, we may keep only forbidden patterns problems that are the largest w.r.t. problem containment. If there is a single forbidden patterns problem that is larger than all the others then we are done.

Otherwise, we assume that φ consists of the disjoint union of several incomparable one colour forbidden patterns problems. We shall see that this case is not possible by deriving a contradiction.

Assume first for simplicity that there are exactly two forbidden patterns problems. That is there are two sets of obstructions $\mathscr{F}_1$ and $\mathscr{F}_2$ with $\Omega_1 := \bigcap_{\mathcal{F}_1 \in \mathscr{F}_1} \{\mathcal{F}_1 \nrightarrow .\}$ and $\Omega_2 := \bigcap_{\mathcal{F}_2 \in \mathscr{F}_2} \{\mathcal{F}_2 \nrightarrow .\}$ such that $\Omega_1 \cup \Omega_2 = \varphi$ and neither $\Omega_1 \subseteq \Omega_2$ nor $\Omega_1 \supseteq \Omega_2$. By Proposition 116, there exists an obstruction $\mathcal{F}_{1,2}$ of $\mathscr{F}_1$ such no obstruction $\mathcal{F}_2$ of $\mathscr{F}_2$ is homomorphic to $\mathcal{F}_{1,2}$, that is $\mathcal{F}_{1,2}$ belongs to Ω_2. Similarly, there exists an obstruction $\mathcal{F}_{2,1}$ of $\mathscr{F}_2$ such that $\mathcal{F}_{2,1} \in \Omega_1$. Since φ is closed under disjoint union it follows that $\mathcal{A} := \mathcal{F}_{1,2} + \mathcal{F}_{2,1} \models \varphi$. However, $\mathcal{A} \notin \Omega_1$ and $\mathcal{A} \notin \Omega_2$ so $\mathcal{A} \notin \Omega_1 \cup \Omega_2$ and $\mathcal{A} \not\models \varphi$, a contradiction.

---

[5] Here φ denotes the set of the yes-instances of the corresponding decision problem, that is $\{\mathcal{A}$ finite s.t. $\mathcal{A} \models \varphi\}$ the set of finite models of φ.

In general, there are allegedly n pairwise incomparable forbidden patterns problems $\Omega_1, \Omega_2, \ldots, \Omega_n$ given by the obstructions sets $\mathscr{F}_1, \mathscr{F}_2, \ldots, \mathscr{F}_n$ such that $\varphi = \Omega_1 \cup \Omega_2 \cup \ldots \cup \Omega_n$. We proceed as above and have a structure $\mathcal{F}_{1,2} + \mathcal{F}_{2,1}$ which models φ but it not in $\Omega_1 + \Omega_2$. By Proposition 116, for $3 \leq i \leq n$ there are structures $\mathcal{F}_{i,1}$ such that $\mathcal{F}_{i,1}$ is an obstruction from $\mathscr{F}_i$ and $\mathcal{F}_{i,1}$ belongs to Ω_1 (and so $\mathcal{F}_{i,1} \models \varphi$). Let $\mathcal{A} := \mathcal{F}_{1,2} + \mathcal{F}_{2,1} + \mathcal{F}_{3,1} + \ldots + \mathcal{F}_{n,1}$. Using the closure under disjoint union, we have $\mathcal{A} \models \varphi$. However, by construction $\mathcal{A} \notin \Omega_1 \cup \Omega_2 \cup \ldots \cup \Omega_n$ and $\mathcal{A} \not\models \varphi$, a contradiction. □

Preservation on a restricted class

A similar result to Theorem 107 holds when one works not on all finite structures but on a restricted class of finite structures $\mathscr{C}$ which is closed under substructures and disjoint union. It holds when $\mathscr{C}$ has bounded tree-width [1] and more generally when $\mathscr{C}$ is quasi-wide [37](to be defined shortly). These results on classes $\mathscr{C}$ are not implied by Rossman's theorem and differs in their proofs. While Rossman's proof is based on saturation, the proofs of the former are based on the density of minimal models and works for classes of sparse structures.

Recall that the Gaifman graph $\mathcal{G}_\mathcal{A}$ of a structure $\mathcal{A}$ has vertex set the domain of $\mathcal{A}$ and an edge between any two vertices if, and only if, they occur in the same tuple of some relation of $\mathcal{A}$. For a structure $\mathcal{A}$, an element a of A and an integer $r \geq 0$, we write $N_r^\mathcal{A}(a)$ to denote the set of elements of A that are at distance at most r from a in $\mathcal{G}_\mathcal{A}$. A set of elements B in a structure $\mathcal{A}$ is r-*scattered* if for every pair of distinct $a, b \in B$ we have $N_r^\mathcal{A}(a) \cap N_r^\mathcal{A}(b) = \emptyset$. Let $f : \mathbb{N} \to \mathbb{N}$ be a function. A class of finite structures $\mathscr{C}$ is *quasi-wide* with margin f if for every r and m there exists an N such that every structure $\mathcal{A}$ with at least N elements in $\mathscr{C}$ contains a set B with at most $f(r)$ elements such that the subgraph of $\mathcal{G}_\mathcal{A}$ (the Gaifman graph of $\mathcal{A}$) induced by $A \setminus B$ contains an r-scattered set of size m. We say that $\mathscr{C}$ is quasi-wide if there is some f such that $\mathscr{C}$ is quasi-wide with margin f.

Theorem 109 ([37]). *Let $\mathscr{C}$ be a class of structures that is quasi-wide and closed under substructures and disjoint union. Let φ be a first-order sentence. The following are equivalent.*

   (i) *There exists a first-order MMSNP sentence ψ such that $\varphi \cap \mathscr{C} = \psi \cap \mathscr{C}$.*

   (ii) *φ is closed under inverse homomorphism on $\mathscr{C}$.*

Corollary 110. *Let $\mathscr{C}$ be a class of structures that is quasi-wide and closed under substructures and disjoint union. Let φ be a first-order sentence. The following are equivalent.*

   (i) *φ captures a forbidden patterns problem with one colour over $\mathscr{C}$, i.e. there exists a finite set of connected obstructions $\mathscr{F}$ such that*

$$\mathscr{C} \cap \varphi = \mathscr{C} \cap \bigcap_{\mathcal{F} \in \mathscr{F}} \{\mathcal{F} \nrightarrow .\}$$

   (ii) *φ is closed under homomorphism and disjoint union on $\mathscr{C}$.*

Proof. The proof is similar to that of Corollary 108. We prove (ii) implies (i) by using Theorem 109 and obtain a first-order MMSNP formula. Next, we write it as the union of one colour forbidden patterns problems. That is, we have one colour forbidden patterns problems $\Omega_1, \Omega_2, \ldots, \Omega_n$ such that $\mathscr{C} \cap \varphi = \mathscr{C} \cap (\Omega_1 \cup \Omega_2 \cup \ldots \cup \Omega_n)$. We may also assume that for every $1 \geq i, j \leq n$ with $i \neq j$, $\Omega_i \cap \mathscr{C}$ is incomparable with $\Omega_j \cap \mathscr{C}$ and there exists a structure $\mathcal{A}_{i,j}$ in $\mathscr{C}$ that is a yes-instance of Ω_j and a no-instance of Ω_i (we may no longer take an obstruction as it may not belongs to $\mathscr{C}$). We build $\mathcal{A}$ as in the proof of Corollary 108 but using $\mathcal{A}_{i,j}$ and the result follows. □

It turns out that for a class $\mathscr{C}$ of graphs, being quasi wide and closed under substructure is equivalent to being *nowhere dense*. A class is not nowhere dense (and is said to be *somewhere dense*) if for some integer $r > 0$, every finite graph can be obtained by contracting edges that are at distance at most r and taking substructure of some graph in $\mathscr{C}$) [86]. Classes of bounded expansions, a notion we will introduced in § 9.3, are nowhere dense. On classes of bounded expansions, we shall see that every problem in MMSNP becomes a CSP.

Preservation for logics that extend FO
Homomorphism preservation in the finite for logics that extend FO has been investigated in the context of Constraint Satisfaction Problems. First, Feder and Vardi [45] provide *effective* homomorphism preservation theorems for monadic SNP, binary SNP and for SNP[6]. They also provide homomorphism preservation theorems for semi-positive Datalog[7] and variable-confined existential infinitary logic. We will give further details on their results concerning SNP in this section. Secondly, Dawar and Kreutzer [38] proved that the homomorphism preservation theorem fails for LFP, both in general and in restriction to finite structures. That is, there is a formula of LFP that is preserved under homomorphisms (in the finite) but is not equivalent (in the finite) to a Datalog program.

We call Binary SNP the fragment of SNP where existentially quantified second order predicates are at most binary.

Theorem 111 ([45]). *Let Φ be a sentence of Monadic SNP with inequalities (respectively, Binary SNP with inequalities). There exists a sentence Ψ of MMSNP (respectively, Monotone Binary SNP without inequalities) such that for any class $\mathscr{C}$ of finite structures that is closed under inverse homomorphism, Φ and Ψ express the same problem, i.e. $\Phi \cap \mathscr{C} = \Psi \cap \mathscr{C}$. Moreover, Ψ can be computed effectively from Φ.*

Remark 112. It is important to note that the formula Ψ does not depend on the class $\mathscr{C}$. When Φ is a sentence of Monadic SNP, the effective transformation has 5

---

[6]The maximal arity r of the existential predicate is not preserved when $r > 2$ as symbols that are signature dependent are added as new existential predicates and their arity could exceed r.

[7]This is denoted by Datalog($\neg, \neq$) in [45] which is the extension of Datalog where inequalities may appear in the body of a rule and the so-called EDB (the input predicates) may appear negatively in the body of a rule.

steps.

In the first step, the negated conjuncts of Φ are expanded until all variables are explicitly assumed to be distinct: if for two variables x and y there is no inequality $x \neq y$ in a negated conjunct then the negated conjunct is replaced by two negated conjuncts, one where $x \neq y$ is added and one where $x = y$ (which we propagate by replacing any occurrence of y by x).

In the second step, the negated conjuncts are split until no input relation occurs with all its argument equal unless it appears as the only input literal (this is achieved by adding a new monadic predicate with the same technique that is used to make negated conjuncts biconnected when computing the normal form of a MMSNP sentence).

In the third step, all negated conjuncts that are trivially satisfied are removed (*e.g.* containing $x \neq x$ or $E(x,y) \wedge \neg E(x,y)$ for an input predicate E).

In the fourth and fifth step, inequalities and negated input predicates are simply erased from negated conjuncts to obtain the formula Ψ which is in MMSNP.

When the sentence Φ is in Binary SNP the construction is mostly the same except for the last step. Inequalities are simulated by an equivalence relation as in Theorem 95. In both cases, a probabilistic argument is used to show equivalence of the sentences in the last step.

When the maximum arity of an existential symbol is 3 or more, Feder and Vardi can not guarantee that this arity is preserved unlike in the monadic and binary case. The proof breaks down in a Ramsey argument which works for graphs and may not work for higher arities. Instead, Feder and Vardi introduce one new existential predicate for each symbol of the input signature, which could exceed the arity of some existential symbol of the original sentence.

Theorem 113 ([45])**.** *Let Φ be a sentence of SNP with inequalities. There exists a sentence Ψ of SNP without inequalities such that for any class $\mathscr{C}$ of finite structures that is closed under inverse homomorphism, Φ and Ψ express the same problem, i.e. $\Phi \cap \mathscr{C} = \Psi \cap \mathscr{C}$. Moreover, Ψ can be computed effectively from Φ.*

8. Deciding Containment

We say that the decision problem Ω_1 is *contained* in the decision problem Ω_2 if, and only if, for any instance $\mathcal{A}$, if $\mathcal{A}$ is a yes-instance of Ω_1 then $\mathcal{A}$ is a yes-instance of Ω_2. We view a decision problem as the set of its yes-instances and simply write $\Omega_1 \subseteq \Omega_2$.

Feder and Vardi sketch that containment is decidable for MMSNP in [44]. It turns out that containment for forbidden patterns patterns that are given by a normal representations is captured precisely by the existence of a recolouring, just like homomorphism capture containment for CSP.

Theorem 114 ([65])**.** *Let Ω_1 and Ω_2 be two forbidden patterns problems given in the normal form over the relational signature σ. Ω_1 is contained in Ω_2 if, and only if, there is a recolouring from Ω_1 to Ω_2.*

We already knew that the existence of a recolouring implies containment (this is the reason they were introduced).

Proposition 115 ([64, 74])**.** *If there is a recolouring r from the representation of a forbidden patterns problem Ω_1 to the representation of a forbidden patterns problem Ω_2 then Ω_1 is contained in Ω_2.*

It remains to show the converse for representations given in the normal form.

8.1 Warm-up

We consider simple cases first.

When there are no colours

Proposition 116 ([64], see also [46])**.** *Let Ω_1 and Ω_2 be two forbidden patterns problems given by two sets of connected obstructions $\mathscr{F}_1$ and $\mathscr{F}_2$. The following are equivalent.*

(i) *Ω_1 is contained in Ω_2*

(ii) *for every obstruction $\mathcal{F}_2$ in $\mathscr{F}_2$ there exists an obstruction $\mathcal{F}_1$ in $\mathscr{F}_1$ such that there is a homomorphism from $\mathcal{F}_1$ to $\mathcal{F}_2$.*

Remark 117. The reader may check that (ii) is a special case of a recolouring (see Definition 100). One can view obstructions has having a single colour (set $\mathcal{T}$ to be the complete one-element structure; and turn each obstruction into a $\mathcal{T}$-coloured structure). The recolouring is trivial and the definition of a recolouring gives precisely (ii).

