UMONS
Université de Mons

Faculté
des Sciences

Mémoire

Data Repairing

Réparation de contraintes

Écrit par : Maxime Van Herzeele
Année académique : 2017-2018
Directeur de mémoire : Jef Wijsen
Section : MAB2 Sciences Informatiques

Faculté des Sciences • University of Mons • Place du Parc 20 • B-7000 Mons

Table des matières

1 Introduction **1**

2 Les contraintes d'intégrité **4**
2.1 Base de données 4
2.2 Contraintes sur les bases de données 7
2.3 Les Denials constraints 8
2.3.1 Quelques définitions et propriétés 10

3 Data Repairing **14**
3.1 Variation sur les denial constraints 19
3.1.1 Parcours dans le treillis 21
3.1.2 Limitation des candidats 22
3.1.3 Denial constraints maximales 22
3.1.4 Limitation par bornes 24
3.1.5 Coût dans le treillis 27
3.1.6 Variation d'un ensemble de contraintes 28
3.2 θ-tolerant model 29
3.3 Réparation au coût minimal 30
3.3.1 Identification des suspects 32

4 Différence par rapport à l'article de base **36**

5 Implémentation **38**
5.1 Choix de langage et format de base de données 38
5.2 Code 39
5.2.1 Classe Predicate et DC 39

6 Conclusion **40**

Table des figures

3.1 Toutes les violations pour φ 17
3.2 All the violation for φ' 20
3.3 Treillis pour un prédicat d'attribut A 21
3.4 Graphe de conflits pour φ 25
3.5 Chaque types de transition aura son propre poids. 27
3.6 Graphe de conflit pour Σ_2 avec : 31
3.7 Toutes les violations pour φ'' 31
3.8 s, t sont deux suspect 33
3.9 Les conditions de suspicions sont 34
3.10 Contexte de réparation 35

4.1 La denial constraint tel que définie dans les articles 37

Liste des tableaux

1.1 2

2.1 Base de données de l'article principal [4] 5
2.2 La table Personne 6
2.3 Element de OP, l'ensemble des parties de $\{<, =, >\}$ 9

3.1 Exemple de réparation I' pour l'attribut 17
3.2 Example of repair with Tax 19
3.3 Correction avec φ_2 : 7 changement nécessaire dans la colonne $NumTel$ . . . 23
3.4 Correction avec φ_1 : seulement 3 changements sont nécessaires. 23

Chapitre 1

Introduction

De nombreuses institutions et entreprises collectent, stockent et utilisent de grandes quantités d'informations. Le traitement de ces données permettent à ces entreprises d'effectuer des statistiques ou de fournir un service à l'un de leurs clients. Mais dans les bases de données, nous pouvons retrouver des données *erronées*. Leur présence peut être due à une erreur lors de l'encodage ou lors de la collecte des informations(appareils imprécis, mauvaises méthodes de collecte, personne qui ment ou se trompe,...). Ces erreurs nuisent au traitement des données, biaisant les statistiques et perturbant les services qui en ont besoin. Afin d'éviter ce problème, les données devraient respecter des *contraintes d'intégrité*. Une contrainte d'intégrité est une règle appliquée aux données. Si une donnée ne respecte pas au moins une de ses contraintes alors cette donnée doit être considérée comme erronée.

Malheureusement, il se peut que ces contraintes soient elles aussi erronées. A cause de ses erreurs sur les contraintes, celles-ci peuvent échouer dans la différenciation des données réellement erronées et des données correctes. Il se peut que certaines données soient identifiées comme étant des violations des contraintes (données erronées) malgré qu'elles ne le devraient pas. Il existe aussi des contraintes qui n'arrivent pas à détecter des violations. Les erreurs à la fois sur les données et sur les contraintes sont un problème pour quiconque souhaite utiliser la base de données

Par exemple, durant mon stage en entreprise, j'ai pu travailler sur un projet associé de près à une base de données ayant un problème de données erronées. Cela a eu un énorme impact sur une partie de mon projet. La réparation de ces données est prévue pour le courant de l'année 2018.

Le terme *Data repairing* ou réparation de données signifie réparer les données mais aussi les contraintes d'intégrité. Il serait naïf de penser que l'on puisse supprimer chaque ligne qui possède au moins une donnée erronée. La perte d'information serait importante parce qu'une telle pratique demanderait d'effacer une ligne complète de la table et ce

	Nome	Année	Salaire (€)	Grade	Bureau
t_1	Dupont	2012	23.000	ouvrier	1B18
t_2	Dupont	2013	24.000	ouvrier	1B14
t_3	Dupont	2014	38.000	chef de service	2B12
t_4	Dupont	2015	42.000	chef de service	2B18
t_5	Dupont	2016	37.000	chef de service	2B18
t_6	Dupont	2017	44.000	chef de service	2B18
t_7	Dupont	2018	28.000	ouvrier	1B14
t_8	Ana	2011	26.000	ouvrier	1B08
...					

TABLE 1.1 –

malgré qu'il n'y ait qu'une seule erreur dans la ligne. Par exemple, si nous avons une base de données des employés d'une entreprise et la contrainte "Le salaire d'un employé augmente chaque année". Cette contrainte n'est pas respectée à la table 1.1. En effet, si nous prenons l'année 2016, nous voyons que monsieur Dupont gagne moins d'argent qu'en 2015 et 2014. Si nous supprimons l'année 2016, nous perdons l'information concernant le bureau que monsieur Dupont occupait.

En outre, les contraintes d'intégrité peuvent aussi être erronées ce qui veut dire que l'on pourrait supprimer une ligne ne contenant que des données correctes. Reprenons la contrainte "Le salaire d'un employé augmente chaque année". Cette contrainte n'est pas respecté à la table 1.1 car en 2017 Dupont gagnait 44.000 €et en 2018, il gagnait 28.000 €de salaire annuel. Nous avons l'une des deux lignes de la table qui contient une donnée erronée. Mais c'est sans compter sur le fait que monsieur Dupont a perdu son grade de chef. Du coup la contrainte est peut être erronée et la contrainte "Le salaire d'un employé augmente chaque année à condition de garder son grade" est correcte. Dans ce cas ci, l'erreur ne vient pas des données mais de la contrainte. Pour cette raison, nous avons besoin de techniques afin de réparer à la fois les données et les contraintes, ce sans perdre trop d'informations tout en évitant d'échouer dans la détection d'erreurs dans les données.

Dans cette thèse de mémoire, nous allons analyser le *modèle de réparation θ-tolérant* comme il a été introduit dans un article scientifique[4]. Nous allons faire varier nos contraintes et chacune de nos variations aura un coût que nous devrons déterminer. Le coût de toutes ces variations ne devra pas excéder θ. Le nouvel ensemble de contraintes sera utilisé pour la réparation des données.

Dans un premier temps, nous allons introduire le concept de *denial constraint*, une forme de contraintes d'intégrité qui va nous aider à définir et comprendre le concept du modèle de réparation θ-tolérant. Nous allons également introduire quelques bases de données que nous utiliserons pour illustrer les différentes notions que nous allons abor-

der. Ensuite, nous allons présenter des techniques pour faire varier nos denial constraint en respectant le modèle de réparation θ-tolérant. Puis, puisque nous resterons critique par rapport au papier scientifique de référence, nous expliquerons en quoi notre démarche est différente et pourquoi nous avons choisi de modifier certaines approches et certains concepts. Enfin, nous terminerons avec une implémentation du modèle θ-tolérant et analyserons ses performances avec quelques exemples.

Chapitre 2

Les contraintes d'intégrité

Dans ce chapitre, nous allons rappeler quelques notions bien connues mais nous allons également introduire de nouveaux concepts. Dans un premier temps, nous allons présenter quelques bases de données que nous utiliserons en tant qu'exemple pour expliquer et illustrer de nombreuses propriétés et définitions. Ces bases de données suivent le modèle relationnel qui a été introduit par E.F. Codd [2]. Ensuite nous allons travailler sur les contraintes d'intégrité et nous allons introduire un nouveau type de contraintes appelées *denial constraint*. Nous allons expliquer plusieurs caractéristiques et propriétés de ces contraintes et expliquer pourquoi nous n'utilisons pas une forme plus conventionnelle de contrainte, comme par exemple les dépendances fonctionnelles.

2.1 Base de données

Dans cette section, nous allons présenter des bases de données que nous allons utiliser comme exemple dans cette thèse de mémoire. Nous utiliserons ces bases de données pour illustrer le modèle de réparation de données θ-tolérant ainsi que d'autres notions que nous définirons.

La première base de données est tirée de l'article principal utilisé dans la bibliographie de cette thèse [4]. Elle représente une table de personnes comprenant le nom de ces personnes, leur date de naissance, leurs numéros de téléphone, leurs revenus, le montant de taxe que ces personnes paient ainsi que l'année à laquelle toutes ces données ont été encodées.

	Nom	Anniversaire	NumTel	Année	Revenu	Taxe
t1	Ayres	8-8-1984	322-573	2007	21k	0
t2	Ayres	5-1-1960	***-389	2007	22k	0
t3	Ayres	5-1-1960	564-389	2007	22k	0
t4	Stanley	13-8-1987	868-701	2007	23k	3k
t5	Stanley	31-7-1983	***-198	2007	24k	0
t6	Stanley	31-7-1983	930-198	2008	24k	0
t7	Dustin	2-12-1985	179-924	2008	25k	0
t8	Dustin	5-9-1980	***-870	2008	100k	21k
t9	Dustin	5-9-1980	824-870	2009	100k	21k
t10	Dustin	9-4-1984	387-215	2009	150k	40k

TABLE 2.1 – Base de données de l'article principal [4]

.

La seconde base de données que nous allons utiliser est inspirée d'une expérience personnelle. Lors d'un stage en entreprise, j'ai pu travailler sur un projet lié à une base de donnée contenant des données erronées. Ces données ne pouvant pas être utilisées en dehors de l'entreprise, nous utiliserons une base de données reprenant l'idée générale. C'est une table appelée "Personne" contenant différentes informations basiques sur des personnes en Belgique [1] .

— **NISS :** Le numéro national de la personne. Un numéro national est unique. En règle général, un NISS est formé de la manière suivante : [1]
  — Il commence avec la date de naissance de la personne dans un format YY-MM-DD. Des exceptions existent pour les étrangers (c'est à dire des personne n'ayant pas la nationalité belge) mais nous n'allons pas considérer ces cas. En effet ces cas peuvent être difficiles à comprendre et ne sont aucunement intéressants pour la suite.
  — Le nombre composé du septième, huitième et neuvième chiffres est pair pour les hommes et impair pour les femmes
  — Le nombre composé des deux derniers chiffres est le resulat de $n \mod 97$ avec n le nombre formé des 9 premiers chiffres
— **Nom :** Nom de famille de la personne.
— **Prénom :** Prénom de la personne.
— **Nai_Date :** Date de naissance de la personne dans le format DD-MM-YYYY.
— **Dec_Date :** Date de décès de la personne dans le format DD-MM-YYYY.
— **Etat_Civil :** État civil courant de la personne, celui ci doit être parmi les suivants : (célibataire, marié, divorcé, décédé, veuf)

1. Les données sont fictives

— **Ville** : La ville où la personne vit.
— **Code_Post** : Le code postal de la ville.
— **Salaire** : Le salaire perçu par la personne en une année.
— **Taxe** : Le montant de taxe payé par la personne en une année.
— **Enfant** : Le nombre d'enfants que la personne a à charge.

	Niss	Nom	Prénom	Nai_Date	Dec_Date	Etat_Civil	Ville	Code_Post	Salaire	Taxe	Enfant
t1	14050250845	Dupont	Jean	14-05-1902	18-05-1962	décédé	Ath	7822	25k	4k	2
t2	08042910402	Brel	Jacques	08-04-1929	09-10-1978	décédé	Schaerbeek	1030	100k	8k	1
t3	45060710204	Merckx	Eddy	07-06-1945	null	décédé	Schaerbeek	1030	125k	9k	2

TABLE 2.2 – La table Personne

.

2.2 Contraintes sur les bases de données

Les bases de données devraient n'accepter que des valeurs qui respectent certaines normes et règles. Ce serait un problème si on pouvait ajouter n'importe quelle valeur à chaque colonne d'une base de données. Pour éviter ce problème, nous avons recours à des règles sur les bases de données. Ces règles sont appelées *contraintes d'intégrité* et fonctionnent de la manière suivante : Si une relation i.e un ensemble de tuples respecte toutes les conditions de ces contraintes alors les données sont acceptables. Si la relation ne respecte pas toutes les conditions alors au moins un tuple de la relation contient au moins une valeur erronée.

Le modèle relationnel des bases de données introduit la notion de *dépendance fonctionnelle* :

Définition 1. Une **dépendance fonctionnelle (DF)** est une expression $X \rightarrow Y$ avec $X, Y \subseteq sort(R)$ et où $sort(R) = \{A_1, A_2, ..., A_n\}$ signifient que pour chaque ensemble de tuples où les attributs de X correspondent, on a les attributs de Y qui correspondent aussi.

