République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieure et de la Recherche Scientifique

Université Ahmed Draia - Adrar

Faculté des Sciences et de la Technologie

Département des Mathématiques et Informatique

Mémoire de fin d'étude, en vue de l'obtention du diplôme de Master en informatique

Option :- Système Intelligent (SI)

Thème

Maximisation d'utilité des données pour la préservation de confidentialité

Préparé par

BENDRISSE Samira et BOUALLALA Zineb

Encadré par

Dr.KABOU Salaheddine

Membres du jury:

Dr.DAHOU Abdelghani	Président
Mr.CHOGUEUR Djilali	Examinateur
Dr.KABOU Salaheddine	Encadreur

Année Universitaire 2019/2020

Résumé

Aujourd'hui, de plus en plus les données du réseau social sont rendues publiquement disponibles à des fins d'analyse des données. Bien que cette analyse soit importante pour les chercheurs, il peut y avoir un risque de violation de la vie privée des utilisateurs constituant ce réseau social.

La préservation de la confidentialité pour les données publiées (PCDP) étudie comment immuniser les données contre les attaques de la vie privée.

L'anonymisation des données est l'une des solutions qui peuvent être utilisées pour préserver la confidentialité des données tout en assurant leurs utilisations.

Dans ce travail, nous décrivons une approche fondée sur les modèles qui guide l'éditeur de données dans son processus d'anonymisation, et les algorithmes qui permettent de minimiser le risque de ré-identification tout en préservant l'utilité des données.

ملخص

اليوم ، يتم إتاحة المزيد والمزيد من بيانات الشبكات الاجتماعية للجمهور لتحليل البيانات. على الرغم من أهمية هذا التحليل للباحثين ، فقد يكون هناك خطر انتهاك خصوصية المستخدمين الذين يشكلون هذه الشبكة الاجتماعية.

يدرس الحفاظ على سرية البيانات المنشورة (PPDP) كيفية جعل البيانات محصنة ضد هجمات الخصوصية.

إخفاء هوية البيانات هو أحد الحلول التي يمكن استخدامها للحفاظ على سرية البيانات مع ضمان استخداماتها.

في هذا العمل ، نصف النهج القائم على النموذج الذي يوجه محرر البيانات في عملية إخفاء الهوية ، والخوارزميات التي تقلل من مخاطر إعادة تحديد الهوية مع الحفاظ على فائدة البيانات.

Abstract

Today, more and more social network data is made publicly available for data analysis. Although this analysis is important for researchers, there may be a risk of invasion of the privacy of the users constituting this social network.

The Preservation of Confidentiality for Published Data (PPDP) studies how to make data immune to privacy attacks.

Data anonymization is one of the solutions that can be used to preserve the confidentiality of data while ensuring its uses.

In this work, we describe a model-based approach that guides the data editor in their anonymization process, and algorithms that minimize the risk of re-identification while preserving the usefulness of the data

REMERCIEMENT:

Nous remercions tout d'abord le bon Dieu, le tout puissant de nous avoir armé de force et de courage pour mener à terme ce projet.

Merci à tous ceux qui ont contribué à ce que la réalisation de ce projet soit possible.

Notre profonde gratitude et mes sincères remerciements vont particulièrement à Dr.KABOU.Salaheddine, qui a accepté de diriger ce travail, Je les remercie pour la confiance et la compréhension qu'ils ont toujours manifestées à mon égard. Si ce travail est mené à terme, c'est grâce à leurs aides et soutien. Qu'ils soient assurés de ma profonde estime.

Enfin, nous ne pouvons pas s'empêcher de remercier tout le corps enseignant département Mathématique et Informatique de l'université Ahmed Draya ADRAR pour la qualité d'enseignement qu'il nous a offert et d'avoir bâti l'édifice intellectuel que nous sommes d'ores et déjà.

Dédicace

C'est grâce à Allah seul que j'ai pu achever ce travail.

Je le dédie à :

A tout ma famille BOUALLALA Allal.