Proof. For the direct implication, note that an obstruction is necessarily a no-instance of a problem. Thus, any obstruction $\mathcal{F}_2$ in $\mathscr{F}_2$ is a no-instance of Ω_2 and consequently of Ω_1. Hence, there is indeed an obstruction $\mathcal{F}_1$ in $\mathscr{F}_1$ such that there is a homomorphism from $\mathcal{F}_1$ to $\mathcal{F}_2$.

For the converse implication, let $\mathcal{A}$ be a yes-instance of Ω_1 and assume for contradiction that $\mathcal{A}$ is a no-instance of Ω_2. That is there exists $\mathcal{F}_2$ in $\mathscr{F}_2$ and a homomorphism from $\mathcal{F}_2$ to $\mathcal{A}$. However, by assumption there exists an obstruction $\mathcal{F}_1$ in $\mathscr{F}_1$ and a homomorphism from $\mathcal{F}_1$ to $\mathcal{F}_2$. Combining the two homomorphisms we obtain a homomorphism from the obstruction $\mathcal{F}_1$ to $\mathcal{A}$, contradicting the fact that $\mathcal{A}$ is a yes-instance of Ω_1. $\qquad\square$

Recolouring implies Containment

proof (of Proposition 115 on the facing page). Let $\mathcal{A}$ be a yes-instance of Ω_1 and let h_1 be a homomorphism from $\mathcal{A}$ to $\mathcal{T}_1$ such that $(\mathcal{A}, h_1)$ is valid w.r.t. Ω_1. Let h_2 be the homomorphism from $\mathcal{A}$ to $\mathcal{T}_2$ such that $h_2 := r \circ h_1$. We claim that $(\mathcal{A}, h_2)$ is valid w.r.t. Ω_2 which implies that $\mathcal{A}$ is a yes-instance of Ω_2.

Assume for contradiction that this claim is false and that there exists a forbidden pattern $(\mathcal{F}_2, f_2)$ in $\mathscr{F}_2'$ and a homomorphism $\tilde{f}_2$ from $(\mathcal{F}_2, f_2)$ to $(\mathcal{A}, h_2)$ (to assist the reader, we provide a picture below of a commutative diagram involving all structures and homomorphisms mentioned in this proof). Let $f_1 := h_1 \circ \tilde{f}_2$. Note that $r \circ f_1 = f_2$. By assumption r is a recolouring and there exists a forbidden pattern $(\mathcal{G}_1, g_1)$ in $\mathscr{F}_1'$ and a homomorphism $\tilde{g}_1$ from $(\mathcal{G}_1, g_1)$ to $(\mathcal{F}_2, f_1)$. Hence we obtain a contradiction as this would mean that $\tilde{f}_2 \circ \tilde{g}_1$ is a homomorphism from a forbidden pattern $(\mathcal{G}_1, g_1)$ in $\mathscr{F}_1'$ to $(\mathcal{A}, h_1)$ and that $(\mathcal{A}, h_1)$ is not valid w.r.t. Ω_1.

$\square$
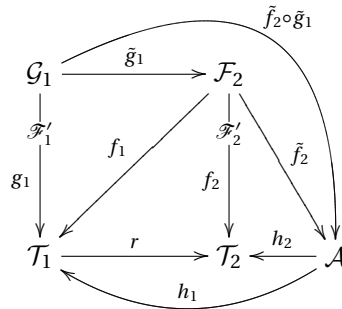
Containment in a Constraint Satisfaction Problem

Another case where it is not too hard to see that the converse of Proposition 115 holds is when Ω_2 is in CSP.

Proposition 118. *Let Ω_1 and Ω_2 be two forbidden patterns problems. If Ω_1 is given in the normal form and Ω_2 is in CSP then Ω_1 is included in Ω_2 if, and only if, there is a recolouring r from Ω_1 to Ω_2.*

Though this case is subsumed by our main result, its proof will serve as a good warm-up. In particular, it will allow us to introduce a key ingredient which is a generalisation by Feder and Vardi of a result due to Erdös, which was used in the proof of Theorem 97. Recall first that the *girth* of a structure is the length of its shortest cycle (and so if there are no cycles then the structure has infinite girth).

Lemma 119 (Erdös lemma [44])**.** *Fix two positive integers r and s. For every structure $\mathcal{B}$, there exists a structure $\mathcal{D}$ such that: the girth of $\mathcal{D}$ is greater than r; there is a homomorphism from $\mathcal{D}$ to $\mathcal{B}$; and for every structure $\mathcal{C}$ of size at most s, there is a homomorphism from $\mathcal{B}$ to $\mathcal{C}$ if, and only if, there is a homomorphism from $\mathcal{D}$ to $\mathcal{C}$.*

Proof (of Proposition 118). When Ω_2 is in CSP this means that we may assume that $\mathscr{F}_2' = \emptyset$. This means that in this case a recolouring is nothing other than a homomorphism from $\mathcal{T}_1$ to $\mathcal{T}_2$. In particular if $\mathcal{T}_1$ is a yes-instance of Ω_1 then we are done. However, this is in fact not true in general.

Assume that Ω_1 is given in the normal form. This means that $\mathcal{T}_1$ is a no-instance of Ω_1 unless $\mathscr{F}_1' = \emptyset$ [74]. We use Erdös lemma: we choose r greater than the largest forbidden patterns in $\mathscr{F}_1'$; s to be $|\mathcal{T}_2|$, the size of $\mathcal{T}_2$; and $\mathcal{B} := \mathcal{T}_1$.

We claim that the structure $\mathcal{D}$ obtained from the lemma in this way is in fact a yes-instance of Ω_1. This is because the homomorphism, say d_1, given by the lemma from $\mathcal{D}$ to $\mathcal{B} = \mathcal{T}_1$ gives us a valid colouring w.r.t. Ω_1. To see this, we use the fact that Ω_1 is given in the normal form: it suffices to show that $(\mathcal{D}, d_1)$ is weakly valid; and, for every forbidden pattern $(\mathcal{F}_1, f_1)$, the structure $\mathcal{F}_1$ is biconnected and must contain a cycle, so it can not occur as a substructure of $\mathcal{D}$ which has a girth greater than the size of any forbidden patterns.

By containment of Ω_1 in Ω_2 it follows that $\mathcal{D}$ is a yes-instance of Ω_2 and that there is a homomorphism from $\mathcal{D}$ to $\mathcal{T}_2$. Hence, by construction of $\mathcal{D}$ this means that there is a homomorphism from $\mathcal{B} = \mathcal{T}_1$ to $\mathcal{T}_2$ and that we are done. $\qquad\square$

8.2 From Forbidden Patterns Problem to CSP and Back

The following result is an adaptation of the ideas used by Feder and Vardi in the computational equivalence of MMSNP with CSP (Theorem 97). There is a small difference here, as the signature of the CSP is now parameterised by a set of patterns that must include the patterns from the forbidden patterns problem considered but may include more. We denote by $\mathrm{CSP}(-, \mathcal{T})$ the (non-uniform) constraint satisfaction problem with template $\mathcal{T}$ and by $\mathrm{CSP}(\mathrm{girth} > \gamma, \mathcal{T})$ its restriction to input of girth greater than γ.

Theorem 120. *Let Ω be a forbidden patterns problem given in the normal form over the relational signature σ. Let $\mathscr{F}$ be a set of biconnected σ-structures that includes all structures involved in patterns forbidden by Ω. Let γ be a fixed integer greater than the largest structure in $\mathscr{F}$.*

There exists a relational signature τ, a τ-structure $\mathcal{T}_\Omega$, and two first-order interpretations Π and Π^{-1} such that:

- *τ extends σ with new symbols, one symbol $R_{\mathcal{F}}$ of arity $|\mathcal{F}|$ for each $\mathcal{F}$ in $\mathscr{F}$;*
- *Π is a quantifier-free first-order interpretation using conjunction only;*
- *Π^{-1} is a first-order interpretation;*
- *$\Pi^{-1} \circ \Pi$ is the identity over σ-structures;*
- *Ω reduces to $CSP(-, \mathcal{T}_{\Omega})$ via Π; and,*
- *$CSP(girth > \gamma, \mathcal{T}_{\Omega})$ reduces to Ω via Π^{-1}.*

We sketch the proof of this result in the remaining of this section, providing an example to help the reader understand the main ideas.

Example 121. We consider the forbidden patterns problem defined by the sentence Ψ_2 in the introduction. It is a variant of the well-known NP-complete problem No-Monochromatic-Triangle. It is given in its normal form on Figure 8.1. The
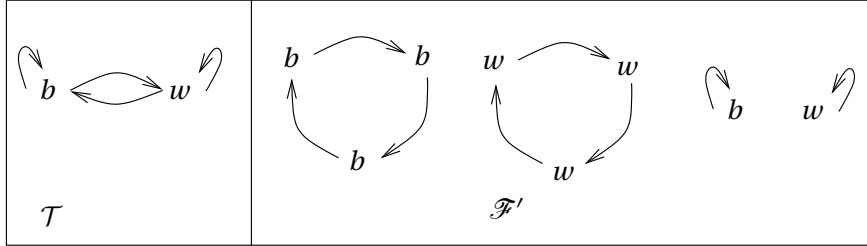

Figure 8.1: No-Monochromatic-Triangle.

signature of this problem is $\sigma = \langle E \rangle$ where E is binary which we extend to a *new signature $\tau = \langle E, R, S \rangle$* where R is ternary and S unary (R encodes the 3-cycles and S the self-loops). The *interpretation Π from σ to τ* is given by:

- $\varphi_R(y_1, y_2, y_3) := E(y_1, y_2) \wedge E(y_2, y_3) \wedge E(y_3, y_1)$;
- $\varphi_S(y_1) := E(y_1, y_1)$; and,
- $\varphi_E(y_1, y_2) := E(y_1, y_2)$.

The *interpretation Π^{-1} from τ to σ* is given by:

$$\psi_E := \big(E(y_1, y_2)\big) \vee \big(y_1 = y_2 \wedge S(y_1)\big) \vee \big(\exists x \ R(y_1, y_2, x) \vee R(x, y_1, y_2) \vee R(y_2, x, y_1)\big).$$

The structure $\mathcal{T}_{\Omega}$ has two elements b and w and, relations $E := \{b, w\}^2$, $S := \emptyset$ and $T := \{b, w\}^3 \setminus \{(b, b, b) \ (w, w, w)\}$.

Signature of the CSP. The problem Ω is represented by a σ-structure $\mathcal{T}$ and a list of forbidden $\mathcal{T}$-coloured structures $\{(\mathcal{F}_1, f_1), (\mathcal{F}_2, f_2), \ldots, (\mathcal{F}_n, f_n)\}$. Let $\mathscr{F}$ be the set of the σ-structures that consists of the structures $\mathcal{F}_i$ considered up to isomorphism. For every $\mathcal{F}$ in $\mathscr{F}$, we introduce a new symbol $R_{\mathcal{F}}$ of arity $|\mathcal{F}|$. Let τ be the signature that consists of the symbol of σ together with the new symbols $R_{\mathcal{F}}$.

Interpretation from the forbidden patterns problem to the CSP. Let $\varphi_{\mathcal{F}}$ be the quantifier-free part of the canonical conjunctive query of $\mathcal{F}$, that is:

$$\varphi_{R_{\mathcal{F}}} := \bigwedge_{R \in \sigma} \bigwedge_{R^{\mathcal{F}}(\bar{x}) \text{ holds}} R(\bar{x})$$

For a symbol R in σ (that is copied to τ), we set $\varphi_R := R(\bar{x})$. Let Π be the interpretation from σ to τ given by the formulae $\varphi_{\mathcal{F}}$ and the formulae φ_R. Note that Π is a quantifier-free interpretation of width one using only conjunction.

Interpretation from the CSP to the forbidden patterns problem. Let Π^{-1} be the interpretation from τ to σ given by reversing in a natural way the interpretation Π:

$$\psi_R := R(\bar{y}) \vee \bigvee_{\mathcal{F} \in \mathscr{F}} \bigvee_{R^{\mathcal{F}}(\bar{y}) \text{ holds}} \exists \tilde{x}\, R_{\mathcal{F}}(\tilde{x}, \bar{y}) \wedge \epsilon(\tilde{x}, \bar{y})$$

In the above sentence $\tilde{x}$ represent the elements of $\mathcal{F}$ not present among $\bar{y}$ and in $R_{\mathcal{F}}(\tilde{x}, \bar{y})$, the reader should understand that the variables $\tilde{x}, \bar{y}$ are reordered in a suitable fashion. The sentence ϵ is a conjunction of equalities between variables among $\tilde{x}, \bar{y}$.

By construction, $\Pi^{-1} \circ \Pi$ is the identity over σ-structures.