En d'autres mots, la contrainte $X \rightarrow Y$ signifie que pour une valeur spécifique de X, il n'y a au plus une valeur possible pour Y. Si la DF est respectée sur la relation R, nous pouvons dire que R satisfait la DF. Prenons quelques exemple sur la table 2.2 :

1. *Un NISS identifie une personne* : En d'autre mot, pour une valeur spécifique du NISS, il n'y a qu'une seule valeur possible pour tout les autres attributs de la table. Cela peut se décrire par la DF suivante : $NISS \rightarrow Nom, Prenom, Nai\_Date, Dec\_Date, Etat\_Civil, Ville, Code\_Post, Salaire, Taxe, Enfant$
2. *Deux personnes avec le même code postal vivent dans la même ville.* : Pour une valeur spécifique de $Code\_Post$ dans notre table il n'y a qu'une valeur possible de $Ville$. Par exemple si la valeur de $Code\_Post$ d'une personne est '7822', la seule valeur possible pour l'attribut $Ville$ est 'Ath'. La dépendance fonctionnelle dans ce cas est $Code\_Post \rightarrow Ville$.

Définition 2. Si pour chaque paire de tuples de la relation R, la DF τ est respectée, nous disons que la relation R *satisfait* τ. Cela ce note $R \models \tau$.

Évidement, certaines bases de données ne contiennent pas qu'une seule contrainte mais plusieurs. Il est important qu'elles soient toutes respectées. Ce qui nous conduit à la définition suivante :

Définition 3. Soit un ensemble Σ de DF sur la relation R. On dit que la relation R satisfait Σ noté $R \models \Sigma$ si pour chaque DF $\tau \in \Sigma$, on a $R \models \tau$

Malheureusement, les dépendances fonctionnelles sont limitées en terme de puissance. En effet, il existe de nombreuses contraintes que nous ne pouvons pas exprimer avec une DF. Par exemple, si nous souhaitons exprimer le fait que *'La date de naissance d'une personne*

doit être antérieure à sa data de décès', nous avons besoin de comparer la Nai_Date et la Dec_Date de la personne et de s'assurer que la date de décès ne soit antérieure à la date de naissance. Les dépendances fonctionnelles ne permettent pas d'utiliser des opérateurs de comparaison, il est donc nécessaire d'exprimer les contraintes d'une autre façon. Pour ce faire, nous allons introduire un nouveau type de contrainte qui répondra bien à nos besoins : les *denial constraints*.

2.3 Les Denials constraints

Dans cette section, nous allons définir ce qu'est une denial constraint. Nous allons aussi expliquer son utilisation dans les bases de données et nous allons également lister et expliquer plusieurs propriétés que peuvent avoir ces contraintes. Commençons d'abord par définir la denial constraint

Définition 4. Considérons un ensemble S fini d'*attribut*. Une *denial constraint (DC)* sur l'ensemble S est une fonction partielle de S vers l'ensemble des parties(aussi appelé ensemble puissance) de $\{<, =, >\}$ noté OP. Nous utiliserons la lettre grecque φ pour représenter une DC. Chaque parties de OP se nomme *opérateur*.

Définition 5. Soit $(\mathbf{dom}, \leq)$ un domaine totalement ordonné contenant au moins deux éléments distincts. Un *tuple sur* S est une fonction totale de S à $\mathbf{dom}$. Une *relation sur S* est un ensemble fini de tuples sur S.

Par définition l'ensemble puissance d'un ensemble S noté $\mathcal{P}(S)$ est l'ensemble de tous les sous-ensembles de S. Cela inclut l'ensemble S lui même mais aussi l'ensemble vide $\emptyset$. Par exemple OP (l'ensemble des parties de $\{<, =, >\}$) est $\mathcal{P}(OP)$ = $\{\emptyset, \{<\}, \{=\}, \{>\}, \{<, =\}, \{=, >\}, \{<, >\}, \{<, =, >\}\}$. Nous utiliserons la lettre grecque ϕ ou θ pour représenter un opérateur. Il existe différentes abréviations pour les éléments de OP, ceux-ci étant répertoriés dans la table 2.3. Nous avons eu besoin d'introduire 2 nouveaux opérateur $\top$ et $\bot$, chacun étant l'abréviation pour l'ensemble $\{<, =, >\}$ et $\emptyset$ respectivement. Nous les définissons comme tel :

Définition 6. $\forall a, b \in \mathbf{dom}$,nous avons $a \bot b$ est toujours faux et $a \top b$ est toujours vrai.

Expliquons maintenant la sémantique qui se cache derrière la denial constraint.

Définition 7. On dit qu'une relation I sur S *satisfait* la DC φ, noté $I \models \varphi$ si il **n'existe pas** deux tuples $s, t \in I$ tel que pour chaque attribut A dans le domaine de φ, nous avons $s(A)\theta t(A)$ avec $\theta = \varphi(A)$

Notons que les tuples s et t ne doivent pas forcément être distinct. La relation I peut être vide. Une relation vide satisfait toutes les DCs.

θ	Abréviation	$\bar{\theta}$	$\hat{\theta}$	implication
$\emptyset$	$\bot$	$\top$	$\bot$	$\{\bot\}$
$\{<\}$	$<$	$\geq$	$>$	$\{<, \leq, \neq, \top\}$
$\{=\}$	$=$	$\neq$	$=$	$\{=, \leq, \geq, \top\}$
$\{>\}$	$>$	$\leq$	$<$	$\{>, \geq, \neq, \top\}$
$\{<, =\}$	$\leq$	$>$	$\geq$	$\{\leq, \top\}$
$\{<, >\}$	$\neq$	$=$	$\neq$	$\{\neq, \top\}$
$\{>, =\}$	$\geq$	$<$	$\leq$	$\{\geq, \top\}$
$\{<, =, >\}$	$\top$	$\bot$	$\top$	$\{\top\}$

TABLE 2.3 – Element de OP, l'ensemble des parties de $\{<, =, >\}$

Exemple 1. Prenons un exemple sur la table 2.1, nous avons $S = \{Nom, Anniversaire, NumTel, Annee, Revenu, Taxe\}$ Une DC pour S est $\varphi = \{(Nom, =), (Anniversaire, =), (NumTel, \neq), (Annee, \top), (Revenu, \top), (Taxe, \top)\}$. Celle-ci est satisfaite par la relation I si il n'existe pas deux tuples $s, t \in I$ tel que $s(Nom) = t(Nom) \wedge s(Anniversaire) = t(Anniversaire) \wedge s(NumTel) \neq t(NumTel) \wedge s(Annee) \top y(Annee) \wedge s(Revenu) \top t(Revenu) \wedge s(Taxe) \top t(Taxe)$.

Lorsqu'une relation I sur S ne satisfait pas une DC φ, on dit que I *viole* φ que l'on note $I \not\models \varphi$.

Soit φ une DC sur S . Nous appellerons *prédicat* P de φ l'expression de la forme (A, θ) avec $A \in S$ que l'on appelle attribut du prédicat et $\theta = \varphi(A)$ que l'on appelle opérateur du prédicat. Soit $pred(\varphi)$ l'ensemble des prédicats de la DC φ. Soit I une relation sur S. Dès lors on peut dire que φ est satisfaite si au moins un des prédicats est faux. Si un prédicat P a pour opérateur $\top$ alors P sera toujours vrai pour tout $t, s \in I$. Dès lors à l'avenir, nous ne noterons plus les prédicats ayant $\top$ pour opérateur par facilité syntaxique. L'exemple précédent s'écrira désormais $\varphi = \{(Nom, =), (Anniversaire, =)?(NumTel, \neq)$. Si un prédicat a pour opérateur $\bot$, il sera toujours faux. Dès lors $I \not\models \varphi$. Notons que la DC $\varphi = \{(A_1, \top), (A_2, \top), ...(A_n, \top)\} \equiv \{\}$ n'est satisfaite par aucune relation excepté par une relation vide.

Si nous prenons l'instance I comme étant la table 2.1, nous avons $I \not\models \varphi$. En effet prenons $s = t_2$ et $t = t_3$ nous avons bien $t_2(Nom) = t_3(Nom) \wedge t_2(Anniversaire) = t_2(Anniversaire) \wedge y_2(NumTel) \neq t_3(NumTel)$. On dit que $\langle t_2, t_3 \rangle$ *viole* la contrainte φ

Pour chaque opérateur dans OP nous pouvons définir son inverse, sa réciproque et son implication. Les valeurs de l'inverse, la réciproque et l'implication de chaque élément de OP se trouve également à la table 2.3.

Définition 8. Soit ϕ un élément de OP
L'inverse de ϕ noté $\overline{\phi}$ est égal à $\{<, =, >\} \setminus \phi$

La réciproque de ϕ noté $\hat{\phi}$ s'obtient en inter-changeant < et > dans ϕ

L'implication de ϕ noté $Imp(\phi)$ est un ensemble d'élément de OP tel que pour n'importe quelle valeur a et b, si $a\phi_2 b$ **implique**[2] **toujours** $a\phi_1 b$ alors $\phi_2 \in Imp(\phi_1)$.

Notons que $\forall \phi_1, \phi_2$, si $\phi_2 \in Imp(\phi_1)$ alors ϕ est un sous ensemble de ϕ_2. Par exemple $\neq \in Imp(>)$ et $\{>\} \in \{<, >\}$.

Une DC peut être *sur-simplifiée* ce qui veut dire qu'une donnée correcte peut être considérée comme une violation. Prenons un exemple sur la table 2.1 avec la denial constraint suivante :

$$\varphi_2 = (Nom, =)(NumTel, \neq)$$

Cette contrainte veut dire que si une personne possède le même nom qu'une autre, alors elle ne peut pas avoir un numéro de téléphone différent. Ceci est bien sur incorrect. En effet, deux personnes différentes ne peuvent avoir le même numero de téléphone. Dans notre relation, le nom seul ne suffit pas à identifier si deux personne sont identiques. Prenons par exemples t_1 et t_2, ils ne satisfont pas φ_2. Si l'on regarde de plus près, on peut facilement comprendre qu'il s'agit de deux personnes différentes. Ces deux personnes n'ont pas le même age i.e elles ont une date d'$Anniversaire$ différent. Si nous souhaitons améliorer la précision de la contrainte et éviter que $\langle t_1, t_2 \rangle$ soit considéré comme une violation, nous avons besoin de regarder l'attribut $Anniversaire$. Une meilleure DC serait :

$$\varphi_2' = (Nom, =), (Anniversaire, =), (NumTel, \neq)$$

Une DC peut être également *sur-raffiné* ce qui entraine qu'une donnée erronée peut être considérée comme correcte par la DC. Prenons un exemple sur la table 2.1 avec la denial constraint suivante :

$$\varphi_2' = (Nom, =), (Anniversaire, =), (NumTel, \neq), (Annee, =)$$

Dans ce cas, l'information $Annee$ n'est pas utile pour distinguer deux personnes différentes. Dans la table, l'attribut année correspond à l'année où les autres attributs ont été encodés. Une même personne peut être encodée deux fois à deux années différentes. Avec cette DC, on ne reconnait pas $\{t_5, t_8\}$ comme étant une violation.

2.3.1 Quelques définitions et propriétés

Dans cette sous-section, nous allons définir quelques notions et propriétés sur les DC qui nous serviront dans les chapitres qui suivront.

2. Tout tuple qui satisfait $a\phi_2 b$ satisfait $a\phi_1 b$

2.3.1.1 Satisfiabilité

Définition 9. Soit φ DC sur S. On dit que φ est *satisfiable* si elle peut être satisfaite par une relation non vide sur S, i.e si $\exists I$ sur S avec I non vide tel que $I \models \varphi$, alors φ est *satisfiable*.

Si φ n'est pas satisfiable, nous dirons qu'il est *insatisfiable*

Il est intéressant de savoir à l'avance si une une denial constraint est satisfiable ou pas. Le lemme suivant nous permet de détecter les DC qui ne sont pas satisfiables.

Lemme 1. Soit φ une denial constraint sur S, alors φ est satisfiable si et seulement si il existe un prédicat $P_i \in pred(\varphi)$ de la forme (A_i, θ_i) tel que θ_i ne contient pas $=$, i.e $\theta_i \notin \{\{=\}, \{<, =\}, \{=, >\}, \{<, =, >\}, \}$

Démonstration.

$\boxed{\Longrightarrow}$ Supposons que pour tout $P \in pred(\varphi)$, θ contient $=$. Alors pour chaque tuple s sur S, pour chaque $P_i \in pred(\varphi)$ on a $s(A_i)\theta_i s(A_i)$. Il s'ensuit que toute relation non vide ne satisfait pas φ

$\boxed{\Longleftarrow}$ Supposons $B \in S$ tel que B ne contient pas $=$. Alors pour chaque tuple s sur S, nous avons que $s(B)\ \theta s(B)$ avec $\theta = \varphi(B)$ faux. Il s'ensuit que n'importe quelle relation avec exactement un tuple satisfait φ.