Et à tous mes fidèles amis, à tous ceux que j'aime…

Bouallala Zineb

Dédicace

C'est grâce à Allah seul que j'ai pu achever ce travail.

Je dédie ce modeste travail

À mon père et ma mère qui se sont sacrifiés pour que je puisse achever mes études.

À mes chers frères et mes chères sœurs.

Ainsi que mes oncles, mes tantes et mes grands parents.

À toute la famille BENDRISSE, RAHMANE .

À ma binôme Zineb, et ma très chère amies Kebir Zohra &Siham frouhate.&Sabah & Zahra & Siham & Fatima &.

À la fill de ma tente Hadjer et À mon frére Bachir

À tous mes amies, à tous les enseignants qui m'ont aidé de prés ou de loin à obtenir mon Master en informatique.

BENDRISSE Samira.

Table de matières:

Chapitre II: Etat de l'art

Chapitre III: L'approche d'anonymisation

Chapitre IV: Implimentation et discussion

Liste des figures:

Liste des Tableaux:

Liste des abréviations:

PCDP	Préservation de la Confidentialité pour les Données Publiées
PPDP	Privacy Preserving Data Publishing
IE	Identifiants Explicites
QID	Quasi-Identifiants
AS	Attributs Sensibles
ANS	Attributs Non Sensibles
TDS	Top Down Specialization
IG	Information Gain
AL	Anonymity Loss
IGPL	Information Gained per each Loss of Privacy
ILPG	Information Loss per each Privacy Gain
DA	Distinctive Attribute
MD	Minimal Mistortion Metric
DM	Discernibility Metric

Introduction générale

Il est connu et reconnu que les données jouent un rôle important dans le développement de la science et de l'innovation. Elles sont aussi, pour les organismes publics et privés.

La confidentialité de ces données doit être préservée avant la publication , c'est à dire aucune information sensible ne doit être divulgués.

La vie privée sur internet est une notion plus importante que celle habituellement admise dans la vie de tous les jours. Il est primordiale de bien comprendre que toute information non sécurisée mise en ligne peut être accessible par tout le monde. Cette prise de conscience de l'universalité d'internet et de sa propension à diffuser rapidement une information important.

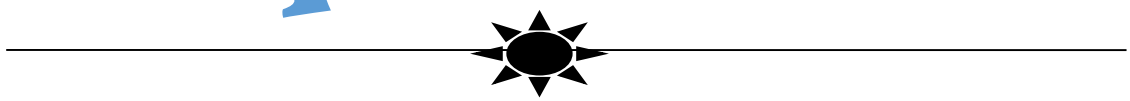
Anonymisation des données est l'une des techniques de la confidentialité qui se traduisent par la conservation de l'information, ce qui rend les données inutiles pour tout le monde sauf les propriétaires.

L'anonymisation touchent deux axes principaux :

- Préserver la confidentialité (anonymisation) des données et affiner la définition de l'anonymisation pour fournir des différentes garanties sur la sécurisation des données.
- Préserver l'utilité de données et minimiser la perte en ce qui concerne la qualité d'information après la publication tout en respectant la définition de l'anonymisation.

Dans cette étude, nous détaillons les caractéristiques de base de la préservation de la confidentialité pour les données publiées. Nous donnons quelques définitions sur tous ce qui concerne l'approche d'anonymisation.

Chapitre 1



Préservation de la confidentialité :Anonymisation

1. Introduction

La collecte d'informations numériques par les gouvernements, les entreprises et les particuliers a créé d'énormes possibilités de prise de décisions fondées sur les connaissances et l'information. Sous l'impulsion d'avantages mutuels ou de réglementations imposant la publication de certaines données, il existe une demande d'échange et de publication de données entre diverses parties. Cependant, les données sous leur forme d'origine contiennent généralement des informations sensibles sur les individus, et la publication de ces données violera la vie privée des individus**. [24]**