Construction of the template of the CSP. We build the τ-structure $\mathcal{T}_\Omega$ as an extension of the σ-structure $\mathcal{T}$ describing the colours of the forbidden patterns problem Ω. So on σ both structures agree and for every n-ary new symbol $R_{\mathcal{F}}$ and for every n-tuples of colours $c_1, c_2, \ldots, c_n$ we set $R_{\mathcal{F}}(c_1, c_2, \ldots, c_n)$ to hold unless,

- it is explicitly forbidden by a pattern $(\mathcal{F}, f)$ where $f(x_i) = c_i$; or,

- ⋆ the coloured structure $(\mathcal{F}, f)$ is implicitly forbidden by $(\mathcal{G}, g)$ in $\mathscr{F}'$ where $\mathcal{G}$ is a substructure of $\mathcal{F}$ and g agrees with $f(x_i) = c_i$ where defined[1].

Computational equivalence. By construction, the forbidden patterns problem Ω reduces to $CSP(-, \mathcal{T}_\Omega)$ via the interpretation Π. The converse interpretation Π^{-1} is not a reduction in general. It is a reduction for the τ-structures that will "not change too much" under $\Pi \circ \Pi^{-1}$. More formally, let $\mathcal{B}$ be the image of a τ-structure $\mathcal{A}$ under $\Pi \circ \Pi^{-1}$. The monotonic nature of the interpretations means that $\mathcal{A}$ is necessarily a substructure of $\mathcal{B}$ and the monotonic nature of the problems under consideration means that we only need to show that if $\mathcal{A}$ is a yes-instance then so is $\mathcal{B}$. The colouring certificate for $\mathcal{A}$ will validate $\mathcal{B}$ provided that if a new tuple involving $R_{\mathcal{G}}$ appeared in $\mathcal{B}$ it is a consequence of a larger tuple $R_{\mathcal{F}}$ where $\mathcal{F}$ and $\mathcal{G}$ are patterns in $\mathscr{F}$ and $\mathcal{G}$ is a substructure of $\mathcal{F}$. This holds because of the condition ⋆ in the construction of $\mathcal{T}_\Omega$.

---

[1] This second case ⋆ allows to channel constraints from one symbol in τ to another as all information regarding the relationship between the forbidden patterns is lost in the new signature τ.

In particular, we can guarantee that $\Pi \circ \Pi^{-1}$ will not change to much a τ-structure $\mathcal{A}$ if it is of sufficiently high girth, say a girth higher than γ, the number of elements of the largest pattern in $\mathscr{F}$ (this is because all patterns in $\mathscr{F}$ are biconnected). This proves that Π^{-1} is a reduction for instances of girth greater or equal to γ.

Note that we may extend $\mathscr{F}$ with any biconnected σ-structure without affecting the constructions or the result. This concludes the proof of Theorem 120.

8.3 Recolouring Captures Containment

We prove Theorem 114 in this section.

Let $\mathscr{F}$ be the set of biconnected structures involved as patterns in both Ω_1 and Ω_2. We use Theorem 120 for each problem, using $\mathscr{F}$ as a parameter, and obtain a τ-structure $\mathcal{T}_{\Omega_1}$ for Ω_1 and a τ-structure $\mathcal{T}_{\Omega_2}$ for Ω_2.

Lemma 122. *If Ω_1 is contained in Ω_2 then $CSP(girth > \gamma, \mathcal{T}_{\Omega_1})$ is contained in $CSP(girth > \gamma, \mathcal{T}_{\Omega_2})$.*

Proof. Let $\mathcal{A}$ be a τ-structure of girth greater than γ such that there is a homomorphism from $\mathcal{A}$ to $\mathcal{T}_{\Omega_1}$. Since Π^{-1} is a reduction to Ω_1, it follows that $\Pi^{-1}(\mathcal{A})$ is a yes-instance of Ω_1. By inclusion of Ω_1 in Ω_2 it follows that $\Pi^{-1}(\mathcal{A})$ is also a yes-instance of Ω_2. Since Π is a reduction from Ω_2 to $CSP(-, \mathcal{T}_{\Omega_2})$, the structure $\mathcal{B} := \Pi \circ \Pi^{-1}(\mathcal{A})$ is a yes-instance of $CSP(-, \mathcal{T}_{\Omega_2})$. Hence, there is a homomorphism from $\mathcal{B}$ to $\mathcal{T}_{\Omega_2}$. Since $\mathcal{A}$ is a substructure of $\mathcal{B}$ by monotonicity of the interpretations, it follows that there is a homomorphism from $\mathcal{A}$ to $\mathcal{T}_{\Omega_2}$. $\qquad \square$

Using Erdös lemma we will derive the following.

Lemma 123. *The following are equivalent.*

(i) *$CSP(girth > \gamma, \mathcal{T}_{\Omega_1})$ is contained in $CSP(girth > \gamma, \mathcal{T}_{\Omega_2})$.*

(ii) *$CSP(-, \mathcal{T}_{\Omega_1})$ is contained in $CSP(-, \mathcal{T}_{\Omega_2})$.*

(iii) *There is a homomorphism from $\mathcal{T}_{\Omega_1}$ to $\mathcal{T}_{\Omega_2}$.*

Proof. The equivalence between (ii) and (iii) is easy and well known. The implication from (ii) to (i) holds trivially. We prove that (i) implies (iii). Let $\mathcal{D}$ be the structure obtained from Erdös lemma from $\mathcal{B} := \mathcal{T}_{\Omega_1}$ with $s := |\mathcal{T}_{\Omega_2}|$ and $g := \gamma$. We know that there is a homomorphism from $\mathcal{D}$ of girth greater than γ to $\mathcal{T}_{\Omega_1}$ (recall that there is a homomorphism from $\mathcal{D}$ to $\mathcal{B}$). It follows from our assumption (i) that there is also a homomorphism from $\mathcal{D}$ to $\mathcal{T}_{\Omega_2}$. Appealing to Erdös lemma again we finally have that there is a homomorphism from $\mathcal{B} = \mathcal{T}_{\Omega_1}$ to $\mathcal{C} = \mathcal{T}_{\Omega_2}$. $\qquad \square$

Lemma 124. *If r is a homomorphism from $\mathcal{T}_{\Omega_1}$ to $\mathcal{T}_{\Omega_2}$ then r is a recolouring from Ω_1 to Ω_2.*

Proof. Recall that $\mathcal{T}_1$ (respectively $\mathcal{T}_2$) the structure used to colour the forbidden patterns of Ω_1 (respectively Ω_2) is by construction the σ-retract of $\mathcal{T}_{\Omega_1}$ (respectively, $\mathcal{T}_{\Omega_2}$). Hence, r is readily a homomorphism from $\mathcal{T}_{\Omega_1}$ to $\mathcal{T}_{\Omega_2}$.

It remains to show that for any $\mathcal{T}_2$-coloured pattern $(\mathcal{F}_2, f_2)$ forbidden by Ω_2, any of its inverse image under r– that is a $\mathcal{T}_1$-coloured structure $(\mathcal{F}_2, f_1)$ such that $f_2 = r \circ f_1$– is not valid w.r.t. Ω_1. Let $(\mathcal{F}_2, f_2)$ and $(\mathcal{F}_2, f_1)$ be as above. By construction of $\mathcal{T}_{\Omega_2}$, the tuple $R_{\mathcal{F}_2}(f_2(\bar{x}))$ does not hold in $\mathcal{T}_{\Omega_2}$. Since r is a homomorphism such that $f_2 = r \circ f_1$, the tuple $R_{\mathcal{F}_2}(f_1(\bar{x}))$ does not hold in $\mathcal{T}_{\Omega_1}$. By construction of $\mathcal{T}_{\Omega_1}$, this is because either a coloured pattern $(\mathcal{G}_1, g_1)$ forbidden by Ω_1 with pattern $\mathcal{F}_2$ or a substructure of $\mathcal{F}_2$ disallowed this tuple. In any case, we have that $(\mathcal{G}_1, g_1)$, which is forbidden by Ω_1 occurs in $(\mathcal{F}_2, f_1)$. This shows that $(\mathcal{F}_2, f_1)$ is not valid w.r.t. Ω_1. $\qquad\square$

We may now prove our main result by combining the three previous lemmas.

proof (of Theorem 114 on page 96). The definition of a recolouring implies containment as proved in Proposition 115. We now prove the converse. Suppose that Ω_1 is contained in Ω_2. By Lemma 122, it follows that CSP(girth $> \gamma, \mathcal{T}_{\Omega_1}$) is contained in CSP(girth $> \gamma, \mathcal{T}_{\Omega_2}$). By Lemma 124, it follows that there is a homomorphism r from $\mathcal{T}_{\Omega_1}$ to $\mathcal{T}_{\Omega_2}$. Finally, by Lemma 124 it follows that r is a recolouring from Ω_1 to Ω_2. $\qquad\square$

We can strengthen our main result by relaxing some hypothesis.

Corollary 125. *Let Ω_1 and Ω_2 be two forbidden patterns problems over the relational signature σ. If Ω_1 is given in a form that satisfies properties $\mathfrak{p}_1$ to $\mathfrak{p}_5$ then Ω_1 is contained in Ω_2 if, and only if, there is a recolouring from Ω_1 to Ω_2.*

Proof. We have not really used in the proof of our main result the property $\mathfrak{p}_6$ from the normal form. In any case, when computing the normal form this condition is enforced last and between a representation satisfying only the first five conditions and the normal representation, there are recolouring in both directions (the normal representation is the core of the former). Since two recolourings may be composed into a recolouring, this means that our main result holds also when the representations are given in a slightly weaker form for which properties $\mathfrak{p}_1$ to $\mathfrak{p}_5$ holds only, which corresponds essentially to the normal form of Feder and Vardi [44]. Moreover, when enforcing that the patterns are biconnected (property $\mathfrak{p}_5$), the colour of an articulation point of a pattern is duplicated, and as we pointed out in [74] the mapping which identifies this colour back to the original one and leaves the other unchanged is a recolouring from the new representation to the old representation of the problem. Similarly, when removing conform patterns (property $\mathfrak{p}_4$), there is also a recolouring from the new representation to the old one. Finally, enforcing the first three properties has no effect on the existence of

---

[1]Note that this is the best we can do as we may not do without property $\mathfrak{p}_5$ as example 104 shows.

recolouring. This means that there is a recolouring from the normal representation to the original one. Thus, we may relax the hypothesis of our main result to allow Ω_2 to be given arbitrarily. □

Example 126. We will see that the above is the best result possible regarding recolouring and containment. Recall Ω_3 and Ω_4 from Example 104. The two problems Ω_3 and Ω_4 coincide and Ω_3 is given in the normal form while Ω_4 is not (its only forbidden pattern fails to be biconnected). The mapping r which sends w and b to the single colour c of Ω_4 is a recolouring from Ω_3. However, there is no recolouring from Ω_4 to Ω_3 as there is no homomorphism from $\mathcal{T}_4$ to $\mathcal{T}_3$, since the former is a self-loop and the latter has no self-loop.

8.4 Complexity of containment

Feder and Vardi argued that MMSNP containment is decidable [44]. However a precise complexity was not given. Every sentence of MMSNP captures a finite union of forbidden patterns problems (see Corollary 99 on page 86) so this motivates us to reformulate the question in terms of forbidden patterns problems.

FPP-Containment:
- Input: forbidden patterns problems Ω_1 and Ω_2 given by $(\mathcal{T}_1, \mathscr{F}_1')$ and $(\mathcal{T}_2, \mathscr{F}_2')$.
- Question: is Ω_1 contained in Ω_2?

It is not difficult to see that the problem is at least NP-hard. Indeed, in the restricted case when the problems have no forbidden patterns, we have in fact the CSP-containment problem (also known as the uniform constraint satisfaction problem) which is NP-complete. In the restricted case when Ω_1 is given by a representation $(\mathcal{T}_1, \mathscr{F}_1')$ which satisfies properties $\mathfrak{p}_1$ to $\mathfrak{p}_5$, the question is equivalent to the following decision problem (by Corollary 125).

Recolouring:
- Input: forbidden patterns problems Ω_1 and Ω_2 given by $(\mathcal{T}_1, \mathscr{F}_1')$ and $(\mathcal{T}_2, \mathscr{F}_2')$.
- Question: is there a recolouring from Ω_1 to Ω_2?

The complexity of this problem is at most in Σ_3^p. This third level of the polynomial hierarchy is obtained directly from the definition of a recolouring. *Guess a homomorphism r, for every inverse image of every forbidden pattern, guess that it is non valid.* There are not many known complete problems in the third level of the polynomial hierarchy to choose from. There are however a myriad of problems in the second level. Using Generalised Graph Colouring [96] we can easily show that.

Proposition 127. *The restriction of Recolouring where Ω_2 has a single colour is Π_2^p-complete. Consequently, Recolouring is Π_2^p-hard.*

Proof. The problem Generalised Graph Colouring is Σ_2^p-complete. It takes as input two graphs $\mathcal{F}$ and $\mathcal{G}$ and asks whether there is a partition into two sets of the vertices of the graph $\mathcal{F}$ such that neither set contains $\mathcal{G}$ as a subgraph.

We reduce this problem to the complement of Recolouring as follows. Choose for $\mathcal{T}_1$ the complete graph with two colours white and black. $\mathscr{F}_1'$ contains a white $\mathcal{G}$ and a black $\mathcal{G}$ as forbidden patterns. Choose for $\mathcal{T}_2$ a self-loop and for $\mathscr{F}_2'$ the pattern induced by $\mathcal{F}$ coloured with the only colour of $\mathcal{T}_2$.