□

2.3.1.2 Implication logique

Définition 10. Soit φ_1, φ_2 deux DC sur S. On dit que φ_1 *implique (logiquement)* φ_2, que l'on note $\varphi_1 \models \varphi_2$, si pour chaque relation I sur S, si $I \models \varphi_2$ alors on a $I \models \varphi_1$. On dira aussi que φ_2 est *plus faible* que φ_1 ou bien que φ_1 est *plus fort* que φ_2

Exemple 2. Soit $S = \{A, B\}$. Soit $\varphi_1 = \{(A, \leq), (B, \neq)\}$ et $\varphi_2 = \{(A, <), (B, >)\}$. Alors φ_1 implique φ_2. En effet, soit I une relation qui satisfait φ_1. Alors pour tout tuples $s, t \in I$, on a $s(A) > t(A)$ ou bien $s(B) = t(B)$ (ou éventuellement les deux en même temps). Il s'ensuit que pour tout tuples $s, t \in I$, on a $s(A) \geq t(A)$ ou bien $s(B) \leq t(B)$ (ou les deux en même temps). Dès lors, I ne contient pas deux tuples s, t tel que $s(A) < t(A)$ et $s(B) > t(B)$. On a donc I qui satisfait φ_2. D'un autre côté, φ_2 n'implique pas φ_1. En effet, considérons la relation I suivante.

I	A	B
	1	2
	1	3

Dès lors, nous avons $I \models \varphi_2$, mais $I \not\models \varphi_1$.

Lemme 2. Soit φ_1 et φ_2 deux denial constraints sur S. Si $\varphi_2(A) \subseteq \varphi_1(A)\ \forall A \in S$, alors $\varphi_1 \models \varphi_2$.

Démonstration. Supposons que $\varphi_2(A) \subseteq \varphi_1(A)$ pour tout $A \in S$. Soit I une relation sur S tel que $I \models \varphi_1$. Nous avons besoin de démontrer que $I \models \varphi_2$. Soit $s, t \in I$. Puisque $I \models \varphi_1$, nous pouvons supposer l'existence d'un $A \in S$ tel que $s(A)\ \theta\ t(A)$ est faux, avec $\theta = \varphi(A)$. Puisque $\varphi_2(A) \subseteq \varphi_1(A)$, alors nous aurons $s(A)\ \theta'\ t(A)$ est faux, avec $\theta' = \varphi_2(A)$. □

Exemple 3. $\{(A, =)\} \models \{(A, \leq)\}$. car $\{=\} \subseteq \{<, =\} \equiv \leq$

2.3.1.3 Trivialité

Une DC peut être inutile et toujours vraie. De telles DC ne devraient pas être présentes dans la base de données puisqu'elles ne détecteront jamais aucune violation. Dans ce cas, on dira que la DC est *triviale*.

Définition 11. Une DC φ est dite *triviale* si $\forall I$ sur S, on a $I \models \varphi$

Lemme 3. Soit φ une DC sur S. alors φ est triviale si et seulement si $\theta = \perp$ pour un prédicat de la forme $P = (A, \theta) \in pred(\varphi)$.

Démonstration.
$\boxed{\Longrightarrow}$ Supposons que pour chaque prédicat $P = (A, \theta) \in pred(\varphi)$ avec $\theta \neq \perp$. Soit s, t deux tuples tel que pour chaque $P = (B, \theta)$, nous avons $s(B)\ \theta\ t(B)$ vrai. Puisque $\theta \neq \perp$ et que **dom** contient au moins deux éléments, s, t peuvent être construit. Dès lors $\{s, t\}$ ne satisfont pas φ puisque φ n'est pas triviale.
$\boxed{\Longleftarrow}$ Supposons qu'il existe un prédicat $P = (A, \theta) \in pred(\varphi)$ tel que $\theta = \perp$. Puisque $s(A) \perp t(A)$ est faux pour tout tuples s, t sur S, aucune relation ne peut contenir deux tuples s, t tel que $s(A) \perp t(A)$ est vrai. □

2.3.1.4 Augmentation

Dans les chapitres suivants, nous verrons comment nous pouvons améliorer les DCs afin qu'elles puissent détecter les données erronées de manière correcte. Pour ce faire nous aurons besoin de modifier des prédicats et donc de changer les opérateurs de ceux ci. Mais modifier un opérateur $\top$ est inutile par la propriété suivante :

Propriété 1. Soit une DC φ sur S et une relation I tel que $I \models \varphi$. Si il existe un prédicat $P_i = (A_i, \theta_i) \in pred(\varphi)$ tel que $\theta_i = \top$ alors pour la DC φ' tel que $pred(\varphi') = pred(\varphi) \backslash P_i$ avec $P'_i = (A_i, \theta_i)$ et $\theta_i \neq \top$, on a $I \models \varphi'$

Cette propriété est triviale. Souvenons nous que φ est un DC tel que $I \models \varphi$ donc $\forall t \in I$ on a φ vrai. Imaginons que φ contienne le prédicat de la forme $(A, \top)$ Prenons φ' une DC qui est la variante de φ tel que $pred(\varphi) = pred(\varphi')$ à l'exception du prédicat $(A, \top)$ qui devient (A, θ) avec $\theta \neq \top$. Puisque φ était satisfaite par I, il y avait déjà un autre prédicat de φ qui était faut $\forall t \in I$. Donc φ' est satisfaite pour qu'importe la valeur de (A, θ).

2.3.1.5 Transitivity

Propriété 2. Soit φ et φ' deux DC sur S et I une relation sur S. Si $\varphi = \{P_1, P_2, ..., P_{i-1}, P_i\}$ est satisfaite par I et $\varphi' = \{P'_i, P_{i+1}, ..., P_n\}$ satisfaite par I, avec P_j prédicat de la forme $(A_j, \theta_j) \forall j \in \{1, ..., n\}$ et $\theta'_i \in Imp(\overline{\theta_i})$, alors $\varphi'' = \{P_1, P_2, ..., P_{i-1}, P_{i+1}, ..., P_n\}$ est également satisfaite par I

En d'autre mots, il est possible de fusionner deux denial constraint satisfaites par I si ces deux contraintes possèdent chacune un prédicat sur un même attribut et si ces prédicats ne peuvent pas être faux en même temps. Alors, la fusion de ces deux contraintes sans les deux prédicats est toujours satisfaite.

2.3.1.6 Raffinement

Dans l'article [4] ils définissent le raffinement d'une denial constraint comme étant :

Définition 12. φ_2 est un **raffinement** de φ_1, noté $\varphi_1 \preceq \varphi_2$, Si pour chaque prédicat $(A, \theta_A) \in pred(\varphi_1)$ on a un prédicat $(A, \theta'_A) \in pred(\varphi_2)$ tel que θ'_A implique θ_A ($\theta'_A \in Imp(\theta_A)$)

Exemple 4. Soit $\varphi_1 = \{(Taxe, \leq), (Revenu, <)\}$ et $\varphi_2 = \{(Taxe, <), (Revenu, <), (Anne, =)\}$ nous avons $\varphi_2 \preceq \varphi_1$ car $\varphi_1(Taxe) \in Imp(\varphi_2(Taxe))$ et $\varphi_1(Revenu) \in Imp(\varphi_2(Revenu))$ et $\varphi_1(Anne) \in Imp(\varphi_2(Anne))$

Notons que φ est raffinement de lui même et que remplacer l'opérateur $\top$ d'un prédicat par n'importe quel autre

Définition 13. Σ_2 est **raffinement** de Σ_1, noté $\Sigma_1 \preceq \Sigma_2$, si pour chaque $\varphi_2 \in \Sigma_2$, Il existe un $\varphi_2 \in \Sigma_1$ tel que $\varphi_1 \preceq \varphi_2$

Si nous voulons changer moins de données, nous pouvons raffiner nos DC. Dans notre exemple précédent, la nouvelle denial constraint est plus faible que la précédente, diminuant le nombre de tuples détecté comme étant une violation.

Chapitre 3

Data Repairing

Les erreurs sont fréquentes dans les bases de données et ces anomalies nuisent à la fiabilité de certaines applications qui les utilisent. Il existe des méthodes dont le but est de détecter ces erreurs mais ces méthodes ne les réparent pas. A la place, les applications pourront filtrer les données et ignorer les erreurs détectées mais malgré ce filtrage, les applications peuvent toujours être non fiables [5]. Au lieu de simplement détecter les erreurs et les filtrer, il est préférable de réparer les données erronées.

Dans le chapitre précédent, nous avons vu comment détecter des erreurs au moyen de denial constraints. Nous avons aussi discuté brièvement de la sur-simplification ou du sur-raffinement de ces DC. Nous allons maintenant aborder la réparation des données mais aussi la réparation des DC.

Le but d'une réparation de données est de trouver une nouvelle relation I' qui est une modification d'une relation I de S. Dans I' il n'y a pas de données erronées. Cela nous amène à définir ce qu'est une base de données sans erreurs. Une base de données sans erreurs est une base de données dont toutes les contraintes sont satisfaites, c'est à dire :

Définition 14. Soit Σ un ensemble de DC sur S. Soit une relation I sur S. On dit que I *satisfait* Σ noté $I \models \Sigma$ si pour chaque DC φ avec $\varphi \in \Sigma$, nous avons $I \models \varphi$

Définition 15. Soit Σ un ensemble de DC sur S. Soit une relation I sur S. Une *réparation de* I est une fonction f de domaine I qui attribue à chaque tuple t de I un nouveau tuple $f(t)$ tel que $f(t) \models \Sigma$. L'ensemble d'arrivée de f est noté $I' = f(I)$. On a donc $I' \models \Sigma$

Donc, lorsque l'on parle de réparation de données, on cherche à trouver I' tel que $I' \models \Sigma$ c'est à dire une nouvelle instance où toutes les violations dans le set de contrainte Σ sont éliminées. Nous allons considérer que lors d'une réparation, nous ne supprimons pas de tuples mais que nous le modifions. En effet, chaque ligne de la table est susceptible de contenir des précieuses informations. Chaque ligne supprimée diminuerait la quantité

d'informations utiles de la base de données. Dans de nombreux cas, seules quelques colonnes de la table contiennent quelques erreurs et les autres colonnes ne présentent aucun problème. Nous n'ajoutons pas de tuples à la relation. Lorsqu'un tuple t est réparé celà peut conduire à avoir deux tuples identiques.

Exemple 5. Soit la relation I suivante :

I

A	B
a	b
a	c

I'

A	B
a	b
a	b

et la DC $\varphi = \{(A, =), (B, \neq\}$. Nous avons $I' \models \varphi$ et $I \not\models \varphi$. I' est une réparation de I, le tuple (a, c) a été modifié en (a, b)

Notons ici la présence de deux lignes identiques dans l'instance I'. Ce n'est pas une pratique courante dans les bases de données, nous sommes donc dans ce cas ci autorisé à supprimer un des lignes ce qui nous donne :

I'

A	B
a	b

La réparation de données suit le principe du changement minimum : La nouvelle relation I' doit minimiser le coût de réparation de données défini comme étant :

Définition 16. Soit I une relation sur S et $f(I)$=I' une réparation de I, alors le coût de réparation est le suivant :

$$\Delta(I, I') = \sum_{t \in I, A \in S} w(t.A).dist(t.A, f(t).A)$$

où :

— $dist(t.A, f(t).A)$ est la distance entre la valeur $t.A$ et sa réparation $f(t).A$.

— $w(t.A)$ est un poid sur la valeur $t.A$.

Dans le coût, nous avons un poid $w(t.A)$ pour un attribut $A \in S$ et un tuple $t \in I$. Ce poid correspond à la confiance que l'on accorde en la valeur de t pour l'attribut A. On peut grâce à cette valeur influencer la réparation de données pour privilégier la réparation d'une valeur plutôt qu'une autre. Pour pouvoir assigner un poids $w(t.A)$, il faut avoir une bonne connaissance du contexte de la base de données d'origine. Il est courant d'avoir la même valeur pour $w(t.A)$ et $w(s.A)$ avec $s, t \in I$ car en général on a connaissance de la confiance pour un attribut en particulier pas pour chaque valeur du tuple. Par exemple, pour la table 2.2, nous pouvons supposer qu'une valeur pour l'attribut *Enfant* est plus susceptible d'être précise que la valeur pour l'attribut *Salaire* ou *Taxe*. Lorsqu'on manque de connaissances sur la base de données, on fixe le même poids à chaque attribut.

Le coût de réparation peut être le nombre de valeurs de I que l'on a changées si nous décidons que :

$$dist(t.A, f(t).A) = \begin{cases} 1\ Si\ t.A \neq f(t).A\ (La\ valeur\ a\ chang\acute{e}) \\ 0\ Sinon\ (aucun\ changement\ n'est\ fait) \end{cases}$$

Nous pouvons aussi décider que la distance est égale à la différence entre les deux valeurs dans le cas d'un attribut numérique. Pour un attribut de type chaine de caractère, nous pouvons utiliser la distance d'édition [1].

Pour réparer une donnée, nous devons remplacer sa valeur par une autre plus adéquate. Mais quel valeur choisir ? Dans l'article de référence principal [4], si la valeur $t.A$ est erronée, ils essayent de trouver une valeur pour $f(t).A$ qui soit dans le $\mathbf{dom}(A)$ et qui respecte les contraintes. Si ce n'est pas possible ils attribuent une *variable fraiche* fv à $t.A$.