The issue of dealing with obstructions via homomorphisms rather than subgraphs can be easily dealt with by amending the signature with a second binary predicate E' which will play the role of $\neq$: make every structure a clique w.r.t. this predicate E' and forbid the self-loop in both problems. The new problems satisfy both property $\mathfrak{p}_1$. $\qquad\square$

It would be interesting to pinpoint more accurately the complexity of Recolouring, which ought to be complete for Σ_3^p. As computing the normal form may blow up significantly the size of the representation, a possible approach would be to find a suitable generalisation of recolourings which works on a representation that is not in the normal form, which would enable to better apprehend the complexity of FPP-Containment.

9. Lifting duality and preservation

The case of first-order MMSNP definable constraint satisfaction problems has attracted some attention in Combinatorics. It turns out that they are precisely those that can be defined in terms of forbidden trees. For example, a directed graph $\mathcal{G}$ omits (in the sense that there is no homomorphism from) the directed path with $n+1$ vertices $\mathcal{P}_{n+1}$ if, and only if, there exists a homomorphism from $\mathcal{G}$ to the transitive tournament on n vertices $\mathcal{T}_n$. The pair $(\mathcal{P}_{n+1}, \mathcal{T}_n)$ is called a *duality pair*. Given a graph from such a duality pair, we shall see that it is possible to compute the other graph. We will see that the notion of duality pairs can be generalised to allow for several obstructions and several templates.

It turns out that when instead of working on arbitrary classes of input, we impose a restriction to a class $\mathscr{C}$ of sufficiently sparse graphs, say of bounded degree, bounded tree-width, planar, or more generally of bounded expansion, then even for obstructions that are not trees, one can build a template showing that we have a restricted constraint satisfaction problem (in this case $\mathscr{C}$ is said to have all restricted dualities).

We will first survey these results. Next, we will see that it is possible to lift these results to MMSNP and show that unlike the general case where we separated MMSNP from CSP, when restricted to sufficiently sparse classes of structures, a problem in MMSNP becomes a restricted CSP. We will show that one can also lift Rossman's preservation theorem from first-order logic to SNP to obtain a result that is similar to Feder and Vardi's preservation theorem (Theorem 113). Our lifted result requires a stronger preservation property (which monotone SNP without inequalities satisfies) but it preserves the arity.

9.1 Homomorphism Duality

In combinatorics, the question of characterising when $(\mathcal{F}, \mathcal{T})$ is a *duality pair* has attracted some attention [87], that is when for any finite structure $\mathcal{A}$, there is no homomorphism from $\mathcal{F}$ to $\mathcal{A}$ if, and only if, there is a homomorphism from $\mathcal{A}$ to $\mathcal{T}$. Schematically,

$$\{\mathcal{F} \not\longrightarrow \cdot\} = \{\cdot \longrightarrow \mathcal{T}\}.$$

The *incidence graph of a structure* is the bipartite graph defined as follows: it has one vertex for each domain element and one vertex for each tuple occurring in some relation; and, a vertex representing a domain element is adjacent to a vertex representing a tuple if and only if the element occurs in the tuple. We define *structures that are trees* as follows. In the case of undirected graphs, we consider the usual definition of an acyclic connected graph. For directed graphs, a tree is

the orientation of a tree. More generally, a relational structure is a tree if it is connected, no tuple repeats an element, and its incidence graph is a tree.

Theorem 128 (duality pairs [87]). *For a core structure $\mathcal{F}$ there exists a structure $\mathcal{T}$ such that $(\mathcal{F}, \mathcal{T})$ is a duality pair if, and only if, $\mathcal{F}$ is a tree. For a tree $\mathcal{F}$, such a structure $\mathcal{T}$, the so-called* dual *is unique up to homomorphism equivalence.*

proof (sketch). The original proof relies on an elegant correspondence between duality pairs and gap pairs. A pair of structures $(\mathcal{T}, \mathcal{E})$ is a gap, when $\mathcal{T}$ is homomorphic to $\mathcal{E}$ but not the converse – which we write $\mathcal{T} < \mathcal{E}$ – and there is no structure $\mathcal{C}$ such that $\mathcal{T} < \mathcal{C} < \mathcal{E}$.

Gap pairs are characterised: the structure $\mathcal{T}$ must be a tree, in which case a construction is provided for $\mathcal{E}$ via the so-called *arrow construction* (the size of $\mathcal{E}$ is typically exponential in the size of $\mathcal{T}$). The corresponding duality pair is $(\mathcal{T}, \mathcal{E}^{\mathcal{T}})$, where $\mathcal{E}^{\mathcal{T}}$ is the so-called *exponential* [62] (the size of the exponential is $|\mathcal{E}|^{|\mathcal{T}|}$). $\square$

A more general notion of *generalised duality pair* $(\mathscr{F}, \mathscr{T})$ between sets of structures has also been considered [46], that is when for every finite structure $\mathcal{A}$, there is no homomorphism from any $\mathcal{F}$ in $\mathscr{F}$ to $\mathcal{A}$ if, and only if, there is a homomorphism from $\mathcal{A}$ to some $\mathcal{T}$ in $\mathscr{T}$. Schematically,

$$\bigcap_{\mathcal{F} \in \mathscr{F}} \{\mathcal{F} \nrightarrow .\} = \bigcup_{\mathcal{T} \in \mathscr{T}} \{. \longrightarrow \mathcal{T}\}.$$

Of course, when $\mathscr{T}$ is a singleton $\{\mathcal{T}\}$, then we can always take for $\mathscr{F}$ the set of the cores of structures that are not homomorphic to $\mathcal{T}$. This set $\mathscr{F}$ of obstructions is infinite in general and some tractable classes of CSP are characterised by having a set of obstructions with good properties, for example, having bounded tree-width. For further details on this kind of duality, see the survey [13]. In the following we assume $\mathscr{F}$ to be a finite set.

Corollary 129 (finitary homomorphism duality [87]). *Let $\mathscr{F} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_m\}$ be a finite nonempty set of structures. The pair $(\mathscr{F}, \{\mathcal{T}\})$ is a generalised duality if, and only if, $\mathcal{T}$ is homomorphically equivalent to $\prod_{i=1,m} \mathcal{T}_i$ and $(\mathcal{F}_i, \mathcal{T}_i)$ is a duality pair for $i = 1, 2, \ldots, m$.*

Theorem 130 (characterisation of generalised duality pairs [46]). *Let $\mathscr{F}$ be a finite set of cores. If $(\mathscr{F}, \mathscr{D})$ is a generalised duality, then all elements of $\mathscr{F}$ are forests and $\mathscr{D}$ is uniquely determined up to homomorphic equivalence by $\mathscr{F}$.*

Deciding given $\mathscr{F}$ whether there exists a singleton $\mathscr{T} := \{\mathcal{T}\}$ such that $(\mathscr{F}, \mathscr{T})$ is a generalised duality pair, amounts to decide whether the forbidden patterns problem with representation $\mathscr{F}$ is a constraint satisfaction problem. When $\mathscr{T}$ is not a singleton, it amounts to decide whether the forbidden patterns problem is a finite union of constraint satisfaction problems. So the above results can be seen as a special case of Theorem 105 and its generalisation to disjoint union of forbidden patterns problems from [74].

In fact the initial construction of the dual used in the proof of Theorem 128, though very elegant, produces very large duals (the construction is doubly exponential in the size of the pattern). If we use our normal form (see Definition 103), we will add a colour for each articulation point of the forbidden pattern, and obtain a smaller dual (this corresponds to the enforcement of step $\mathfrak{p}_5$, *c.f.* Example 104 to get the intuition). Another construction which gives also smaller dual has been given in [88], and it is proved that for some cases there is no smaller dual. The historical survey [82] lists several other possible constructions.

A more general notion of *shadow duality* was considered and characterised in [59]. This characterisation amounts to a non effective form of Theorem 105.

9.2 Detecting First-order Constraint Satisfaction Problems

So far we have seen how given a set of obstructions, one can decide whether it corresponds to a CSP or a finite union of CSP. The converse question, "given a template $\mathcal{T}$, is there a finite set of obstructions $\mathscr{F}$ such that $(\mathscr{F}, \{\mathcal{T}\})$ is a generalised duality pair?" is NP-complete [61]; and, becomes tractable when $\mathcal{T}$ is a core.

We say that a function f from $\mathcal{T}^n$ to $\mathcal{T}$ is a *one-tolerant polymorphism* if for any relation symbol R and any tuples $t_1, t_2, \ldots, t_n$ on which R holds, but possibly for one tuple, R holds on the tuple $f(t_1, t_2, \ldots, t_n)$.

Together with Rossman's theorem (see Theorem 107), putting together several results from [61], we get the following.

Theorem 131. *Let $\mathcal{T}$ be a finite structure. The following are equivalent.*

(i) *The problem CSP($\mathcal{T}$) is first-order definable.*

(ii) *There exists a finite set of structures $\mathscr{F}$ such that $(\mathscr{F}, \mathcal{T})$ is a generalised duality pair.*

(iii) *For some integer n, there exists a one-tolerant n-ary polymorphism of $\mathcal{T}$.*

(iv) *There exists a hyper-endomorphism from $\mathcal{T} \times \mathcal{T}$ to $\mathcal{T}$.*

Moreover, given $\mathcal{T}$, deciding whether this is the case is NP-complete if $\mathcal{T}$ is not a core and in P otherwise.

Remark 132. In [61], the authors phrase (iv) in terms of "*dismantability of the square of the template to its diagonal*". This means that there exists a sequence of homomorphisms, starting from the square $\mathcal{T} \times \mathcal{T}$ and ending with its diagonal (which is isomorphic to $\mathcal{T}$). Each homomorphism sends a vertex b to a vertex a such that a *dominates* b. Their definition of domination can be restated as follows. The unary hyper-operation which sends b to $\{a, b\}$ and fixes the other vertices is a hyper-endomorphism. This justifies our reformulation.

Note that a problem is expressible in first-order MMSNP if, and only if, its complement is expressible in $\{\exists, \wedge, \vee\}$-FO. We have seen that hyper-endomorphisms caracterise the complexity of $\{\exists, \wedge, \vee\}$-FO(see Theorem 61). So the reformulation of (iv) does not come as a complete surprise on a superficial level, but it would

be interesting to have more direct evidence that explains why the concept arises here.

9.3 Restricted dualities

Let $\mathscr{C}$ be a class of structures. We say that $\mathscr{C}$ has *all restricted dualities* if, and only if, for every finite set of connected structures $\mathscr{F}$ there exists a *finite* structure $\mathcal{T}$, such that

$$\mathscr{C} \cap \bigcap_{\mathcal{F} \in \mathscr{F}} \{\mathcal{F} \not\longrightarrow .\} = \mathscr{C} \cap \{. \longrightarrow \mathcal{T}\}.$$

Note that $\mathcal{T}$ is not required to belong to $\mathscr{C}$. The first example of a restricted duality theorem is due to Häggvist and Hell.

Theorem 133. *[51] Let b be an integer and $\mathscr{C}$ be a class of graphs. If every graph in $\mathscr{C}$ has bounded degree then $\mathscr{C}$ has all restricted dualities.*

proof (sketch). Let $\mathscr{F}$ be a set of connected obstructions of size at most p. We will build $\mathcal{T}$ such that $\mathscr{C} \cap \bigcap_{\mathcal{F} \in \mathscr{F}} \{\mathcal{F} \not\longrightarrow .\} = \mathscr{C} \cap \{. \longrightarrow \mathcal{T}\}$.

To ascertain that a vertex c of an instance $\mathcal{A}$ is not involved in a forbidden pattern and is a "valid vertex", we only need to check $N_p^{\mathcal{A}}(c)$, the neighbourhood of c of vertices that are at distance at most p from c. Since the graph has bounded degree, any such neighbourhood has at most ℓ vertices where $\ell = 1 + b. \sum_{i=1}^{p}(b-1)^i$ and b denotes the maximum degree of the graphs in $\mathscr{C}$. We may therefore colour each vertex with a label from $\{1, 2, \ldots, \ell, \ell+1\}$ such that every vertex in such a neighbourhood has a different label.

We consider the finitely many rooted graphs $(c, \mathcal{G})$ with domain set $\{1, 2, \ldots, \ell, \ell+1\}$ that are valid w.r.t. $\mathscr{F}$. The template graph $\mathcal{T}$ has one vertex for each such valid rooted graph. Two vertices $(c, \mathcal{G})$ and $(c', \mathcal{G}')$ are adjacent whenever $\mathcal{G}$ and $\mathcal{G}'$ contain both c and c' as adjacent vertices, and can be "glued" adequately: the p-neighbourhood centred at c (respectively, c') are the same in both graphs.

A valid instance $\mathcal{A}$ from $\mathscr{C}$ is homomorphic to $\mathcal{T}$. We simply label its vertices with $\{1, 2, \ldots, \ell, \ell+1\}$ such that every vertex in a p-neighbourhood has a different label and map each vertex c of the labelled graph to the vertex $(c, \mathcal{G})$ of $\mathcal{T}$, where $\mathcal{G}$ is the labelled p-neighbourhood of c in $\mathcal{A}$.