Définition 17. Une *variable fraiche* fv est une valeur qui ne satisfait aucun prédicat c'est à dire : soit φ une DC sur S et I une relation sur S. Une variable fraiche fv pour $A \in S$ est une valeur tel que $fv \notin \mathbf{dom}(A)$ le prédicat $(A, \theta_A) \in pred(\varphi)$ est toujours faux quelque soit l'opérateur θ_A donc pour tout tuple t avec $t \in I$, si $t.A = fv$ alors $t \models \varphi$.

Nous allons procéder de manière un peu différente par rapport à l'article de référence [4]. Nous allons automatiquement attribuer une nouvelle valeur fv pour chaque donnée erronée. Par la suite, nous allons essayer de trouver une valeur adéquate pour remplacer la valeur fraiche fv si possible. Une des motivations derrière ce changement vient du fait qu'après une réparation, certaines valeurs du domaine n'y figurent plus (ces valeurs étaient erronées). De plus il n'est pas pertinent de croire que si une valeur du domaine peut réparer une valeur erronée, alors elle est forcément la bonne valeur. Une valeur en dehors du domaine peut tout aussi bien convenir.

Si nous attribuons la variable fraiche $fv_{t.A}$ pour l'attribut A d'un tuple t et la valeur fraiche $fv_{s.A}$ pour l'attribut A d'un tuple s, ces deux variables peuvent être différentes ou identiques et ce pour tout $s, t \in I$. Par défaut, nous considérons dans un premier temps que toutes les variables fraiches sont différentes.

Exemple 6. Prenons un exemple pour relation I la table 2.1. Supposons que notre denial constraint est la suivante :

$$\varphi = \{(Revenu, >), (Taxe, \leq)\}$$

1. Le nombre minimum d'opération nécessaire pour transformer la chaine de caractère initiale en la chaine de caractère cible

/	t_β									
t_α	/	1	2	3	4	5	6	7	8	9	10
	1	/	V	V	V	V	V	V	V	V	V
	2	F	/	V	V	V	V	V	V	V	V
	3	F	V	/	V	V	V	V	V	V	V
	4	V	V	V	/	V	V	V	V	V	V
	5	F	F	F	F	/	V	V	V	V	V
	6	F	F	F	F	V	/	V	V	V	V
	7	F	F	F	F	F	F	/	V	V	V
	8	V	V	V	V	V	V	V	/	V	V
	9	V	V	V	V	V	V	V	V	/	V
	10	V	V	V	V	V	V	V	V	V	V

FIGURE 3.1 – Toutes les violations pour φ

En d'autres termes, on suppose par cette contrainte que si une première personne perçoit un revenu annuel plus élevé qu'une seconde personne, alors la première personne doit payer un montant de taxe annuelle plus élevé. Nous avons $(t_2, t_1) \not\models \varphi$ parce que $t_2.Revenu > t_1.Revenu$ et $t_2.Taxe \leq t_1.Taxe$. Nous avons également $(t_3, t_1) \not\models \varphi, (t_5, t_1) \not\models \varphi, ...$ Toutes les violations de φ peuvent être trouvé à la figure 3.1. Une réparation I' de I pourrait être la table 3.1.

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_18	t_9	t_10
Taxe	0	fv_1	fv_2	3k	fv_3	fv_4	fv_5	21k	21k	40k

TABLE 3.1 – Exemple de réparation I' pour l'attribut Tax for φ.

Nous avons choisi fv_1 comme variable fraiche pour t_2. Bien que ce soit une variable fraiche, nous savons plusieurs choses à son propos. En effet nous savons les choses suivantes :

1. $I(t_1.Taxe) = 0$ donc $I'(t_2.Taxe) > 0$, i.e $fv_1 > 0$ car $I(t_1.Revenu) < I(t_2.Revenu)$
2. $I(t_4.Taxe) = 3$ donc $I'(t_2.Taxe) < 3k$, i.e $fv_1 < 3k$ car $I(t_2.Revenu) < I(t_4.Revenu)$

Nous avons assigné une variable fraiche fv_1 comme valeur pour $t_2.Taxe$ mais nous savons quand même que $0 < fv_1 < 3k$ i.e $fv_1 \in]0, 3k[$ grâce à 1 et 2. Nous pouvons utiliser la même logique pour connaitre les valeurs possible pour les autres variables fraiches. Nous pouvons dire que $fv_2 = fv_1$, $3k < fv_3 < 21k$, $fv_4 = fv_3$ et $fv_3 < fv_5 < 21k$. Maintenant que nous avons identifié les valeurs erronées ainsi qu'un ensemble de valeurs pour remplacer les variables fraiches, il reste à savoir quelle valeur finale on peut prendre. Pour cela nous avons plusieurs solutions possibles.

— Prendre une valeur de $\mathbf{dom}(A)$. C'est la solution envisagée dans le principal article de référence que nous utilisons. Il ne sera pas toujours possible de prendre une valeur dans le domaine, par exemple il n'y a pas de valeur dans le $dom(Taxe) = \{0, 3k, 21k, 40k\}$ qui puisse satisfaire la condition fv_1. Dès lors, dans le cas où aucune valeur de dom(A) n'est attribuable à la variable fraiche, celle-ci est conservée dans la base de données. Cette pratique n'est pas la plus logique puisque même si nous avions une valeur dans $\mathbf{dom}(Taxe)$ qui puisse satisfaire fv_1, il y a plusieurs valeurs en dehors de $\mathbf{dom}(Taxe)$ qui sont tout aussi correctes.

— Prendre une valeur aléatoire mais respectant les conditions sur la variable fraiche. C'est une très mauvaise idée puisque nous avons une chance de s'éloigner de la vraie valeur. Cela peut impacter énormément l'ajout de tuples dans la relation après que la réparation soit effectuée. Ces nouveaux tuples malgré qu'ils soient corrects pourraient être perçus comme contenant des données erronées.

— Garder les variables fraiches dans la base de données tout en conservant les informations que l'on connait à propos de celles-ci. Si nous n'avons qu'une seule valeur possible alors nous privilégions cette valeur à fv. Cette solution est celle qui respectera le mieux l'intégrité des données. Le seul problème qu'apporte cette solution est que de nombreuses applications ne pourront plus fonctionner correctement avec des variables fraiches. De nombreux SGBD ne permettent pas de stocker ces variables. Les informations que l'on connait sur elles peuvent changer au fur et à mesure que la relation se remplit.

Exemple 7. Nous pouvons calculer le coût de réparation pour la relation de la table 2.1 en considérant les distances suivantes :

$$\forall a \in dom(A)\ avec\ a \neq b. \begin{cases} dist(a,a) = 0 \\ dist(a,b) = 1 \\ dist(a,fv) = 1.5 \\ dist(fv,fv) = 1.5 \\ dist(fv,b) = 1 \end{cases}$$

Lorsqu'on ne change pas la valeur, la distance est bien évidement égale à zéro. $dist(a, fv)$ soit être supérieure à $dist(a, b)$ pour privilégier les valeurs aux variables fraiches. $dist(fv, fv)$ représente le fait qu'on avait déjà une variable fraiche venant d'une réparation antérieure, et qu'on garde une variable fraiche. $dist(fv, b)$ représente le changement d'une variable fraiche venant d'une réparation antérieure par une valeur précise. Cela peut arriver lorsque l'on possède de plus amples informations sur la valeur fraiche,par exemple grâce à de nouveaux tuples dans la relation. Nous avons besoin que $dist(fv, fv) > dist(fv, b)$ pour favoriser la correction par une vraie valeur quand c'est possible. Dans l'exemple précédent, avec les valeurs susmentionnées, nous pouvons calculer un coût de réparation $\Delta(I, I')$ =$5 * dist(a, a) + 5 * dist(a, fv)$ = $7, 5$.

3.1 Variation sur les denial constraints

Nous avons vu précédemment qu'une denial constraint peut être sur-raffinée, échouant donc dans la détection d'erreurs. Une DC peut aussi être sur-simplifiée ce qui conduit à considérer des données correctes comme étant erronées. Parce que les contraintes peuvent être imprécises et inexactes, nous avons besoin de les corriger. En modifiant ces contraintes, nous pouvons obtenir une meilleure réparation

Exemple 8. Par exemple prenons la DC suivante :

$$\varphi = \{(Revenu, >), (Taxe, \leq)\}$$

La denial constraint φ exprime le fait que si je reçois un revenu plus élevé qu'une autre personne, alors je dois payer un montant de taxe strictement plus élevé. Nous allons modifier φ pour obtenir une nouvelle denial constraint φ' en changeant le prédicat $(Taxe, \leq)$ en $(Taxe, <)$. Dès lors, maintenant φ' exprime le fait que si je reçois un revenu plus élevé qu'une autre personne, alors je dois payer un montant de taxe plus élevé ou équivalent. Cela permet d'exempter de taxe plusieurs personnes ayant un Revenu faible, mais différent les uns des autres.

$$\varphi' = \{(Revenu, >), (Taxe, <)\}$$

Avec cette nouvelle contrainte, nous avons moins de violations détectées. Toutes les violations peuvent être trouvées à la figure 3.2. Les modifications que nous avons faites dans la table 3.1 sont :

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_18	t_9	t_10
Taxe	0	0	0	fv_1	0	0	0	21k	21k	40k

TABLE 3.2 – Example of repair with Tax

.

Contrairement à la réparation de φ qui proposait une correction avec 5 variables fraiches (voir table 3.1), nous 'avons ici qu'une seule variable fraiche. De plus, nous savons certaines choses à propos de cette variable fraiche :

1. $t_1.Taxe = 0$ donc $f(t_4).Taxe \geq 0$ car $t_1.Revenu < t_4.Revenu$
2. $t_5.Taxe = 0$ donc $f(t_4).Taxe \leq 0$ car $t_4.Revenu < It_5.Revenu$

Nous avons donc $0 \leq fv_1 \leq 0$, dès lors la seule valeur possible pour remplacer fv_1 est $fv_1 = 0$. Nous avons donc un coût de réparation de $\Delta(I, I') = 1$ en reprenant les valeurs de distance de l'exemple 7.

/	t_β									
t_α	/	1	2	3	4	5	6	7	8	9	10
	1	/	V	V	V	V	V	V	V	V	V
	2	V	/	V	V	V	V	V	V	V	V
	3	V	V	/	V	V	V	V	V	V	V
	4	V	V	V	/	V	V	V	V	V	V
	5	V	V	V	F	/	V	V	V	V	V
	6	V	V	V	F	V	/	V	V	V	V
	7	V	V	V	F	V	V	/	V	V	V
	8	V	V	V	V	V	V	V	/	V	V
	9	V	V	V	V	V	V	V	V	/	V
	10	V	V	V	V	V	V	V	V	V	V

FIGURE 3.2 – All the violation for φ'

Nous voyons donc que la modification de la contrainte conduit à un coût de réparation plus faible. Mais modifier une contrainte doit aussi avoir un coût. En effet , si on ne donne pas de coût à la modification d'une contrainte, il suffirait de sur-raffiné toutes nos contraintes pour détecter moins de violations et donc faire moins de modifications. Nous allons aussi considérer que nous n'ajoutons pas DC ni n'en retirons de Σ pour obtenir Σ' et donc $|\Sigma| = |\Sigma'|$.

Nous devons désormais définir une marche à suivre concernant les variations de DC. Nous devons trouver un moyen de définir un coût de variation. Considérons le treillis tel qu'illustré à la figure 3.3 pour un attribut A de S. Chaque nœud du treillis représente un prédicat différent pour A, donc A associé à un opérateur, un des éléments de l'ensemble puissance de OP. Descendre dans une branche du treillis consiste à ajouter un élément de OP à l'opérateur. Par exemple passer de $(A, <)$ à $(A, \leq)$ consiste à ajouter l'opérateur $=$ à $\{<\}$. Lorsque l'on monte dans le treillis, on retire un élément de OP à l'opérateur. Par exemple, passer de $(A, \geq)$ à $(A, =)$ consiste à retirer l'opérateur $>$ à $\{>, =\}$.

Nous devons modifier nos DC de telle manière qu'une contrainte sur-raffinée ou sur-simplifiée ne le soit plus. Si une DC est sur-raffinée, nous devons descendre dans le treillis. En effet , plus on descend dans le treillis, plus le prédicat sur l'attribut A sera faible, i.e la DC sera faible. A l'inverse, si une DC est sur-simplifiée ; nous devons monter dans le treillis. En effet plus nous montons dans le treillis, plus le prédicat sur l'attribut A sera fort, i.e la DC sera plus forte. Plus nous changeons de prédicats, plus le *coût de variation des contraintes* sera important. Le *coût de variation des contraintes* se définit comme tel :

Définition 18. Soit Σ un ensemble de DC sur un ensemble S. Pour une variante Σ' de Σ, la fonction de calcul du *coût de variations des contraintes* de Σ est :

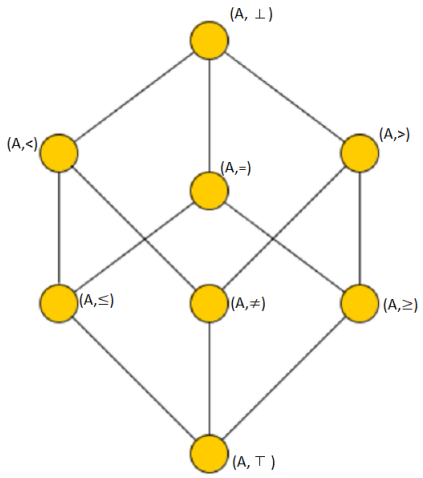

FIGURE 3.3 – Treillis pour un prédicat d'attribut A

$$\Theta(\Sigma, \Sigma') = \sum_{\varphi in \Sigma} edit(\varphi, \varphi')$$

avec φ' une variante de φ et $edit(\varphi, \varphi')$ est le coût pour changer φ en φ'.

La fonction $edit(\varphi, \varphi')$ qui indique le coût pour changer φ en φ' est défini comme étant :

Définition 19.

$$edit(\varphi, \varphi') = \sum_{A \in S} path(\varphi(A), \varphi'(A))$$

avec $path(\varphi(A), \varphi'(A))$ le coût du chemin emprunté dans le treillis.

Si nous considérons notre treillis comme un graphe, il reste maintenant à trouver le poids de chacun des arcs de ce graphe.

3.1.1 Parcours dans le treillis

Maintenant que nous avons défini le poids de variation des contraintes, regardons le coût pour se déplacer dans le treillis. Soit φ une DC sur S et φ' une variante de φ. Soit A un attribut sur S. Le cout pour passer de $\varphi(A)$ à φ', i.e la valeur de $path(\varphi(A), \varphi'(A)$ est le coût du plus court chemin du treillis. Le treillis doit être vu comme un graphe pouvant être parcouru dans les deux sens. Le poids de chaque arc est de $c(A)$, $c(A)$ étant une valeur symbolisant la confiance qu'on accorde à la valeur de A. Si on n'a aucune connaissance sur la base de données, $c(A) = c(B)$, pour tout $A, B \in S$.

Exemple 9. Soit $\varphi = \{(Taxe, \leq), (Revenu, <)\}$ et $\varphi' = \{(Taxe, >)(Revenu, <)\}$ et c(A) =1.
Pour passer de $(Taxe, \leq)$ à $(Taxe, >)$, un chemin possible est $(Taxe, \leq) \rightarrow (Taxe, <) \rightarrow (Taxe, \perp) \rightarrow (Taxe, >)$ et ce chemin à un coût de 3.

Par observation, on peut facilement se rendre compte que le plus court chemin du treillis est de longueur 3 maximum. Dans les sous-sections qui suivent, nous verrons qu'il sera important de limiter nos candidats, ce qui modifiera éventuellement le poids de certains arcs.

3.1.2 Limitation des candidats

Dans notre treillis, nous avons 8 nœuds. Pour une DC sur un ensemble S de taille 1, nous avons 8 différentes variantes de DC. Pour un ensemble S de taille n nous avons 8^n variations. Ce nombre est très important et envisager toutes les variations de contraintes serait coûteux en terme de complexité. Si le calcul d'une variation se fait en $O(1)$ alors considérer toutes les variantes possible est en $O(8^n)$. Si l'ensemble Σ de contraintes est de taille l, puisqu'il faut considérer toutes les variantes pour chaque DC $\varphi \in \Sigma$, la complexité serait de $O((8^n)^l)$ Nous avons donc besoin de limiter le nombre de variations à considérer. Pour ce faire, nous allons utiliser des propriétés sur les DC pour limiter le nombre de contraintes à considérer.

Nous savons déjà par le chapitre précédent que les contraintes triviales sont inutiles. En effet, une contrainte triviale sera toujours vraie et donc considérera tous les tuples comme étant corrects. Le Lemme 3 nous permet de rejeter toutes les DC dont au moins un des prédicats est de la forme $(A, \perp)$.

Nous avons également besoin que la denial constraint soit satisfiable. Si une DC n'est pas satisfiable alors elle possède uniquement des prédicats avec l'opérateur $\top$ ou $=$. Nous pouvons donc rejeter toutes les DC ne possédant aucun opérateur parmi $\{\leq, \geq, \neq, <, >\}$

3.1.3 Denial constraints maximales

Nous avons besoin également que les DC soit *maximales*. Dans [4] ils définissent une DC maximale comme étant :

Définition 20. La variante φ' d'une DC φ est dite *maximale* si $\varphi \preceq \varphi'$ [2] est qu'il n'existe pas de DC φ'' tel que $\varphi' \preceq \varphi'$ et que $edit(\varphi, \varphi'') = edit(\varphi, \varphi')$

Propriété 3. [4] Soit φ une DC sur S. Soit φ' une variante de φ sur S. Pour chaque attribut A de S tel que $\varphi(A) = \{\top, =, <, >\}$, si $\varphi'(A) \in \{\neq, \leq, \geq\}$ alors φ' n'est pas maximal.

2. voir définition du raffinement chapitre 2.3.1.6

Cette propriété vient de la définition de $Imp(\varphi)$. Soit φ_1, φ_2 deux DC sur S Soit une relation I sur S. Pour deux tuples $s, t \in I$ et un attribut A de S, si $\varphi_1 \in Imp(\varphi_2)$ alors $s.A\Theta_1 t.A$ implique $s.A\Theta_2 t.A$ avec $\Theta_1 = \varphi_1(A)$ et $\Theta_2 = \varphi_2(A)$. Les opérateurs $\leq$, $\geq$ et $\neq$ sont les seuls opérateurs qui impliquent d'autres opérateurs qu'eux même comme on peut le voir à la table 2.3(c'est également le cas pour l'opérateur $\top$, mais celui-ci est exclu de la propriété).

Grâce à cette propriété, nous savons maintenant qu'il est inutile de considérer toutes les insertions [3] de prédicats possibles. Nous avons seulement besoin d'insérer des prédicats avec les opérateurs $\{<, =, >\}$ lorsque que l'on considère les variantes de φ. Illustrons cela avec un exemple.

Exemple 10. Nous allons prendre deux DC sur la relation de la table 2.1 :

$$\varphi_1 : \{(Nom, =), (Revenu, =), (Numtel, \neq)\}$$

$$\varphi_2 : \{(Nom, =), (Revenu, \leq), (Numtel, \neq)\}$$

Nous savons que $\leq \in Imp(=)$ (voir table 2.3) donc nous avons $\varphi_2 \preceq \varphi_1$. Grâce à la propriété 3, φ_2 n'est pas une denial constraint maximale parce qu'elle contient l'opérateur $\leq$. Dans ce scénario, nous pouvons calculer un coût de réparation pour l'attribut $NumTel$ de 7 pour la DC φ_2 et la DC φ aura un coût de réparation de $Numtel$ de valeur 3.

	Nom	NumTel	Revenu
t1	Ayres	564-389	22k
t2	Ayres	564-389	22k
t3	Ayres	564-389	22k
t4	Stanley	930-198	24k
t5	Stanley	930-198	24k
t6	Stanley	930-198	24k
t7	Dustin	824-870	100k
t8	Dustin	824-870	100k
t9	Dustin	824-870	100k
t10	Dustin	824-870	100k

TABLE 3.3 – Correction avec φ_2 : 7 changement nécessaire dans la colonne $NumTel$

.

	Nom	NumTel	Revenu
t1	Ayres	322-573	21k
t2	Ayres	564-389	22k
t3	Ayres	564-389	22k
t4	Stanley	868-701	23k
t5	Stanley	930-198	24k
t6	Stanley	930-198	24k
t7	Dustin	179-924	25k
t8	Dustin	824-870	100k
t9	Dustin	824-870	100k
t10	Dustin	387-215	150k

TABLE 3.4 – Correction avec φ_1 : seulement 3 changements sont nécessaires.

.

Nous voyons que la DC raffinée a un coût de réparation plus faible. Ce n'est pas une coincidence puisque le Lemme suivant existe : [4]

3. Par insertion de prédicats, comprenons ici que nous passons de $\varphi(A) = \top$ à $\varphi(A) \neq \top$

Lemme 4. Soit Σ un ensemble de DC sur S et I une relation sur S. Soit 2 variantes de Σ noté Σ_1, Σ_2 et Σ_2 est un raffinement de Σ. Si $\Sigma \preceq \Sigma_1$ et $\Sigma_1 \preceq \Sigma_2$ alors $\Delta(I, I_1) \geq \Delta(I, I_2)$ avec I_1 la relation résultante de la réparation avec Σ_1 et I_2 la relation résultante de la réparation avec Σ_2.

Comme conséquence à ce Lemme, tous les ensembles de denial constraint Σ qui ne sont pas maximals ne donneront pas la meilleure réparation.

3.1.4 Limitation par bornes

Nous avons vu précédemment que pour une seule DC, nous avions $8^{|S|}$ variantes possibles (voir treillis à la figure 3.3). Nous avons donné des pistes pour limiter le nombre de candidats. Pour continuer dans cette lancée, nous allons déterminer deux bornes entre lesquelles le coût de la réparation la moins coûteuse se trouve. Nous avons la borne inférieure notée $\delta_l(\Sigma, I)$ et la borne supérieure notée $\delta_u(\Sigma, I)$. Nous savons que le coût de réparation $\Delta(I, I')$ pour la relation I sur S avec l'ensemble de DC Σ sera situé entre les deux bornes i.e $\delta_l(\Sigma, I) \leq \Delta(I, I') \leq \delta_u(\Sigma, I)$. L'idée étant qu'on puisse calculer les deux bornes très rapidement, là où $\Delta(I, I')$ ne peut être calculer qu'après avoir essayer la réparation. Considérons la propriété suivante :

Propriété 4. Soit deux ensemble de denial constraints Σ_1 et Σ_2 pour une relation I sur R, si $\delta_u(\Sigma_1, I) < \delta_l(\Sigma_2, I)$ alors Σ_2 n'est pas un bon candidat et peut être ignoré.

Démonstration.

Soit I une relation sur S et Σ un ensemble de DC.

Soit Σ_1, Σ_2 deux ensemble de DC, variantes de Σ.

Soit I_1, I_2 deux relations résultantes de la réparation de I avec les ensembles de DC Σ_1, Σ_2 respectivement.

Par hypothèse nous avons $\delta_l(\Sigma_1, I) < \Delta(I, I_1) < \delta_u(\Sigma_1, I)$ et $\delta_l(\Sigma_2, I) < \Delta(I, I_2) < \delta_u(\Sigma_2, I)$ et $\delta_u(\Sigma_1, I) < \delta_l(\Sigma_2, I)$.

Dès lors nous avons $\delta_l(\Sigma_1, I) < \Delta(I, I_1) < \delta_u(\Sigma_1, I) < \delta_l(\Sigma_2, I) < \Delta(I, I_2) < \delta_u(\Sigma_2, I)$.

Alors par transitivité $\Delta(I, I_1) < \Delta(I, I_2)$. I_2 est une réparation plus couteuse que I_1 i.e Σ_1 est un meilleur candidat que Σ_2. □

Cela veut dire que si la pire borne (borne supérieure) de la réparation de Σ_1 est moins bonne que la meilleure borne (borne inférieure) de la réparation de Σ_2 alors Σ_1 sera un meilleur ensemble DC pour la réparation que Σ_2 et donc Σ_2 peut être abandonnée. C'est intéressant pour l'élaboration d'un algorithme cherchant à trouver le meilleur ensemble de contraintes. En effet, si le calcul de ces bornes est moins couteux que le calcul du cout de réparation $\Delta(I, I')$ alors on peut réduire la complexité de cet algorithme.

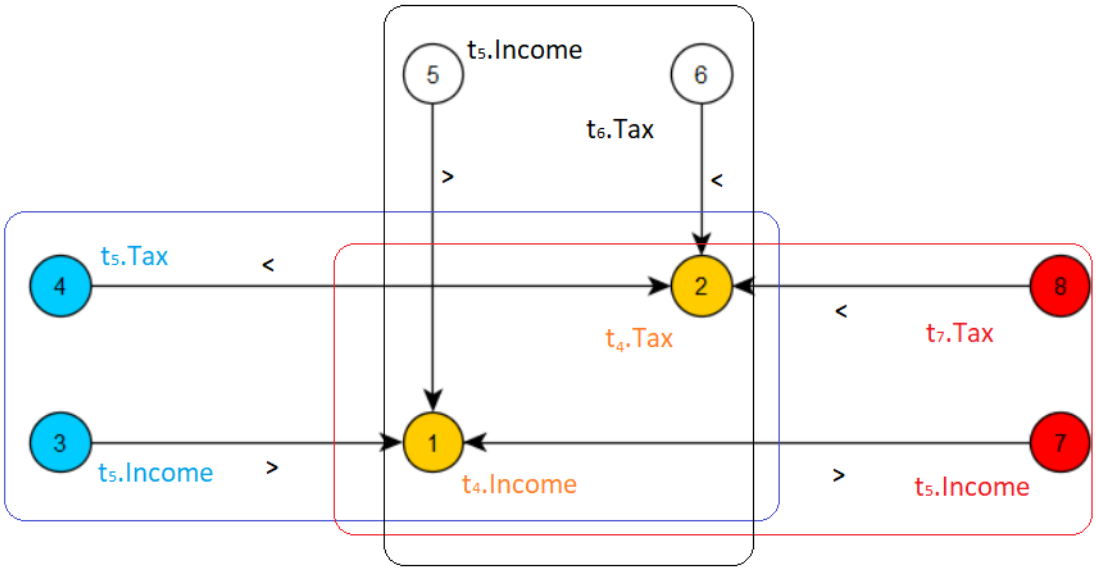

FIGURE 3.4 – Graphe de conflits pour φ

3.1.4.1 Graphe de conflits

Maintenant, nous allons introduire le *graphe de conflits* qui peut représenter les violations dans une relation I sur S. Dans un premiers temps, nous avons besoins de trouver l'ensemble de toutes les violations et ensuite obtenir nos deux bornes grâce à cet ensemble. On définition l'*ensemble des violations* comme étant : [4]

Définition 21. L'*ensemble des violations* $viol(I, \varphi) = \{(s,t)|(s,t) \not\models \varphi$ avec $s,t \in I\}$ est un ensemble de couple de tuples qui viole φ.L'ensemble de violation de Σ est $viol(I,\Sigma) = \cup_{\varphi\in\Sigma} viol(I,\varphi)$.