By construction, $\mathcal{T}$ is valid so a non valid structure can not map to $\mathcal{T}$ by construction and we are done. $\square$

More recently, Nešetřil and Ossona de Mendez gave a duality theorem for proper minor closed classes.

Theorem 134. *[84] Let M be a graph and $\mathscr{C}$ be a class of graphs. If no graph in $\mathscr{C}$ admits M as a minor then $\mathscr{C}$ has all restricted dualities.*

One of the key notions in the proof of the above is that elements of $\mathscr{C}$ have a *low tree-depth decomposition*: that is, for every integer $p > 0$, there exists an integer $N > 0$ such that for every $\mathcal{C}$ in $\mathscr{C}$, there exists an N-colouring [1] such that any subgraph $\mathcal{H}$ of $\mathcal{G}$ induced by at most p colour classes has tree-depth at most p. Tree-depth is a graph measure which can be defined in several equivalent ways. We will give here a definition for structures, not just for graphs. Let F be a rooted forest (disjoint union of rooted trees). We define the *closure of F* w.r.t. a relational signature σ to be the σ-structure with domain $|F|$ and all tuples $R_i(x_1, x_2, \ldots, x_{r_i})$ such that the elements mentioned in this tuple $\{x_i | 1 \leq i \leq r_i\}$ form a chain w.r.t. the ancestor relation on F. The *tree-depth* of S, denoted by $\mathrm{td}(S)$, is the minimum height of a rooted forest F such that S is a substructure of the closure of F. Alternatively, the tree-depth can be seen as the quantifier rank of the canonical $\{\exists, \wedge\}$-FO-sentence of $\mathcal{S}$ (when rewriting this sentence in a non prenex-form).

The same authors have introduced the notion of classes of graphs of bounded expansion, which encompasses both classes of graphs of bounded degree and proper minor closed classes. A class of graphs $\mathscr{C}$ has *bounded expansion* if there exists a function $f : \mathbb{N} \to \mathbb{R}$ such that for every graph G in $\mathscr{C}$ and every $r > 0$, $\nabla_r(G)$, the so-called *grad of G of rank r* is bounded by $f(r)$, where $\nabla_r(G) = \max \frac{|E(G|\mathcal{P})|}{|\mathcal{P}|}$. Here, $\mathcal{P}$ is a set of disjoint sets of vertices of G, each of which induce a connected subgraph of G of radius at most r; and, $E(G|\mathcal{P})$ denotes the edge set of the minor of G constructed by identifying the vertices inside each set into a single vertex and deleting other vertices and edges. These authors proved that classes of graphs of bounded expansions have also low tree-depth decomposition[2] and proved the following general result.

Theorem 135. *[85] Let $\mathscr{C}$ be a class of graphs. If $\mathscr{C}$ has bounded expansion then $\mathscr{C}$ has all restricted dualities.*

Restricted Dualities for Structures
We extend Theorem 135 to structures. Recall that $\mathcal{G}_\mathcal{A}$ denotes the *Gaifman graph* of a structure $\mathcal{A}$. We will say that a class of structures $\mathscr{C}$ has *bounded expansion* if, and only if, the class of graphs $\{\mathcal{G}_\mathcal{A} \text{ s.t. } \mathcal{A} \in \mathscr{C}\}$ has bounded expansion.

Theorem 136. *Let $\mathscr{C}$ be a class of structures. If $\mathscr{C}$ has bounded expansion then $\mathcal{K}$ has all restricted duality.*

We sketch the proof in the remaining of this section, which follows exactly that of Theorem 135, with adaptation in places to deal with structures. A crucial step of the proof is to note that.

Lemma 137. *For a fixed relational signature, there are only finitely many cores of bounded tree-depth.*

---

[1] In the sense of graph theory: adjacent vertices have different colours.

[2] In fact they later showed that the two coincide and studied even more general classes that show a certain sparsity, see the survey [83].

proof (sketch). By induction on the tree-depth. We will omit the proof here and only give the intuition. If a structure is large and has bounded tree-depth $p + 1$ witnessed by a rooted forest of height $p + 1$. Then, since the structure is large so is the forest and it must have one tree with a large degree node or a large number of trees. In the former case, when one look at the structures induced by the subtrees rooted at the large degree nodes, at least two of them must have the same core by induction. In the later case, we may assume that we have a forest with no node of large degree, and necessarily two structures corresponding to two trees must have the same core. $\square$

For the second step, we use the fact that classes of graphs of bounded expansions have low-tree-depth decompositions. Observing that a low tree-depth decomposition of the Gaifman graph of a structure induces a low tree-depth decomposition of the structure, we get.

Lemma 138. *Let $\mathscr{C}$ be a class of structures. If $\mathscr{C}$ has bounded expansion then $\mathscr{C}$ has low tree-depth decompositions.*

So given a set of forbidden patterns $\mathscr{F}$ of size less than p, we will use the fact that $\mathscr{C}$ has bounded expansion to derive the existence of a low tree-depth decomposition, that is a partition of any $\mathcal{A}$ in $\mathscr{C}$ into some $q \geq p$ parts such that any p parts induce a structure of tree-depth at most p.

Next, we consider the set of connected cores of tree-depth at most p that are valid w.r.t. $\mathscr{F}$. Since the patterns are connected (it follows that the corresponding forbidden patterns problem is closed under disjoint union and) their disjoint union $\mathcal{U}$ is also valid w.r.t. $\mathscr{F}$, and any structure of tree-depth at most p that is valid w.r.t. $\mathscr{F}$ is homomorphic to $\mathcal{U}$. By Lemma 137, the structure $\mathcal{U}$ is finite.

By the low-tree depth decomposition from Lemma 138 and the construction of $\mathcal{U}$, a structure $\mathcal{A}$ in $\mathscr{C}$ is valid w.r.t. $\mathscr{F}$ if, and only if, every structure induced by p parts of the low-tree depth decomposition is valid w.r.t. $\mathscr{F}$ (the direct implication follows by monotonicity, the other direction follows from the fact that an obstruction in $\mathscr{F}$ has at most p elements).

When $q = p$, we can set $\mathcal{T} = \mathcal{U}$ and we are done. When $q > p$, we will need a new notion, which will allow to translate the property of "having all partial homomorphisms from some parts" into "having a (global) homomorphism".

Let $\mathcal{A}$ be a structure and $p \geq 2$ be an integer. We define the pth *truncated product* of $\mathcal{A}$, to be the structure $\mathcal{B}$ defined as follows. Its domain is $\bigcup_{i=1}^{p} W^i$ where,

$$W^i := \{(a_1, a_2, \ldots, a_{i-1}, \star, a_{i+1}, \ldots, a_p) \text{ s.t. } \forall 1 \leq k \leq p, k \neq i \implies a_k \in A\}$$

The symbol $\star$ denotes a new element (*i.e.* $\star \notin A$) that plays the role of a "don't care" symbol w.r.t. the usual product definition. That is, for every relation symbol R of arity r, we set $R(w^{i_1}, w^{i_2}, \ldots, w^{i_r})$ to hold in $\mathcal{B}$, where for every $1 \leq k \leq r$, w^{i_k} belongs to W^{i_k} and is of the form $w^{i_1} = (a_1^{i_1}, \ldots, a_{i_1-1}^{i_1}, \star, a_{i_1+1}^{i_1}, a_p^{i_1})$ if, and only if, for every $1 \leq i \leq p$, with $i \notin \{i_1, i_2, \ldots, i_r\}$, $R(a_i^{i_1}, a_i^{i_2}, \ldots, a_i^{i_r})$ holds in $\mathcal{A}$. We denote the pth truncated product by $\mathcal{A}^{\Uparrow p}$.

Lemma 139. *Let $p \geq 2$ and let $\mathcal{F}$ be a structure with $F < p$. For any structure $\mathcal{A}$, if $\mathcal{F}$ is homomorphic to $\mathcal{A}^{\Uparrow p}$ then $\mathcal{F}$ is homomorphic to $\mathcal{A}$. Consequently, if a structure $\mathcal{A}$ is valid w.r.t. $\mathcal{F}$ then $\mathcal{A}^{\Uparrow p}$ is also valid w.r.t. $\mathcal{F}$.*

Proof. Let h be a homomorphism from $\mathcal{F}$ to $\mathcal{A}^{\Uparrow p}$. Since $F < p$, there exists $1 \leq i_0 \leq p$ such that $h(W^{i_0}) \cap W^{i_0} = \emptyset$. Note that the projection to the i_0th coordinate π_0 is a homomorphism from the substructure of $\mathcal{A}^{\Uparrow p}$ induced by the removal of W^{i_0}. Thus, $\pi_0 \circ h$ is a homomorphism from $\mathcal{F}$ to $\mathcal{A}$. $\square$

Lemma 140. *Let $\mathcal{A}$ and $\mathcal{T}$ be structures. Let $A_1, A_2, \ldots, A_p$ be a partition of the domain A of $\mathcal{A}$. If for every substructure $\widetilde{A}_i$ of $\mathcal{A}$ induced by the removal of A_i there exists a homomorphism $\widetilde{s}_i$ from $\widetilde{A}_i$ to $\mathcal{T}$, then there exists a homomorphism from $\mathcal{A}$ to $\mathcal{T}^{\Uparrow p}$.*

Proof. For a_i in A_i, we set $s(a_i) := \big(\widetilde{s}_1(a_i), \ldots, \widetilde{s}_{i-1}(a_i), \star, \widetilde{s}_{i+1}(a_i), \ldots, \widetilde{s}_p(a_i)\big)$. It follows from the definition of the truncated product that s is a homomorphism from $\mathcal{A}$ to $\mathcal{T}^{\Uparrow p}$. $\square$

We are now ready to conclude the proof of Theorem 136.

If $q = p + 1$, then we set $\mathcal{T} := \mathcal{U}^{\Uparrow p+1}$. If there is a homomorphism from $\mathcal{A}$ to $\mathcal{T}$ then $\mathcal{A}$ is valid by Lemma 139. Conversely, if $\mathcal{A}$ is valid, then all substructures induced by p parts have a homomorphism to $\mathcal{U}$. By Lemma 140, it follows that $\mathcal{A}$ has a homomorphism to $\mathcal{T}$.

If $q > p + 1$, we may use the same proof principle as Lemma 140 to show that "having all homomorphism from p parts" to a structure $\mathcal{B}$ implies "having all homomorphisms from $p + 1$ parts" to $\mathcal{B}^{\Uparrow p+1}$. The result follows by setting $\mathcal{T} := \Big(\ldots \big((\mathcal{U}^{\Uparrow p+1})^{\Uparrow p+2} \big) \ldots \Big)^{\Uparrow q}$.

9.4 What input restrictions of forbidden patterns problems makes them constraint satisfaction problems?

We would like to know which classes $\mathscr{C}$ of structures have the property that for any forbidden patterns problem Ω, there exists a finite template $\mathcal{T}$ such that:

$$\Omega \cap \mathscr{C} = \mathrm{CSP}(\mathcal{T}) \cap \mathscr{C}.$$

When this is the case we will say that $\mathscr{C}$ has *all coloured dualities*. We shall see that this is the case when the structures in $\mathscr{C}$ have bounded degree [36], when their Gaifman graphs fall within a proper minor closed class (this includes structures of bounded tree-width) [67] and more generally when their Gaifman graphs have bounded expansion [68]. Our original proofs followed the lines of the proofs of the theorems of the previous section, but taking colours into account. We shall see that it is not necessary and that one can provide a shorter proof in the spirit of Theorem 98 where we lifted a result about first-order MMSNP sentences to arbitrary MMSNP sentences.

Formalising colours with monadic predicates

It will be convenient to view coloured structures, not as structures together with a homomorphism to a fixed template, but as extensions of the structure with monadic predicates. Let us introduce some notation we will use in the rest of this section. Let σ be a fixed relational structure. Let μ^+ be a signature of monadic symbols that are not in σ, symbols which we will write with a superscript $+$ for convenience. Let σ' be $\sigma \cup \mu^+$. Let σ'' be $\sigma \cup \mu^+ \cup \mu^-$ where $\mu^- := \{M^-$ s.t. M^+ is a symbol of $\mu\}$. For a class $\mathscr{C}$ of σ-structures, we denote by $\mathscr{C}'$ the class of all σ'-structures that are extensions of a structure in $\mathscr{C}$.

Let $\mathscr{F}'$ be a set of σ'-structures describing the coloured obstructions of a forbidden patterns problem. We will assume that these structures are connected, but we will no longer insist that they must have a tuple in some relation in σ, as we may need to forbid certain colour configurations, as in our MMSNP sentence Φ_2 in § 7.1. In this formalism, a structure $\mathcal{A}$ is accepted if there exists a valid extension $\mathcal{A}'$: that is, such that there is no homomorphism from any $\mathcal{F}'$ in $\mathscr{F}'$ to $\mathcal{A}'$ that is *full* w.r.t. symbols in μ^+.