Définition 22. L' *ensemble cellules impliquées dans les violation* noté $cell(\varphi)$ est définie comme étant $cell(\varphi) = \{t.A, s.A|s,t \in I, (s,t) \in viol(I,\varphi), \varphi(A) \neq \top\}$. Les cellules de deux tuples $s,t \in I$ impliquées dans les violations de φ est définie par $cell(s,t\varphi) = \{t.A, s.A|s,t \in I, \varphi(A) \neq \top\}$.

Avec le graphe de conflits G nous pouvons représenter les violations dans la relation I. Pour chaque $(s,t) \in viol(I,\varphi)$ il y a un arc pour $cell(s,t,\varphi)$ dans G. Une bonne réparation I' consiste à réparer la base de données de telle manière à ce que tout les arcs soient supprimés. Dès lors après une réparation, le graphe est vide.

Exemple 11. Prenons un exemple sur la table 2.1 ainsi qu'une DC que nous avons déjà utilisé auparavant :

$$\varphi' = \{(Revenu, >), (Taxe, <)\})$$

Pour notre relation, l'ensemble des violations est (voir figure 3.2) :

$$viol(I, \varphi') = \{(t_5, t_4), (t_6, t_4), (t_7, t_4)\}$$

Dans notre graphe, $(t_5, t_4) \in viol(I, \varphi')$ est représenté par $cell(t_5, t_4; \varphi')$ qui est égale à $\{t_5.Revenu, t_4.Revenu, t_5.Taxe, t_4.Taxe\}$. Nous voulons éliminer les conflits, ce qui se traduit par éliminer les arcs du graphe. Introduisons d'abord deux Lemmes et une définition : [4]

Définition 23. On note $\min_a dist(t.A, a)$ le poid d'un arc t.A, i.e, le coût minimum qui devrait être payé pour réparer t.A avec une valeur a.

Définition 24. $\mathbb{V}(G)$ est la **couverture de poids minimum** du graphe G correspondant à Σ avec le poids

$$||\mathbb{V} * (\mathbb{G})|| = \sum_{t.A \in \mathbb{V}(G)} min_a dist(I(t.A), a)$$

Lemme 5. Soit I une relation sur S. Pour n'importe quelle réparation I' de I, i.e, $I' \models \Sigma$, nous avons $\Delta(I, I') \leq ||V * (G)||$.

Définition 25. Soit I une relation sur S. Soit Σ un ensemble de DC. Soit s,t deux tuples de I. Nous définissons degré de Σ noté $Deg(\Sigma)$ comme étant :

$$Deg(\Sigma) = \sum_{\varphi \in \Sigma} |cell(s, t, \varphi)|$$

Dans [4] ils définissent la borne supérieure et la borne inférieure du coût de réparation comme étant :

$$\delta_l(\Sigma, I) = \frac{||\mathbb{V}(G)||}{Deg(\Sigma)}$$

$$\delta_u(\Sigma, I) = \sum_{t.A \in \mathbb{V}(G)} dist(I(t.A), fv)$$

Si nous revenons à notre example et que nous supposons que nous avons les distances suivantes :

$$\forall a \in dom(A)\ avec\ a \neq b. \begin{cases} dist(a, a) = 0 \\ dist(a, b) = 1 \\ dist(a, fv) = 1.5 \\ dist(fv, fv) = 1.5 \\ dist(fv, b) = 1 \end{cases}$$

Donc si chaque arc a un poids de 1 (=$dist(a, b)$) et si nous posons $\mathbb{V}(G)$ = $\{t_4.Tax\}$ nous avons $||\mathbb{V}(G)||$ = 1. Nous avons aussi $Deg(\Sigma)$ = 4, donc en utilisant les formules pour le calcul des bornes supérieure et inférieure : $\delta_l(\Sigma, I)$=$\frac{||\mathbb{V}(G)||}{Deg((\Sigma)}$ = $\frac{1}{4}$ = 0.25 and $\delta_u(\Sigma, I)$ = $\sum_{t.A \in \mathbb{V}(G)} dist(I(t.A), fv)$ = $dist(a, fv)$=1.1 .

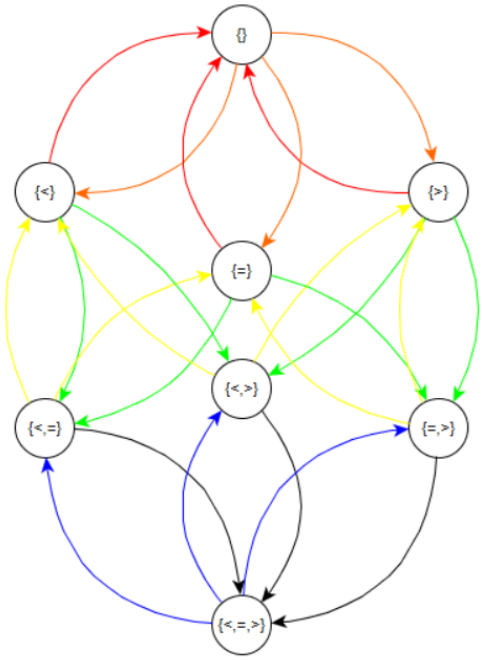

FIGURE 3.5 – Chaque types de transition aura son propre poids.

3.1.5 Coût dans le treillis

Nous pouvons donc modifier une DC φ en modifiant ses prédicats. Nous avons vu précédemment que les opérateurs peuvent être représenter dans un treillis et modifier l'opérateur consiste à se déplacer dans le treillis. Le coût dépend donc du nombre d'arcs parcourus dans le treillis. Nous verrons dans cette sous-section quel est le poids pour chacun de ces arcs.

Il est possible de représenter le treillis comme à la figure 3.5. Chaque couleur représente un type de changement. Par exemple chaque arc orange représente un changement depuis l'opérateur $\{\} \equiv \perp$ vers un élément de OP de taille 1. Discutons maintenant du poids que l'on peut donner à chacun de ces arcs.

Définition 26. Soit I une relation sur S. Soit $\mathbb{T}$ un treillis pour un attribut $A \in S$. On note $w(a, b)$ le poids d'un arc de $\mathbb{T}$ depuis un élément de OP de taille a vers un élément de OP de taille b.

Nous avons vu précédemment qu'une DC ne doit pas contenir l'opérateur $\perp$ car sinon la DC est triviale (voir Lemma 3). De ce fait nous pouvons dire :

— Il ne faut pas transformer un opérateur en $\perp$. Les arcs allant vers $\perp$, i.e les arcs rouges dans la figure 3.5, ne doivent jamais être empruntés. Pour cela, le poids de

ces arcs sera $w(1,0) = +\infty$

— Une DC ne devrait jamais avoir de prédicat dont l'opérateur est $\perp$. Si c'est le cas, il est nécessaire de corriger ce prédicat. Pour aider à la correction, nous posons $w(0,1) < 0$, dès lors emprunter un arc orange ne coute rien, c'est même positif et donc on encourage à faire cette modification. On force ainsi la correction de tels prédicats.

Nous avons également vu que si une DC n'est pas maximale, alors elle ne donnera pas la meilleure réparation. Il faut donc sanctionné les transitions qui partent d'un prédicat maximal vers un prédicat non maximal. Par la propriété 3, nous savons que les prédicats qui contiennent les opérateurs $\leq, \geq$ ou $\neq$ ne sont pas maximals. Pour encourager à rendre une DC maximale, nous poserons $w(1,2)$ = + ∞ et $0 < w(2,1) < \alpha$ avec α un seuil que nous devrons définir plus tard.

Il nous reste à évaluer $w(2,3)$ et $w(3,2)$. $w(3,2)$ représente une insertion de prédicat tandis que $w(2,3)$ est une suppression de prédicat. A priori, il n'y a aucune raison de mettre un poid particulier sur ces transitions. Nous poserons que $0 < w(3,2) = w(2,3) < \alpha$.

3.1.6 Variation d'un ensemble de contraintes

Nous avons jusqu'à présent abordé le sujet de la variation d'une seule contrainte. Pour faire varier un ensemble de contraintes Σ, il suffit d'appliquer le même raisonnement à au moins une des DCs de Σ. Mais nous devons statuer si on peut ajouter des DC ou même en supprimer. Nous déciderons d'aller dans le même sens que l'article sur lequel nous nous sommes initialement basés [1] et donc nous n'ajouterons pas de DC ni n'en retirerons. Retirer une DC diminue le nombre d'erreurs détectées (l'ensemble des DC est sur-raffiné)et on peut supposer que chaque DC avait un sens lors de sa création (mais peut évidement être imprécis). L'ajout de DC est bien plus compliqué a mettre en oeuvre et on peut rapidement obtenir des DC qui n'ont pas de sens et qui détecteraient des erreurs qui ne le sont pas (l'ensemble de DC est sur-simplifié).

Puisque nous avons décidé de ne pas supprimer des DCs dans un ensemble de DC Σ, nous devons faire attention à la manière dont nous modifions Σ. Si cet ensemble contient au moins deux DC et qu'on modifie une des contraintes jusqu'à ce qu'elle soit identique à une autre, alors ça revient à faire une suppression.

Exemple 12. Soit S = $\{A, B, C\}$ Prenons $\Sigma_1 = \{\varphi_1 = \{(A, <), (B, \neq)\}\varphi_2 = \{(A, <), (C, \neq)\}\}$ et la variation $\Sigma_2 = \{\{(A, <), (B, \neq)\}, \{(A, <), (B, \neq)\}\}$ nous voyons que la variation contient deux fois la même DC, et donc on peut réécrire $\Sigma_2 = \{\{(A, <), (B, \neq)\}\}$ ce qui équivaut à une suppression.

Notons qu'il est possible de faire des variations inutiles :

Exemple 13. Reprenons S = $\{A, B, C\}$ et $\Sigma' = \{\varphi_1 = \{(A, <), (B, \neq)\}\varphi_2 = \{(A, <), (C, \neq)\}\}$ et la variation $\Sigma' = \{\varphi'_1 = \{(A, <), (C, \neq)\}\varphi'_2 = \{(A, <), (B, \neq)\}\}$, les deux DC ont été modifiés mais nous avons le même ensemble !

3.2 θ-tolerant model

Dans cette section nous allons enfin aborder le *modèle de réparation θ-tolérant*. θ représente ici un seul de variation sur l'ensemble des contraintes Σ, nous ne voulons donc pas une variante de contraintes dont le coût serait plus grand que θ : $\Theta(I, I') \leq \theta$. L'idée est d'éviter le sur-raffinement et donc d'éviter de laisser certaines données erronées. Afin d'éviter la sur-simplification, nous utilisons le principe du changement minimum. Nous avons donc besoin d'une relation réparée I' de la relation originale I et minimisant le coût de réparation $\Delta(I, I')$.

Trouver la meilleure réparation suivant le modèle θ-tolérant, i.e trouver la réparation de coût minimum est un problème de la classe NP-difficile. La classe de problèmes NP-difficile est une classe dont les problèmes sont au moins aussi difficile que les problèmes les plus difficiles de la classe NP [4]. Nous devons donc retenir ici que ce n'est pas possible de résoudre la réparation θ-tolérante en un temps polynomial. L'approche naïve est de récupérer toutes les variantes Σ' de Σ et de ensuite calculer $\Theta(I, I')$ avec I' la relation résultante de la réparation suivant les DC de Σ'. Si $\Theta(I, I') \leq \theta$, on calcule le coût de réparation. Ensuite on compare le coût de la réparation I' pour chaque Σ' et on garde la réparation la moins coûteuse. Cette méthode est évidement très haute en complexité.

Nous avons vu qu'on remplaçait chaque donnée erronée par une variable fraiche fv puis nous essayons de remplacer certaines de ces variables fraiches par des vraies valeurs. Plus une réparation privilégies les vraies valeurs aux valeurs fraiches, plus elle aura de chance de minimiser le coût de réparation.

Maintenant, considérons $\mathbb{D} = \Sigma'_1 * \Sigma'_2 * ...\Sigma'_{|\Sigma|}$ où chaque $\Sigma'_i \in \mathbb{D}$ est une variante de Σ obtenu par variations. Considérons que ces variantes sont limitées par θ, donc nous avons $\Theta(\Sigma, \Sigma') \leq \theta$. L'algorithme 1 retourne la meilleure relation I_{min} de l'ensemble des contraintes Σ_{min}. L'algorithme est simple : pour chaque Σ_i , si la borne inférieure est plus petite que la borne supérieure (voir la propriétés 4), nous mettons à jour la valeur de δ_{min} car une meilleure réparation I_i a été trouvée.

Exemple 14. Pour prendre un exemple, imaginons que nous avons un $\theta = \frac{?}{?}$ et un ensemble de variations de contraintes $\mathbb{D} = \{\Sigma_1, \Sigma_2\}$ avec comme premier ensemble $\Sigma_1 = \{\varphi'\}$ et comme second ensemble $\Sigma_2 = \{\varphi''\}$ avec :

$$\varphi' = \{(Revenu, >), (Taxe, <)\}$$

$$\varphi'' = \{(Revenu, >), (Taxe, =)\}$$

4. Les problèmes NP sont des problèmes dont une solution peut être vérifiée comme étant bonne en un temps polynomial

Algorithme 1 : θ-TolerantRepair($\mathbb{D}, \Sigma, I$)

Entréé : Relation I, ensemble de DC Σ, ensemble $\mathbb{D}$ de variantes de DC bornées par θ

Sortie : Une relation réparée I_{min}

```
δ_min = δ_u(Σ, I)
pour chaque variante de contrainte Σ_i ∈ 𝔻 faire
    si δ_l(Σ_i, I) ≤ δ_min alors
        I_i = DATAREPAIR(Σ_i, I, 𝕍(G_i), δ_min)
        si Δ(I, I_i) ≤ δ_min alors
            δ_min = Δ(I, I_i)
            I_min = I_i
return I_min
```

Nous avons déjà fait le graphe de conflit pour la DC φ' que l'on retrouve à la figure 3.4 et on sait également que $\delta_u(\Sigma_1, I) = 1.1$

Pour $\Sigma 2$ nous obtenons le graphe de conflit de la figure 3.6 (et les violations se retrouvent à la figure 3.7) avec $\mathbb{V}(G_2) = \{t_2.Tax, t_3.Tax, t_5.Tax, t_6.Tax, t_7.Tax\}$. Nous $Deg(\Sigma_2) = Deg(\varphi')$ [5], donc nous avons $\delta_l(\Sigma_2, I)$= $\frac{6}{2}$ = 1,5. Rappelons tout de même que $\delta_u(\Sigma_1, I)$ =1.1, donc nous avons $\delta_u(\Sigma_1, I) < \delta_l(\Sigma_2, I)$ ce qui signifie que nous pouvons ignorer $\Sigma 2$ et donc ne pas appeler la fonction DATAREPAIR pour Σ_2. Nous parlerons de la fonction DATAREPAIR plus tard.

Parlons maintenant de la complexité de l'algorithme. Disons que l est le nombre de contraintes impliquées dans Σ dès lors nous pouvons dire que la construction du graphe de conflits G_i pour chaque $\Sigma_i \in \mathbb{D}$ est en $O(|I|^l)$. L'algorithme de réparation de données a une complexité en temps de $O(|I|^l)$ et l'algorithme 1 a une complexité en temps de $O(|I|^l|\mathbb{D}|)$.

3.3 Réparation au coût minimal

Après avoir utilisé le modèle θ-tolérant, nous connaissons enfin quel ensemble de denial constraints Σ' (obtenu en faisant varier Σ) nous devons utiliser. Mais pour l'instant, nous n'avons pas encore vu comment effectuer la réparation. Nous avons seulement expliqué que chaque mauvaise donnée est remplacée par une variable fraiche et ensuite remplacée par une vraie valeur si possible. Dans cette section, nous allons nous concentrer sur la réparation de données en minimisant le coût et en se basant sur Σ'. Pour ce faire, nous allons devoir nous assurer que nous ne créons pas de nouvelle violation après avoir

5. Même raisonnement que précédement, nous avons 4 cellules impliquées.

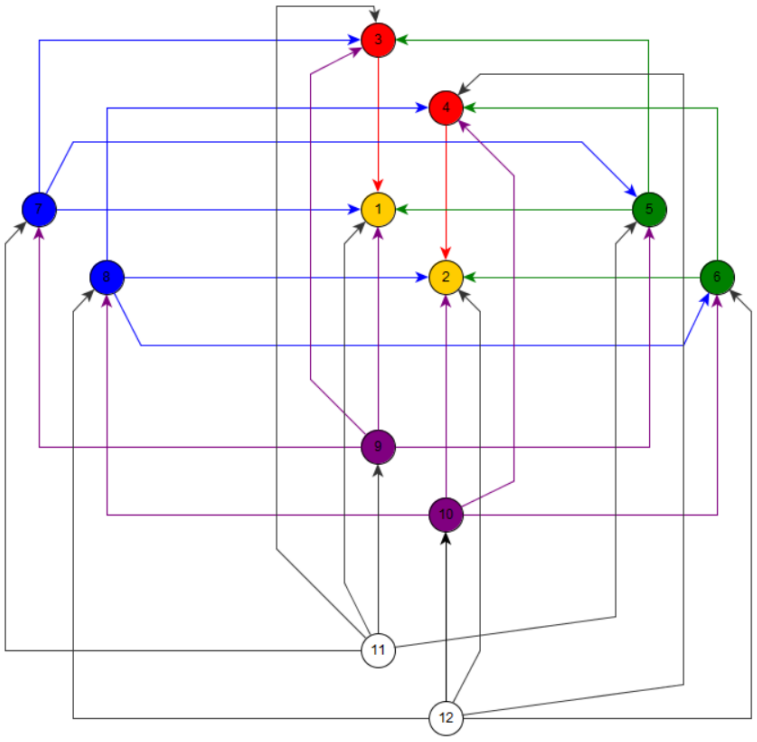

FIGURE 3.6 – Graphe de conflit pour Σ_2 avec :
nombre pair pour les Taxe, nombre impair pour le Revenu.
t_1 en jaune, t_2 en rouge, t_3 en vert,
t_4 en bleu, t_5 en violet et t_6 en blanc.

/	t_β										
t_α	/	1	2	3	4	5	6	7	8	9	10
	1	/	V	V	V	V	V	V	V	V	V
	2	F	/	V	V	V	V	V	V	V	V
	3	F	V	/	V	V	V	V	V	V	V
	4	V	V	V	/	V	V	V	V	V	V
	5	F	F	F	V	/	V	V	V	V	V
	6	F	F	F	V	V	/	V	V	V	V
	7	F	F	F	V	F	F	/	V	V	V
	8	V	V	V	V	V	V	V	/	V	V
	9	V	V	V	V	V	V	V	V	/	V
	10	V	V	V	V	V	V	V	V	V	V

FIGURE 3.7 – Toutes les violations pour φ''

corrigé une donnée. Par exemple, si nous choisissons de changer la valeur de $t_5.Taxe$ à $22k$, nous réglons la violation sur (t_5, t_4) que nous avions avec la denial constraint φ = $\{(Revenu, >)(Taxe, <)\}$. En changeant cette valeur de cette manière, on crée une nouvelle violation (t_8, t_5).

Il est important de rappeler que trouver une réparation de coût minimum est un problème NP-difficile. Pour ces problèmes, il est important de trouver une approximation. Pour la suite de cette section, nous allons noter $\mathbb{C}$ les cellules sélectionnées dans $\mathbb{V}(G)$.

3.3.1 Identification des suspects

Trouver la réparation de coût minimal à partir d'un ensemble Σ' de DC est un problème NP-complet [4], nous avons donc besoin d'un algorithme d'approximation pour trouver les solutions en un temps polynomial. Nous allons commencer par travailler avec la couverture $\mathbb{V}(G) = \mathbb{C}$ et nous assurer qu'aucune nouvelle valeur n'introduit de nouvelle violations. Nous savons que les variables fraiches n'introduisent aucune violation puisque par définition une variable fraiche rend tout prédicat faux et donc la DC est vraie. Mais puisqu'on tente de trouver une valeur à ces variables, nous devons détecter quels tuples sont les plus susceptibles d'introduire de nouvelles violations après une correction des cellules de $\mathbb{C}$.

Définition 27. Les *suspects de* φ noté $susp(\mathbb{C}, \varphi)$ est un ensemble de couple de tuples qui satisfont tous les prédicats de φ qui n'implique pas de cellules impliquées de $\mathbb{C}$, i.e les tuples qui satisfont la *condition de suspicion :*

$$sc(s, t, \varphi) = \{s.A\Theta t.A | s, t \in I; P : (A, \Theta) \in pred(\varphi); s.A, t.A \notin \mathbb{C}\}$$

Soit deux tuples $s, t \in I$ on peut utiliser le graphe à la figure 3.8 pour représenter les attributs et le lien avec les suspects et les cellules impliquées. Les cercles représentent les éléments de $\mathbb{C}$ i.e les éléments qui vont changer et les carrés représentent les éléments qui ne sont pas de $\mathbb{C}$ i.e des éléments qui ne doivent pas changer. Les flèches représentent les opérateurs de nos prédicats. Par exemple la flèche allant de 1 à 5 représentent l'opérateur Θ_A du prédicat (A, Θ_A). Si toutes les flèches noires de φ sont satisfaites alors nous avons un suspect.

Le lemme suivant nous permet d'affirmer qu'avoir la liste des suspects d'une DC contient toutes les violations de cette même DC [4] :

Lemme 6. Soit I une relation sur S , Σ un ensemble de DC ,φ une DC de Σ et $\mathbb{C}$ une approximation de la couverture de poids minimum pour un graphe de conflit G correspondant à Σ, I nous avons toujours $viol(I, \varphi) \subseteq susp(\mathbb{C}, \varphi)$

Démonstration. Soit I une instance sur S , deux tuples $s, t \in I$ et φ une DC d'un ensemble de contrainte Σ pour I.
Supposons que nous avons $(s, t) \in viol(I, \varphi)$ alors tous les prédicats de φ sont vrais pour (s, t) i.e $(s, t) \models \varphi$ et donc ils satisfont la condition de suspicion □

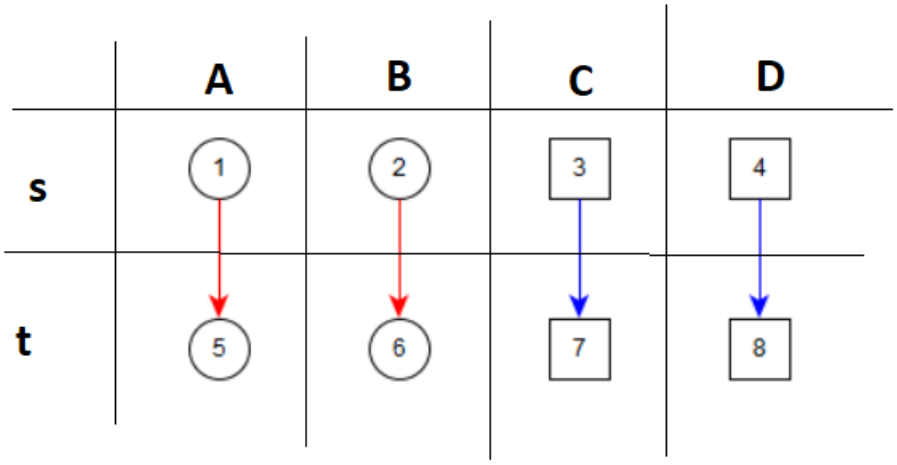

FIGURE 3.8 – s, t sont deux suspect

Exemple 15. Reprenons l'exemple sur $\varphi' = \{(Revenu, >), (Taxe, >)\}$ avec comme instance I le tout en relation avec le graphe de conflits à la figure 3.4. Nous allons uniquement changer $t_4.Taxe$ comme nous l'avons fait à la table 3.1. Donc nous prenons $\mathbb{C}$ = $\{t_4.Taxe\}$ and $susp(\mathbb{C}, \varphi') = \{(t_4, t_1), (t_4, t_2), (t_4, t_3), (t_5, t_4), (t_6, t_4), (t_7, t_4), (t_8, t_4), (t_9, t_4), (t_{10}, t_4)\}$ Regardons en détails (t_4, t_1) à la figure 3.9. Nous avons $t_4.Taxe \in \mathbb{C}$ représenté par un cercle et $t_1.Taxe, t_4.Revenu, t_1.Revenu \notin \mathbb{C}$ et donc représenté par un carré. Le prédicat $t_4.Taxe > t_1.Taxe$ a une flèche rouge car il contient un cercle i.e une valeur de Taxe qui doit être changée et son opérateur $>$ est remplacé par $<$ (pour représenté la situation actuelle) et $t_4.Revenu > t1.Revenu$ ne contient pas de cercle donc sa flèche est noire i.e aucune valeur à changer.

Nous avons comme condition de suspicion sur (t_4, t_1) : $sc(t_4, t_1, \varphi) = \{t_4.Revenu > t_1.Revenu\}$. En se référant à la table 2.1, nous avons $t_4.Revenu > t_1.Revenu$ ce qui implique que tous les prédicats avec une flèche bleue, c'est à dire ceux impliquant une valeur dans $\mathbb{C}$ sont satisfaits. Nous avons donc bien (t_4, t_1) suspect , donc une réparation sur $t_4.Taxe$ peut introduire de nouvelles violations avec $t_1.Taxe$, i.e $f(t_4.Taxe) < t_1.Taxe$ avec f fonction de réparation pour la relation I.