Given a structure $\mathcal{A}'$, we let $\mathcal{A}''$ be its σ''-extension where M^- is $A \setminus M^{\mathcal{A}'}$. We call such structures *complementative*. We denote by $\mathscr{F}''$ the set obtained from applying this translation to each structure of $\mathscr{F}'$. Note that $\mathcal{A}$ is valid w.r.t. $\mathscr{F}'$ if, and only if, there exists an extension $\mathcal{A}''$ that is complementative and such that there is no homomorphism from any $\mathcal{F}''$ in $\mathscr{F}''$ to $\mathcal{A}''$.

Lifting

Lemma 141. *Let $\mathscr{C}$ be a class of relational structures over some signature σ. If for any additional monadic symbols μ^+, the class $\mathscr{C}''$ has all restricted dualities then $\mathscr{C}$ has all coloured dualities.*

Proof. Let μ^+ be the signature of the additional monadic symbols for some finite coloured obstruction set $\mathscr{F}'$. We follow the same proof principle as in Theorem 98.

By assumption, there exists a structure $\mathcal{T}''$ such that $(\mathscr{F}'', \{\mathcal{T}''\})$ is a generalised duality pair. We consider the complementative substructure $\mathcal{T}''_c$ of $\mathcal{T}''$, that is the substructure induced by those vertices for which, for every pair of new monadic symbol, either $M^+(x) \wedge \neg M^-(x)$ holds or $\neg M^+(x) \wedge M^-(x)$ holds. Let $\mathcal{T}'_c$ be the σ'-reduct of $\mathcal{T}''_c$ and $\mathcal{T}_c$ its σ-reduct. We claim that $\mathrm{FPP}(\mathscr{F}') = \mathrm{CSP}(\mathcal{T}_c)$.

Assume that $\mathcal{A} \in \mathrm{FPP}(\mathscr{F}')$. This means that there exists a complementative extension $\mathcal{A}''$ such that there is no homomorphism from any $\mathcal{F}''$ in $\mathscr{F}''$ to $\mathcal{A}''$. By duality, $\mathcal{A}''$ is homomorphic to $\mathcal{T}''$. Since it is complementative, it must be homomorphic to $\mathcal{T}''_c$. Taking σ-reducts, we obtain that $\mathcal{A}$ is homomorphic to $\mathcal{T}_c$.

Conversely, if there is a homomorphism h from a structure $\mathcal{A}$ to $\mathcal{T}_c$ then we can consider the extension such that h is a homomorphism from $\mathcal{A}''$ to $\mathcal{T}''_c$ (and therefore to the larger $\mathcal{T}''$). By duality, $\mathcal{A}''$ is valid w.r.t. $\mathscr{F}''$ and since it is complementative, we can conclude that $\mathcal{A} \in \mathrm{FPP}(\mathscr{F}')$. $\qquad\square$

Theorem 142 ([68])**.** *Let $\mathscr{C}$ be a class of structures. If $\mathscr{C}$ has bounded expansion then $\mathscr{C}$ has all coloured dualities.*

Proof. We have seen that when $\mathscr{C}$ has bounded expansion then it has low tree-depth decomposition. Since monadic predicates have no bearing on the tree-depth, it follows that $\mathscr{C}''$ has also low tree-depth decomposition. By Theorem 136, $\mathscr{C}''$ has all restricted dualities. By Lemma 141, it follows that $\mathscr{C}$ has all coloured dualities. □

If instead of considering colours on vertices for the obstructions, we were to consider colours on edges or more generally for structures colour tuples of the relations, then we would derive that structures of bounded expansion have all "edge-coloured dualities". Indeed, it may be seen that we are not really using monadicity in Lemma 141; and, that if we expand a structure by relations (to code for the edge colours), the tuple of which all belong to an existing relation (the edge we are colouring), then we are not changing the tree-depth.

Courcelle investigated the difference in expressivity that adding edge set quantification provided to *Monadic Second Order logic*: he proved that MSO_2 (with edge set quantification) is more expressive than MSO_1 (with the more usual vertex set quantification, often denoted by MSO) in general. However, he also showed that under certain restriction, edge set quantification does not provide more expressivity.

Theorem 143 ([29])**.** *On each of the following classes of simple graphs: those of degree at most k, those of tree-width at most k, for each k, planar graphs, and, more generally, every proper minor closed class, every sentence in MSO_2 is logically equivalent to a sentence of MSO_1.*

In [68], I introduced the logic $MMSNP_2$ which captures union of forbidden patterns problems with edge colours, and pointed out that the following holds (we denote MMSNP, by $MMSNP_1$ for notational consistency).

Theorem 144 ([68])**.** *If a class $\mathscr{C}$ has bounded expansion then $MMSNP_1$ and $MMSNP_2$ are equally expressive when restricted to inputs from $\mathscr{C}$. These logics define precisely finite unions of constraint satisfaction problems; and, in particular if $\mathscr{C}$ contains connected structures only then these logics define precisely constraint satisfaction problems.*

Courcelle has recently extended Theorem 143 to hypergraphs, which can be stated as follows in the case of graphs.

Theorem 145. *[30] Let $k > 0$. Every sentence of MSO_2 is logically equivalent to a sentence of MSO_1 over uniformly k-sparse graphs.*

Recall that a graph $\mathcal{G}$ is *uniformly k-sparse* if, and only if, every subgraph $\mathcal{H}$ of $\mathcal{G}$ is k-sparse, that is $|E(\mathcal{H})| \leq k.|V(\mathcal{H})|$. This definition is equivalent to the following condition: $\mathcal{G}$ has an orientation such that every vertex has in-degree at most k (see Lemma 3.1 in [30]).

113

Remark 146. It follows directly from the definitions that a class of graphs with bounded expansion is uniformly k-sparse for some fixed k. However, we know that the converse implication can not hold as 2-sparse graphs do not have all restricted dualities. Indeed, we prove in Proposition 147 that there exists a problem definable by a first-order sentence of MMSNP_1 that is not a CSP even when restricted to uniformly 2-sparse graphs. However, this does not exclude that MMSNP_1 and MMSNP_2 are also equally expressive when restricted to uniformly k-sparse graphs.

Proposition 147. *Uniformly* 2*-sparse graphs do not have all restricted dualities.*

Proof. Consider the problem "no triangle" given by Ψ_1 in § 7.1 and the structure $\mathcal{A}$ used to prove that it is not a CSP. This structure $\mathcal{A}$ is a graph with n special elements, such that every pair of distinct special elements are linked by a path of length three (using additional vertices). We give an orientation to each edge on each of the path as follows: edges with a special vertex become arcs originating from this special vertex; and other edges are oriented arbitrarily. Note that every special vertex has in-degree zero and every non-special vertex has in-degree at most 2 (since it has degree at most 2). This shows that our graph is uniformly 2-sparse.

Using the same argument as in § 7.1, it follows that the restriction of Ψ_1 to uniformly 2-sparse graphs is not a CSP. $\qquad\square$

9.5 Lifting Preservation Theorems

We try to lift Rossman's theorem (Theorem 107) – using it as a black box – as much as we can. We are able to prove a weakening of Feder and Vardi's preservation for monadic SNP and binary SNP (see Theorem 111). Our result is weaker because we assume a stronger preservation property than just closure under inverse homomorphism, and our construction of the MMSNP sentence is not effective (as it relies on the non-effective theorem of Rossman). More interestingly, we are able to prove a preservation theorem for r-ary SNP which preserves arity (Feder and Vardi did not prove such a theorem, see Theorem 113). However, we assume also a stronger preservation property and are not effective.

We say that a sentence Φ in ESO, where $\Phi = \exists \bar{S} \varphi$ and φ is first-order is *closed under inverse homomorphism* (in the finite) whenever for any (finite) structures $\mathcal{A}$ and $\mathcal{B}$ such that $\mathcal{A}$ is homomorphic to $\mathcal{B}$, $\mathcal{B} \models \Phi$ implies $\mathcal{A} \models \Phi$. We say that Φ is *closed under inverse certificate-strongly-preserving homomorphism* (in the finite) whenever for any (finite) structures $\mathcal{A}'$ and $\mathcal{B}'$ where $\mathcal{A}' := \langle \mathcal{A}, \bar{S}^{\mathcal{A}} \rangle$ such that there is a homomorphism h from $\mathcal{B}'$ to $\mathcal{A}'$ that is full w.r.t. the predicates in $\bar{S}$, $\mathcal{B}' \models \varphi$ implies $\mathcal{A}' \models \varphi$. We say that Φ is *closed under inverse certificate-preserving homomorphism* (in the finite) when φ is closed under inverse homomorphism (in the finite).

It is easy to check that a sentence in monotone SNP without $\neq$ is closed under inverse certificate-strongly-preserving homomorphism. There are sentences Φ in existential MSO that are closed under inverse homomorphism but not under inverse certificate-preserving homomorphism. For example, consider 2-colorability: a single monadic predicate M will code the two vertex colours and the graph is 2-coloured w.r.t. these colours except for a special vertex u that has colour M but whose colour should really read $\neg M$ rather than M. A suitable MSO sentence achieving this together with examples of two suitable structures are depicted on Figure 9.1.

$$\exists M \forall v, v' \neg\big(E(v,v') \wedge \neg M(v) \wedge \neg M(v')\big) \wedge \exists u\, M(u) \wedge \forall v \neg\big(E(u,v) \wedge \neg M(v)\big)$$
$$\wedge \forall u', u'' \neg\big(E(u',u'') \wedge u' \neq u \wedge M(u') \wedge u'' \neq u \wedge M(u'')\big).$$
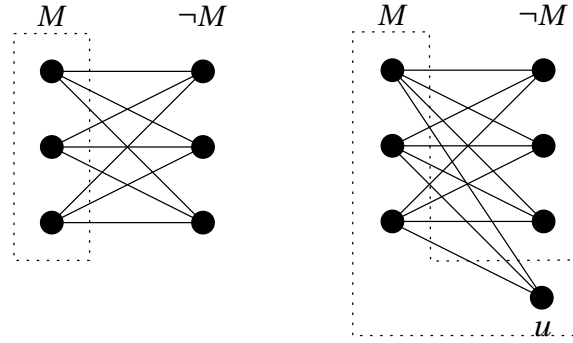

Figure 9.1: a sentence preserved under inverse homomorphism but not for the certificate: on the right, a structure with a valid certificate is depicted; and, on the left, a substructure with a non valid certificate (though we have a certificate-strongly-preserving homomorphism from left to right).

Using Rossman's theorem, we are able to obtain a weak preservation result for Monadic SNP with inequalities.

Theorem 148. *Let Φ be a sentence of Monadic SNP with inequalities. The following are equivalent.*

(i) *Φ is closed under certificate-strongly-preserving homomorphism.*

(ii) *Φ is logically equivalent to a sentence Ψ of MMSNP.*

Proof. It is a direct consequence of the definitions that (ii) implies (i). We now prove the converse. For simplicity we assume that Φ is of the form $\exists M^+ \forall \bar{x} \varphi$ where φ is quantifier-free (the proof is the same with more than one monadic predicate but the notation would be heavier). Let σ denote the input signature and $\sigma' := \sigma \cup \{M^+\}$.

The construction. We assume w.l.o.g. that φ is written in CNF, and rewrite each clause as a negated conjunction (in keeping with the way in which we have written MMSNP sentences thus far). Moreover, we will insist that in any negated conjunct $\neg C$ that mentions a variable x, then either $\neg M^+(x)$ or $M^+(x)$ appears. If it were not the case, then we would duplicate C, replacing it by the two negated conjunct $\neg(M^+(x) \wedge C)$ and $\neg(\neg M^+(x) \wedge C)$ (note that this transformation does not affect (i)).

We consider the extended signature $\sigma'' := \sigma \cup \{M^+\} \cup \{M^-\}$ (intuitively, M^- stands for the complement of M^+). Let φ'' be the sentence obtained from φ by replacing every occurrence of $\neg M^+(x)$ in a negated conjunct by $M^-(x)$ and adding the negated conjunct $\neg(M^+(x) \wedge M^-(x))$.

We will prove shortly that φ'' is closed under inverse homomorphism in order to apply Rossman's theorem to obtain an equivalent universal negative sentence ψ''. Reversing the above transformation, we will turn occurrences of $M^-(x)$ back into $\neg M^+(x)$ in ψ'' and obtain a σ'-sentence ψ. We will finally set $\Psi := \exists M \psi$ which is in MMSNP.

Some observations. Seeing a σ''-structure $\mathcal{A}''$ as a coloured σ-structures, note that when $\mathcal{A}'' \models \varphi''$, vertices of $\mathcal{A}''$ may only take three colours: $M^+ \wedge \neg M^-$ or $\neg M^+ \wedge M^-$ or $\neg M^+ \wedge \neg M^-$. We say that a σ''-structure is *complementative* if it has only vertices of the colours $M^+ \wedge \neg M^-$ or $\neg M^+ \wedge M^-$.

By construction of φ'', $\mathcal{A}'' \models \varphi''$ if, and only if, $\mathcal{A}''_c \models \varphi''$, where $\mathcal{A}''_c$ is the largest complementative substructure of $\mathcal{A}''$ (here, we rely on the fact that φ, and consequently φ'' is universal and that in every negated conjunct of φ'', for any variable x, either $M^+(x)$ or $M^-(x)$ occurs).