3.3.1.1 Contexte de réparation

Pour chaque couple de tuples $(s, t) \in susp(\mathbb{C}, \varphi)$ nous pouvons définir un *contexte de réparation* de la DC φ qui se note $rc(s, t\varphi)$. Ce contexte de réparation nous permettra d'assurer que les réparations ne vont pas introduire de nouvelles violations. Le contexte de réparation d'une DC φ est défini comme étant :

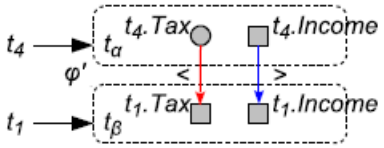

FIGURE 3.9 – Les conditions de suspicions sont représentées par des flèches bleues et le contexte de réparation est représenté par des flèches rouges (avec l'opérateur inverse)

Définition 28. Soit I une relation sur S, soit φ une DC sur I, soit $s, t \in I$, et f une fonction de réparation pour I le contexte de réparation de φ est noté $rc(s, t, \varphi)$ et vaut :

$$\begin{aligned} rc(s,t,\varphi) = &\{t.A\overline{\theta}f(s.A) | (A,\Theta) \in pred(\varphi), s,t \in I\} \cup \\ &\{f(t.A)\overline{\theta}s.A | (A,\Theta) \in pred(\varphi), s,t \in I\} \cup \\ &\{f(t.A)\overline{\theta}f(s.A) | (A,\Theta) \in pred(\varphi), s,t \in I\} \end{aligned}$$

Si nos reprenons notre figure 3.8 ou 3.9, les arcs rouges représentent le contexte de réparation.

Propriété 5. Soit I une relation sur S et Σ un ensemble de DC pour I. Toute relation I' issue de la réparation de I i.e $I' = f(I)$ est *valide* si I' satisfait tous les contextes de réparations.

Exemple 16. (suite) Pour le couple de suspect $(t_4, t_1) \in susp(\mathbb{C}, \varphi)$ nous avons le contexte de réparation $rc(t_4, t_1, \varphi) = \{t_4.Taxe \geq t_1.Taxe\}$ que l'on obtient en prenant l'opposé de l'opérateur du prédicat $(Taxe, <)$ (voir définition) ce qui est représenté par une flèche rouge dans la figure 3.8. En considérant toutes les paires de tuples de $susp(\mathbb{C}, \varphi)$, nous obtenons l'entièreté du contexte de réparation de φ visible graphiquement à la figure 3.10. Par les contextes de réparations de φ : $f(t_4.Taxe) > t_1.Taxe = 0$ et $f(t_4.Taxe) \leq t_5.Taxe = 0$ par transitivité , nous pouvons affirmer avec certitude que la variable fraiche fv que nous avions assigné vaut $fv = 0$.

Nous avons défini le contexte de réparation pour une DC φ. Nous pouvons maintenant définir le contexte de réparation pour un ensemble Σ de DC :

Définition 29.

$$rc(\mathbb{C}, \Sigma) = \bigcup_{(s,t) \in susp(\mathbb{C},\varphi), \varphi \in \Sigma} rc(s, t, \varphi)$$

Nous pouvons exprimer le problème de réparation d'une relation I comme étant un problème de minimisation à résoudre :

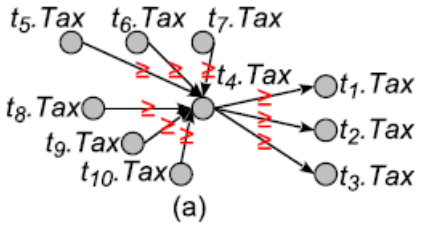

FIGURE 3.10 – Contexte de réparation

$\min \sum_{t.\in\mathbb{C}} dist(t.A, f(t.A))$
sous la contrainte$rc(\mathbb{C}, \Sigma)$

Il est possible d'utiliser la programmation linéaire pour résoudre ce problème d'optimisation pour les valeurs numériques et une "value frequency map" pour les chaines de caractères [4].

Nous pouvons décomposer $\mathbb{C}$ en plusieurs sous ensemble $\mathbb{C}_1, \mathbb{C}_2, ..., \mathbb{C}_n$ tel que $\forall s, t \in I$, il n'existe pas de $f(t.A)\overline{\Theta_A} f(s.A) \in rc(s, t, \varphi)$ pour $(s, t) \in susp(\mathbb{C}, \varphi)$ avec $\varphi \in \Sigma$ et Σ un ensemble de DC pour la relation I sur S. Nous pouvons dire que $rc(\mathbb{C}, \Sigma) = rc(\mathbb{C}_1, \Sigma) \cup ... \cup rc(\mathbb{C}_n, \Sigma)$ et donc nous pouvons résoudre notre problème en résolvant chaque $rc(\mathbb{C}_k, \Sigma)$ individuellement.

Chapitre 4

Différence par rapport à l'article de base

Dans ce chapitre, nous parlerons des différences entre l'article de référence sur le modèle θ-tolérant et ce mémoire. Certaines notions et définitions ont été revues, et le traitement théorique a été amélioré. Certaines erreurs ont été corrigées notamment des erreurs concernant les variations de contraintes et coût de réparation de la base de données.

Pour la rédaction de ce mémoire, nous nous sommes inspiré de deux articles. Le premier est un article sur le modèle θ-tolérant, le second est un article qui porte sur les denial constraints. Ils utilisent une approche différente et une définition différente de la notre. Dans les deux articles [4, 3] ils définissent une DC comme étant :

Définition 30. Considérons un schéma de relation R avec comme attributs $att(R)$. Soit l'espace de prédicat $\mathbb{P}$ qui est un ensemble de prédicat P de la forme $v_1 \phi v_2$ ou $v_1 \phi c$ avec $v_1, v_2 \in t_x.A,\ x \in \{\alpha, \beta\}$, $t_\alpha, t_\beta \in R$, $A \in attr(R)$, c une constante et $\phi \in \{=, <, >, \leq, \geq, \neq\}$ est un opérateur. Une *denial constraint* (DC)

$$\varphi : t_\alpha, t_\beta, ... \in R, \neg(P_1 \wedge P_2 \wedge ... \wedge P_m)$$

signifie que pour tout tuples t_α, t_β dans R, tous les prédicats $P_i \in pred(\varphi)$, i = 1,....,m, ne devraient pas être tous vrais en même temps.

Une DC peut donc être vue comme une conjonction de prédicats et l'un de ses prédicats doit être faux afin que la DC soit vraie, i.e si pour deux tuples chaque prédicat est vrai, alors il y a au moins une donnée erronée dans l'un des deux tuples.

La définition bien qu'étant différente d'un point syntaxique représente toutes les deux la même chose. Cette définition-ci permet en plus de comparer un attribut à une constante. Bien qu'il puisse être intéressant d'avoir des denial contraints qui fixe un salaire minimum par exemple $t.Revenu > 10k$ ou exprime le fait qu'un revenu ou une taxe ne peut être négative $(t.Revenu > 0) \wedge (t.Taxe > 0)$ cela crée plusieurs problèmes.

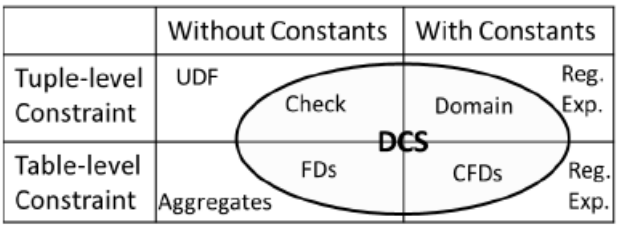

FIGURE 4.1 – La denial constraint tel que définie dans les articles de référence peut exprimer plusieurs types d'autres contraintes [3]

Dans notre définition, nous avons un nombre de prédicats maximum qui est la norme de S. Ici nous ne sommes pas limités et ce à cause des constantes. Et cela peut générer des problèmes lorsque l'on tente de changer une DC lors de la réparation. En effet, si notre DC est de la forme $\varphi = \neg(t.A > 10)$ doit-on considérer toutes les variations possibles ($\varphi' = \neg(t.A > 11)$,$\varphi'' = \neg(t.A > 12)$) ? Et malgré que cela puisse être un problème, ce n'est jamais abordé dans l'article [4].

Concernant les variations de contraintes, pour les auteurs de l'article, modifier un prédicat consiste à le supprimer puis le remplacer par un autre différent. Ce concept comporte un problème : toutes les modifications ne devraient pas être traités de la même façon. Par exemple si notre prédicat est de $(A, <)$ il devrait être moins couteux de le modifier en $(A, \leq)$ que de le modifier en $(A, \geq)$. En effet, pour modifier $\{<\}$ en $\leq \equiv \{<, =\}$, il suffit d'ajouter l'opérateur $=$. Pour modifier $\{<\}$ en $\geq \equiv \{>, =\}$ il faut ajouter $=$ mais aussi ajouter $>$, et retirer $<$. Notre système de treillis reflète mieux le fait que la seconde modification est plus grande et donc plus couteuse.

Chapitre 5

Implémentation

Nous avons décidé d'implémenter l'algorithme θ-tolérant en utilisant les différents concepts vus lors des chapitres précédents pour tester son efficacité. Dans ce chapitre, nous expliquerons comment nous avons implémenté les différentes notions. Nous ne donnerons pas toutes les explications du code car le code joint avec ce mémoire comporte des commentaires suffisamment explicites. Nous détaillerons les idées générales, les choix d'implémentations tels que le langage choisi, le format des bases de données, ... Nous terminerons par une discussion à propos de l'efficacité de l'algorithme, du choix de la valeur de θ , etc.

5.1 Choix de langage et format de base de données

Le langage de programmation que nous avons choisi est le Python et le développement de l'outil s'est fait avec *Jupyter Notebook*, une application web et open-source qui permet de créer et partager du code , des équations, etc. Cet outil est utilisé entre autres pour la visualisation de données, le machine learning, des simulations, transformation et nettoyage de données, etc. Il supporte plus de 40 langages de programmation incluant entre autre le Python, le Scala et R.

Le langage choisi, le python a l'avantage d'être plus facilement lisible. Puisque le but de l'implémentation est de regarder si l'algorithme est performant et fonctionne bien, un code lisible est nécessaire. Le débug est plus facile que d'autres langages comme le C++ et le Java. Puisqu'une interface utilisateur graphique n'est pas nécessaire, le python se présentait comme étant une très bonne solution.

La première chose dont nous avons besoin est de bases de données. Le choix a été fait de les stocker dans des fichiers $.txt$. L'avantage de ce format est que ne nous sommes soumis à aucune contrainte particulière d'un SGBD, nous pourrons donc stocker des variables fraiches, des entiers et des chaines de caractères comme bon nous semble. Le but n'étant pas d'utiliser l'algorithme développé pour faire des réparations concrètes et réelles, nous n'avions donc aucunement besoin de choisir un autre format. Les fichiers textes doivent être formatés de la façon suivante :

— La première ligne contient le nom des attributs séparés chacun d'un espace.
— La seconde ligne contient le type de variable parmi str , int et $float$.
— Chaque ligne qui suit contient un tuple de la base de données. Chaque attribut est séparé avec un espace.

Parmi les bases de données , nous retrouvons bien évidement la relation de la table 2.1.

5.2 Code

5.2.1 Classe Predicate et DC

Le premier choix fut de faire deux classes pour représenter les prédicats et les denial constraint.

La classe prédicat représente donc un prédicat comme nous l'avons défini, il y a donc un attribut et un opérateur pour ce prédicat. Parmi les opérateurs, nous ne retrouvons que les 6 opérateurs suivants : $<, >, =, \neq, \geq$ et $\leq$, les opérateurs $\top$ et $\bot$ n'étant pas nécessaire. En effet, un prédicat avec l'opérateur $\top$ est toujours vrai, nous ne les écrivons même pas dans l'écriture abrégée dans les chapitres théoriques 2 et 3. L'opérateur $\bot$ étant toujours faux, aucune DC ne devrait l'avoir.

La classe DC est une collection (liste) de prédicats. Différentes fonctions permettent de modifier une DC ou de vérifier si une DC est respectée.

Chapitre 6

Conclusion

TODO

Bibliographie

[1] Description des données du registre national et du registre bcss. https://www.ksz-bcss.fgov.be/sites/default/files/assets/services_et_support/cbss_manual_fr.pdf. accessed : 2018-02-15.

[2] ics relational database model. http://databasemanagement.wikia.com/wiki/Relational_Database_Model. Accessed : 2018-02-13.

[3] Xu Chu, Ihab F. Ilyas, and Paolo Papotti. Discovering denial constraints. *PVLDB*, 6(13) :1498–1509, 2013.

[4] Shaoxu Song, Han Zhu, and Jianmin Wang. Constraint-variance tolerant data repairing. In *SIGMOD Conference*, pages 877–892. ACM, 2016.

[5] Aoqian Zhang, Shaoxu Song, Jianmin Wang, and Philip S. Yu. Time series data cleaning : From anomaly detection to anomaly repairing. *PVLDB*, 10(10) :1046–1057, 2017.