Let $\mathcal{B}'$ be the σ'-reduct of a complementative $\mathcal{B}''_c$. Trivially, $\mathcal{B}' \models \varphi$ if, and only if, $\mathcal{B}''_c \models \varphi''$. Moreover, if $\mathcal{A}''$ is the complementative σ''-expansion of a σ'-structure $\mathcal{A}'$ then $\mathcal{A}' \models \varphi$ if and only if $\mathcal{A}'' \models \varphi''$. Note that the above holds also for sentences ψ and ψ'' where ψ is obtained from ψ'' by replacing $M^-(x)$ by $\neg M^+(x)$, provided that ψ'' entails $\forall x \neg(M^+(x) \wedge M^-(x))$.

Applying Rossman's theorem. We are now ready to prove that φ'' is closed under inverse homomorphism. Let $\mathcal{A}''$ and $\mathcal{B}''$ be two σ''-structures and h a homomorphism from $\mathcal{A}''$ to $\mathcal{B}''$. Note that the restriction g of h to the domain of $\mathcal{A}''_c$ is a homomorphism to $\mathcal{B}''_c$ that must be strong w.r.t. M^+ (and also M^-). Consequently, g is a certificate-strongly-preserving homomorphism from $\mathcal{A}'_c$ to $\mathcal{B}'_c$, the respective σ'-reducts of $\mathcal{A}''_c$ and $\mathcal{B}''_c$.

Assume that $\mathcal{B}'' \models \varphi''$. Equivalently, $\mathcal{B}''_c \models \varphi''$ which is in turn equivalent to $\mathcal{B}'_c \models \varphi$. Since g is a certificate-strongly-preserving homomorphism from $\mathcal{A}'_c$ to $\mathcal{B}'_c$, it follows that $\mathcal{A}'_c \models \varphi$. Equivalently, $\mathcal{A}''_c \models \varphi''$ which is in turn equivalent to $\mathcal{A}'' \models \varphi''$ and concludes the proof that φ'' is closed under inverse homomorphisms.

By Rossman's theorem, let ψ'' be the universal negative sentence that is logically equivalent to φ''. Let ψ be the corresponding σ'-sentence obtained from ψ

by replacing every occurrence of M^- by $\neg M^+$.

We show that ψ is logically equivalent to φ. Assume that $\mathcal{A}' \models \varphi$. Equivalently, $\mathcal{A}'' \models \varphi''$ where $\mathcal{A}''$ is the complementative σ''-expansion of $\mathcal{A}'$. Via Rossman's theorem, this is equivalent to $\mathcal{A}'' \models \psi''$. Since $\mathcal{A}''$ is complementative this is equivalent to $\mathcal{A}' \models \psi$. $\qquad\square$

Note that we can not hope to extend Theorem 148 to MSO as the following universal-monadic-second-order sentence captures digraph-acyclicity (in the sense that a digraph contains no directed cycle).

$$\forall U \left(\forall x \, \neg U(x) \vee \exists x \, U(x) \right) \wedge \forall y \, \left(\neg E(x,y) \vee \neg U(y) \right)$$

Indeed, if this sentence holds for a non trivial choice of U, it computes a vertex x which has no outgoing neighbour in U. This means that if $U \setminus \{x\}$ is acyclic, then so is U. This sentence is closed under disjoint union and inverse homomorphism (as the corresponding problem is the constraint satisfaction problem with ω-categorical template $(\mathbb{Q}, \leq)$). It is not expressible in MMSNP. An MMSNP sentence Φ with n monadic predicates and forbidden patterns of size at most l can not capture this problem. Indeed, consider a very long directed path and some partition of this path with 2^n different colours witnessing that Φ holds. We will find two contiguous regions of size greater than p of this coloured path which coincide and we may identify two vertices in these regions, which provides us with a structure that is accepted by Φ, yet no longer acyclic (see [4] for further details).

We may extend our arguments to obtain a weak preservation result for r-ary SNP, for any r.

Theorem 149. *Let Φ be a sentence of r-ary SNP with inequalities. The following are equivalent.*

(i) *Φ is closed under inverse certificate-strongly-preserving homomorphism.*

(ii) *Φ is logically equivalent to a sentence Ψ of monotone r-ary SNP without inequalities.*

Proof. The proof works almost in the same way. Assume for notational convenience that Φ is of the form $\exists S^+ \forall \bar{x} \varphi$ where φ is quantifier-free and S^+ has arity 2. Let σ denote the input signature and $\sigma' := \sigma \cup \{S^+\}$. We proceed similarly to the monadic case and enforce that for every pair of variables x, y in a negated conjunct of φ either $\neg S^+(x,y)$ or $S^+(x,y)$ occurs. Consequently, a σ''-structure is not satisfied if, and only if, a negated conjunct occurs in a part that is complementative. However, instead of a single complementative substructure of $\mathcal{B}$, we have now possibly several such substructures and $\mathcal{B}'' \models \varphi''$ if, and only if, every complementative substructure $\mathcal{B}''_c$ of $\mathcal{B}''$, $\mathcal{B}''_c \models \varphi''$ if, and only, if the reduct of every complementative $\mathcal{B}'_c \models \varphi$. Apart from this minor difference, the proof remains the same. $\qquad\square$

9.6 Some questions

We have shown that the Recolouring problem is Π_2^p-hard and in Σ_3^p. It would be interesting to settle more accurately its complexity.

Question 150. *Is the Recolouring problem Σ_3^p-complete?*

For every constraint satisfaction problem with template $\mathcal{T}$, there exists a MMSNP sentence $\Phi_{\mathcal{T}}$ which defines it. It is possible to decide given $\mathcal{T}$, and construct if it exists a sentence $\Phi_{\mathcal{T}}$ that is first-order (see Theorem 131). In general, it could be interesting to find a sentence $\Phi_{\mathcal{T}}$ with a minimal number of monadic predicates (or minimal number of colours, where a colour is one of the 2^n combination induced by n monadic predicates, and we would only count those colours that are not explicitly disallowed). Indeed, this could lead to interesting search algorithms as the number of colours would influence directly the degree of the search tree.

Question 151. *Given a structure $\mathcal{T}$, can we compute the minimal number? Can we at least decide it is less than some constant r? Can we compute a sentence $\Phi_{\mathcal{T}}$ with minimal number of colours?*

We have seen that $MMSNP_2$ and $MMSNP_1$ coincide over sufficiently sparse inputs. However, we have not proved that $MMSNP_2$ is strictly more expressive than $MMSNP_1$ in general. Note that this would require for a specific problem Ω of $MMSNP_2$, for example the problem Edge-no-monochromatic-triangle [3], and for any formula Φ in $MMSNP_1$ that there exists a structure $\mathcal{A}$ such that $\mathcal{A}$ is in $(\Phi \setminus \Omega) \cup (\Omega \setminus \Phi)$. Since on a class of sparse graphs $\mathscr{C}$ the problem Ω is a CSP with some template $\mathcal{T}$, for the corresponding MMSNP formula $\Phi_{\mathcal{T}}$, one would need a non sparse graph $\mathcal{A}$ to distinguish Ω from $\Phi_{\mathcal{T}}$.

Question 152. *Is $MMSNP_2$ strictly more expressive than $MMSNP_1$ over finite structures? For example, is the problem Edge-no-monochromatic-triangle expressible in $MMSNP_1$?*

We have seen that $MMSNP_1$ and $MMSNP_2$ do not collapse to CSP for uniformly k-sparse graphs. They might nonetheless have the same expressivity over this class as is the case for MSO_1 and MSO_2.

Question 153. *Are $MMSNP_1$ and $MMSNP_2$ equally expressive over uniformly k-sparse graphs?*

We have seen that there exists a universal MSO sentence that is closed under inverse homomorphism and is not logically equivalent to an MMSNP sentence, so the most we can hope for is to extend Theorem 148 to existential MSO.

---

[3]This graph problem asks whether there exists a bipartition of the edge set so that no triangle occurs in any part.

Question 154. *Is every existential* MSO *sentence closed under inverse certificate-strongly-preserving homomorphism logically equivalent to a sentence of MMSNP in the finite?*

We can also wonder if the preservation condition can be relaxed.

Question 155. *Is every existential* MSO *sentence closed under inverse homomorphism equivalent to a sentence of MMSNP in the finite?*

Bibliography

[1]   Albert Atserias, Anuj Dawar, and Phokion G. Kolaitis. On preservation under homomorphisms and unions of conjunctive queries. *J. ACM*, 53(2):208–237, 2006. 93

[2]   Libor Barto, Marcin Kozik, and Todd Niven. The CSP dichotomy holds for digraphs with no sources and no sinks (a positive answer to a conjecture of bang-jensen and hell). *SIAM J. Comput.*, 38(5):1782–1802, 2009. 72

[3]   M. Benedetti, A. Lallouet, and J. Vautard. Quantified constraint optimization. In *CP*, pages 463–477, 2008. 17

[4]   M. Bodirsky. *Constraint Satisfaction with Infinite Domains*. PhD thesis, Humboldt-Universität zu Berlin, 2004. 117

[5]   M. Bodirsky. Constraint satisfaction problems with infinite templates. In Creignou et al. [32], pages 196–228. 7

[6]   Manuel Bodirsky. Complexity classification in infinite-domain constraint satisfaction. *CoRR*, abs/1201.0856, 2012. 79, 82

[7]   Manuel Bodirsky, Hubie Chen, and Tomás Feder. On the complexity of MM-SNP. *SIAM J. Discrete Math.*, 26(1):404–414, 2012. 79

[8]   Manuel Bodirsky and Víctor Dalmau. Datalog and constraint satisfaction with infinite templates. In Bruno Durand and Wolfgang Thomas, editors, *STACS*, volume 3884 of *Lecture Notes in Computer Science*, pages 646–659. Springer, 2006. 85, 86

[9]   F. Börner. Basics of galois connections. In Creignou et al. [32], pages 38–67. 17

[10]   F. Börner, A. A. Bulatov, H. Chen, P. Jeavons, and A. A. Krokhin. The complexity of constraint satisfaction games and QCSP. *Inf. Comput.*, 207(9):923–944, 2009. 16, 50, 52, 74

[11]   Ferdinand Börner. Total multifunctions and relations. In *AAA60: Workshop on General Algebra, Dresden, Germany*, 2000. 33

[12]   A. A. Bulatov. H-coloring dichotomy revisited. *Theor. Comput. Sci.*, 349(1):31–39, 2005. 6

[13]   A. A. Bulatov, A. A. Krokhin, and B. Larose. Dualities for constraint satisfaction problems. In Creignou et al. [32], pages 93–124. 8, 106

[14] A. A. Bulatov and D. Marx. Constraint satisfaction problems and global cardinality constraints. *Commun. ACM*, 53(9):99–106, 2010. 19

[15] A. A. Bulatov and M. Valeriote. Recent results on the algebraic approach to the CSP. In Creignou et al. [32], pages 68–92. 15, 72

[16] Andrei A. Bulatov. A dichotomy theorem for constraint satisfaction problems on a 3-element set. *J. ACM*, 53(1):66–120, 2006. 15

[17] Andrei A. Bulatov. Complexity of conservative constraint satisfaction problems. *ACM Trans. Comput. Log.*, 12(4):24, 2011. 15

[18] Andrei A. Bulatov. On the CSP dichotomy conjecture. In Alexander S. Kulikov and Nikolay K. Vereshchagin, editors, *CSR*, volume 6651 of *Lecture Notes in Computer Science*, pages 331–344. Springer, 2011. 72

[19] Ashok K. Chandra and Philip M. Merlin. Optimal implementation of conjunctive queries in relational data bases. In John E. Hopcroft, Emily P. Friedman, and Michael A. Harrison, editors, *STOC*, pages 77–90. ACM, 1977. 29

[20] H. Chen. A rendezvous of logic, complexity, and algebra. *ACM Comput. Surv.*, 42(1), 2009. 50, 55

[21] H. Chen and M. Grohe. Constraint satisfaction with succinctly specified relations. *J. Comput. Syst. Sci.*, 76(8):847–860, 2010. 19

[22] Hubie Chen. The complexity of quantified constraint satisfaction: Collapsibility, sink algebras, and the three-element case. *SIAM J. Comput.*, 37(5):1674–1701, 2008. 16, 74

[23] Hubie Chen. Meditations on quantified constraint satisfaction. *CoRR*, abs/1201.6306, 2012. 16, 75

[24] Hubie Chen, Florent R. Madelaine, and Barnaby Martin. Quantified constraints and containment problems. In *LICS*, pages 317–328. IEEE Computer Society, 2008. 20, 44, 45

[25] G. Cherlin, S. Shelah, and N. Shi. Universal graphs with forbidden subgraphs and algebraic closure. *Adv. in Appl. Math.*, 22(4):454–491, 1999. 85, 86

[26] David Cohen, editor. *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*. Springer, 2010. 125, 126

[27] M. C. Cooper, P. G. Jeavons, and A. Z. Salamon. Generalizing constraint satisfaction on trees: Hybrid tractability and variable elimination. *Artif. Intell.*, 174(9-10):570–584, 2010. 19

[28] B. Courcelle. Graph rewriting: An algebraic and logic approach. In *Handbook of Theoretical Computer Science, Volume B: Formal Models and Sematics (B)*, pages 193–242. 1990. 4

[29] Bruno Courcelle. The monadic second-order logic of graphs VI: On several representations of graphs by relational structures. *Discrete Applied Mathematics*, 63(2):199–200, 1995. 113

[30] Bruno Courcelle. The monadic second-order logic of graphs XIV: uniformly sparse graphs and edge set quantifications. *Theor. Comput. Sci.*, 1-3(299):1–36, 2003. 113

[31] N. Creignou, S. Khanna, and M. Sudan. *Complexity classifications of boolean constraint satisfaction problems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001. 6, 16, 50

[32] N. Creignou, P. G. Kolaitis, and H. Vollmer, editors. *Complexity of Constraints - An Overview of Current Research Themes [Result of a Dagstuhl Seminar]*, volume 5250 of *Lncs*. Springer, 2008. 2, 120, 121, 122, 126

[33] N. Creignou and H. Vollmer. Boolean constraint satisfaction problems: When does Post's lattice help? In Creignou et al. [32], pages 3–37. 16

[34] V. Dalmau, P. G. Kolaitis, and M. Y. Vardi. Constraint satisfaction, bounded treewidth, and finite-variable logics. In P. Van Hentenryck, editor, *CP*, volume 2470 of *Lncs*, pages 310–326. Springer, 2002. 9, 10

[35] Victor Dalmau. Some dichotomy theorems on constant-free quantified boolean formulas. Technical Report LSI-97-43-R., Departament LSI, Universitat Pompeu Fabra., 1997. 16, 50

[36] Stefan S. Dantchev and Florent R. Madelaine. Bounded-degree forbidden patterns problems are constraint satisfaction problems. In Dima Grigoriev, John Harrison, and Edward A. Hirsch, editors, *CSR*, volume 3967 of *Lecture Notes in Computer Science*, pages 159–170. Springer, 2006. 111

[37] Anuj Dawar. Homomorphism preservation on quasi-wide classes. *J. Comput. Syst. Sci.*, 76(5):324–332, 2010. 93

[38] Anuj Dawar and Stephan Kreutzer. On datalog vs. LFP. In Luca Aceto, Ivan Damgård, Leslie Ann Goldberg, Magnús M. Halldórsson, Anna Ingólfsdóttir, and Igor Walukiewicz, editors, *ICALP (2)*, volume 5126 of *Lecture Notes in Computer Science*, pages 160–171. Springer, 2008. 94

[39] R. Dechter. Bucket elimination: A unifying framework for reasoning. *Artif. Intell.*, 113(1-2):41–85, 1999. 4

[40] R. G. Downey and M.R. Fellows. *Parameterized Complexity*. Monographs in Computer Science. Springer, 1999. 10

[41] H.-D. Ebbinghaus and J. Flum. *Finite Model Theory*. Perspective in Mathematical Logic. Springer-Verlag, 1995. 23

[42] R. Fagin. Generalized first-order spectra and polynomial-time recognizable sets. In R. M. Karp, editor, *Complexity of Computation*, pages 43–73, 1974. 79, 80

[43] T. Feder, F. R. Madelaine, and I. A. Stewart. Dichotomies for classes of homomorphism problems involving unary functions. *Theor. Comput. Sci.*, 314(1-2):1–43, 2004. 8

[44] 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. 6, 7, 8, 14, 15, 72, 79, 82, 83, 96, 98, 102, 103

[45] Tomás Feder and Moshe Y. Vardi. Homomorphism closed vs. existential positive. In *LICS*, pages 311–320. IEEE Computer Society, 2003. 94, 95

[46] Jan Foniok, Jaroslav Nešetřil, and Claude Tardif. Generalised dualities and maximal finite antichains in the homomorphism order of relational structures. *Eur. J. Comb.*, 29(4):881–899, 2008. 96, 106

[47] E. C. Freuder. A sufficient condition for backtrack-bounded search. *J. ACM*, 32(4):755–761, 1985. 2, 4

[48] Erich Grädel, P. G. Kolaitis, L. Libkin, M. Marx, J. Spencer, Moshe Y. Vardi, Y. Venema, and Scott Weinstein. *Finite Model Theory and Its Applications (Texts in Theoretical Computer Science. An EATCS Series)*. Springer, 2007. 29

[49] M. Grohe. The complexity of homomorphism and constraint satisfaction problems seen from the other side. *J. ACM*, 54(1), 2007. 10

[50] M. Grohe and D. Marx. Constraint solving via fractional edge covers. In *SODA*, pages 289–298. ACM Press, 2006. 11

[51] Roland Häggkvist and Pavol Hell. Universality of A-mote graphs. *Eur. J. Comb.*, 14(1):23–27, 1993. 108

[52] P. Hell and J. Nešetřil. On the complexity of H-coloring. *J. Combin. Theory Ser. B*, 48, 1990. 52, 72

[53] P. Hell and J. Nešetřil. *Graphs and Homomorphisms*. OUP, 2004. 5, 87

[54] P. Jeavons. On the algebraic structure of combinatorial problems. *Theoretical Computer Science*, 200(1–2):185–204, 1998. 11

[55] Peter Jeavons, David A. Cohen, and Marc Gyssens. Closure properties of constraints. *J. ACM*, 44(4):527–548, 1997. 14

[56] Phokion G. Kolaitis and Moshe Y. Vardi. Conjunctive-query containment and constraint satisfaction. *J. Comput. Syst. Sci.*, 61(2):302–332, 2000. 24

[57] M. Krasner. Une généralisation de la notion de corps. *Journal de Mathématiques Pures et Appliquées*, 9:367–385, 1938. 52

[58] Gábor Kun. Constraints, MMSNP and expander relational structures. arXiv:0706.1701v1, June 2007. 83, 84

[59] Gábor Kun and Jaroslav Nešetřil. Forbidden lifts (NP and CSP for combinatorialists). *Eur. J. Comb.*, 29(4):930–945, 2008. 107

[60] R. E. Ladner. On the structure of polynomial time reducibility. *J. ACM*, 22(1):155–171, 1975. 6, 83

[61] Benoit Larose, Cynthia Loten, and Claude Tardif. A characterisation of first-order constraint satisfaction problems. *Logical Methods in Computer Science*, 3(4), 2007. 107

[62] L. Lovász. Operations with structures. *Acta Mathematica Hungarica*, 18:321–328, 1967. 10.1007/BF02280291. 106

[63] Nancy A. Lynch. Log space recognition and translation of parenthesis languages. *J. ACM*, 24(4):583–590, 1977. 51, 57

[64] F. Madelaine. *Constraint satisfaction problems and related logic.* PhD thesis, University of Leicester, Department of Maths and Computer Science, March 2003. 19, 79, 87, 90, 96

[65] F. R. Madelaine. On the containment of forbidden patterns problems. In *CP*, pages 345–359, 2010. 7, 20, 96

[66] Florent Madelaine. De la complexité des problèmes de contraintes. In Arnaud Lallouet, editor, *Actes des Septièmes Journées Francophones de Programmation par Contraintes*, pages 201–211, 2011. 20

[67] Florent R. Madelaine. Universal structures and the logic of forbidden patterns. In Zoltán Ésik, editor, *CSL*, volume 4207 of *Lecture Notes in Computer Science*, pages 471–485. Springer, 2006. 111

[68] Florent R. Madelaine. Universal structures and the logic of forbidden patterns. *Logical Methods in Computer Science*, 5(2), 2009. 8, 20, 79, 111, 112, 113

[69] Florent R. Madelaine and Barnaby Martin. The complexity of positive first-order logic without equality. In *LICS*, pages 429–438. IEEE Computer Society, 2009. 20

[70] Florent R. Madelaine and Barnaby Martin. A tetrachotomy for positive first-order logic without equality. In *LICS*, pages 311–320. IEEE Computer Society, 2011. 18, 20, 57

[71] Florent R. Madelaine and Barnaby Martin. The complexity of positive first-order logic without equality. *ACM Trans. Comput. Log.*, 13(1):5, 2012. 18, 20, 57, 60

[72] Florent R. Madelaine and Barnaby Martin. Containment, equivalence and coreness from CSP to QCSP and beyond. *CoRR*, abs/1204.5981, 2012. 20, 29, 47, 64, 76

[73] Florent R. Madelaine and Iain A. Stewart. Some problems not definable using structure homomorphisms. *Ars Comb.*, 67, 2003. 79

[74] Florent R. Madelaine and Iain A. Stewart. Constraint satisfaction, logic and forbidden patterns. *SIAM J. Comput.*, 37(1):132–163, 2007. 7, 20, 79, 86, 87, 88, 89, 90, 91, 96, 98, 102, 106

[75] B. Martin. Dichotomies and duality in first-order model checking problems. *CoRR*, abs/cs/0609022, 2006. 52

[76] Barnaby Martin. First-order model checking problems parameterized by the model. In Arnold Beckmann, Costas Dimitracopoulos, and Benedikt Löwe, editors, *CiE*, volume 5028 of *Lecture Notes in Computer Science*, pages 417–427. Springer, 2008. 17, 52

[77] Barnaby Martin. The lattice structure of sets of surjective hyper-operations. In Cohen [26], pages 368–382. 69

[78] Barnaby Martin. QCSP on partially reflexive forests. In Jimmy Ho-Man Lee, editor, *CP*, volume 6876 of *Lecture Notes in Computer Science*, pages 546–560. Springer, 2011. 16, 74

[79] Barnaby Martin and Florent R. Madelaine. Towards a trichotomy for quantified H-coloring. In Arnold Beckmann, Ulrich Berger, Benedikt Löwe, and John V. Tucker, editors, *CiE*, volume 3988 of *Lecture Notes in Computer Science*, pages 342–352. Springer, 2006. 16, 72

[80] Barnaby Martin and Jos Martin. The complexity of positive first-order logic without equality II: The four-element case. In Anuj Dawar and Helmut Veith, editors, *CSL*, volume 6247 of *Lecture Notes in Computer Science*, pages 426–438. Springer, 2010. 18, 57

[81] D. Marx. Tractable hypergraph properties for constraint satisfaction and conjunctive queries. In L. J. Schulman, editor, *STOC*, pages 735–744. ACM, 2010. 11

[82] Jaroslav Nešetřil. A surprising permanence of old motivations (a not-so-rigid story). *Discrete Mathematics*, 309(18):5510–5526, 2009. 107

[83] Jaroslav Nešetřil and Patrice Ossona de Mendez. *Sparse Combinatorial Structures: Classification and Applications*, chapter 162, pages 2502–2529. 109

[84] Jaroslav Nešetřil and Patrice Ossona de Mendez. Tree-depth, subgraph coloring and homomorphism bounds. *Eur. J. Comb.*, 27(6):1022–1041, 2006. 108

[85]   Jaroslav Nešetřil and Patrice Ossona de Mendez. Grad and classes with bounded expansion III. restricted graph homomorphism dualities. *Eur. J. Comb.*, 29(4):1012–1024, 2008. 109

[86]   Jaroslav Nešetřil and Patrice Ossona de Mendez. First order properties on nowhere dense structures. *J. Symb. Log.*, 75(3):868–887, 2010. 94

[87]   Jaroslav Nešetřil and Claude Tardif. Duality theorems for finite structures (characterising gaps and good characterisations). *J. Comb. Theory, Ser. B*, 80(1):80–97, 2000. 105, 106

[88]   Jaroslav Nešetřil and Claude Tardif. Short answers to exponentially long questions: Extremal aspects of homomorphism duality. *SIAM J. Discrete Math.*, 19(4):914–920, 2005. 107

[89]   G. Nordh and B. Zanuttini. Frozen boolean partial co-clones. In *ISMVL*, pages 120–125, 2009. 17

[90]   Christos H. Papadimitriou. *Computational complexity*. Addison-Wesley, 1994. 23

[91]   Omer Reingold. Undirected connectivity in log-space. *J. ACM*, 55(4), 2008. 72

[92]   Eric Rosen. Some aspects of model theory and finite structures. *Bulletin of Symbolic Logic*, 8(3):380–403, 2002. 91

[93]   Benjamin Rossman. Homomorphism preservation theorems. *J. ACM*, 55(3), 2008. 91

[94]   John E. Savage. Computational work and time on finite machines. *J. ACM*, 19(4):660–674, 1972. 82

[95]   F. Scarcello, G. Gottlob, and G. Greco. Uniform constraint satisfaction problems and database theory. In Creignou et al. [32], pages 156–195. 11

[96]   Marcus Schaefer and Christopher Umans. Completeness in the polynomial-hierarchy: part I. *SIGACT news*, 33(3):32–49, 2002. SIGACT news complexity theory column 37, Guest column, introduced by Lane A. Hemaspaandra. 103

[97]   T.J. Schaefer. The complexity of satisfiability problems. In *STOC*, 1978. 16, 50

[98]   B. ten Cate, L. Chiticariu, P. G. Kolaitis, and W. Chiew Tan. Laconic schema mappings: Computing the core with sql queries. *PVLDB*, 2(1):1006–1017, 2009. 19

[99]   B. ten Cate, P. G. Kolaitis, and W. Chiew Tan. Database constraints and homomorphism dualities. In Cohen [26], pages 475–490. 19

126

[100] Moshe Y. Vardi. The complexity of relational query languages (extended abstract). In Harry R. Lewis, Barbara B. Simons, Walter A. Burkhard, and Lawrence H. Landweber, editors, *STOC*, pages 137–146. ACM, 1982. 23, 26