L'un de ses points clés était d'établir un système national de dossiers médicaux électroniques qui encourage le partage des connaissances médicales par le soutien d'une décision clinique assistée par l'ordinateur. Les données détaillées qui spécifient un individu contiennent souvent des informations sensibles, ce qui fait que la publication de ces données risque de violer sa confidentialité. La pratique actuelle repose sur le principe des politiques et des lignes directrices qui restreignent les types de données publiables et ainsi sur les accords et le stockage des données sensibles. La limitation de cette approche nécessite un niveau de confiance plus élevé dans la plus part des scénarios de partage de données. **[1]**

La tâche la plus importante est de développer des méthodes et des outils pour la publication de données de sorte que ces données doivent rester pratiquement utiles tout en préservant leurs confidentialités. Ce concept est appelé : PCDP, **la Préservation de la Confidentialité pour les Données Publiées** (Privacy preserving data publishing,).

2. La préservation de la confidentialité pour les données publiées

Le PCDP nettoie les données personnelles (par exemple les dossiers de santé électroniques) qui sont très susceptibles de les mettre à la disposition des agences ou du public. Comme le montre la figure I.2 ci-dessous, l'attaquant peut être toute personne (destinataire des données) qui obtient les informations personnelles sur un individu. Ainsi, il est vital pour l'éditeur de données d'appliquer diverses mesures de protection de la vie privée pour contrôler les informations des données publiées en les modifiant avant leur publication. **[27]**

Le terme information doit être pris au sens le plus large : il recouvre non seulement les données elles-mêmes, mais aussi les flux d'information et la connaissance de l'existence des données ou des communications. Assurer la confidentialité d'un système est donc une tâche complexe. Il faut analyser tous les chemins qu 'une

information particulière peut prendre dans le système pour s'assurer qu'ils sont sécurisés. Il importe également de prendre en compte les connaissances qu'un ou plusieurs utilisateurs peuvent déduire à partir des informations qu'ils acquièrent. Il faut donc contrôler non seulement les informations présentes dans le système, mais aussi les liens logiques qui peuvent les relier entre elles ou à des informations publiques**. [8]**

Les principaux aspects de la confidentialité sont les suivants :

2.1.L'Anonymat (Anonymity) : peut être définie comme la propriété garantissant qu'un utilisateur peut utiliser une ressource ou un service sans révéler son identité!; autrement dit, l'impossibilité (pour d'autres utilisateurs) de déterminer le véritable nom de cet utilisateur**. [2]**

2.2. La pseudonymat (Pseudonymity): est la suppression des champs qui identifiées les enregistrements, et remplacer ces champs dans chaque enregistrement par un nouveau champ, appelé pseudonyme, dont la caractéristique est qu'il doit rendre impossible tout lien entre cette nouvelle valeur et la personne réelle. Pour créer ce pseudonyme, on utilise souvent une fonction de hachage que l'on va appliquer à l'un des champs identifiants (par exemple le numéro de sécurité sociale), qui est un type de fonction particulier qui rend impossible (ou tout du moins extrêmement difficile) le fait de déduire la valeur initiale. On voit ainsi que deux entités possédant des informations sur une même personne, identifiée par son numéro de sécurité sociale, pourraient partager ces données de manière anonyme en hachant cet identifiant. Il est également possible d'utiliser tout simplement une fonction aléatoire pour générer un identifiant unique pour chaque personne, mais nous verrons plus bas que cela ne résout pas tous les problemes. **[3]**

2.3. La non-chaînabilité (Unlinkability): est l'incapacité de trouver un lien entre différents opérations réalisées par un même utilisateur! par exemple, en interdisant la fusion ou le croisement d'informations à partir de différents fichiers ou bases de données. **[2]**

2.4. La non-observabilité (Unobservability): consiste à ce que les utilisateurs ne puissent pas déterminer si une opération est en cours. La non-observabilité assure la protection de l'activité d'un utilisateur contre un tiers qui ne peut pas présumer qu'une ressource ou un service est utilisé. **[28]**

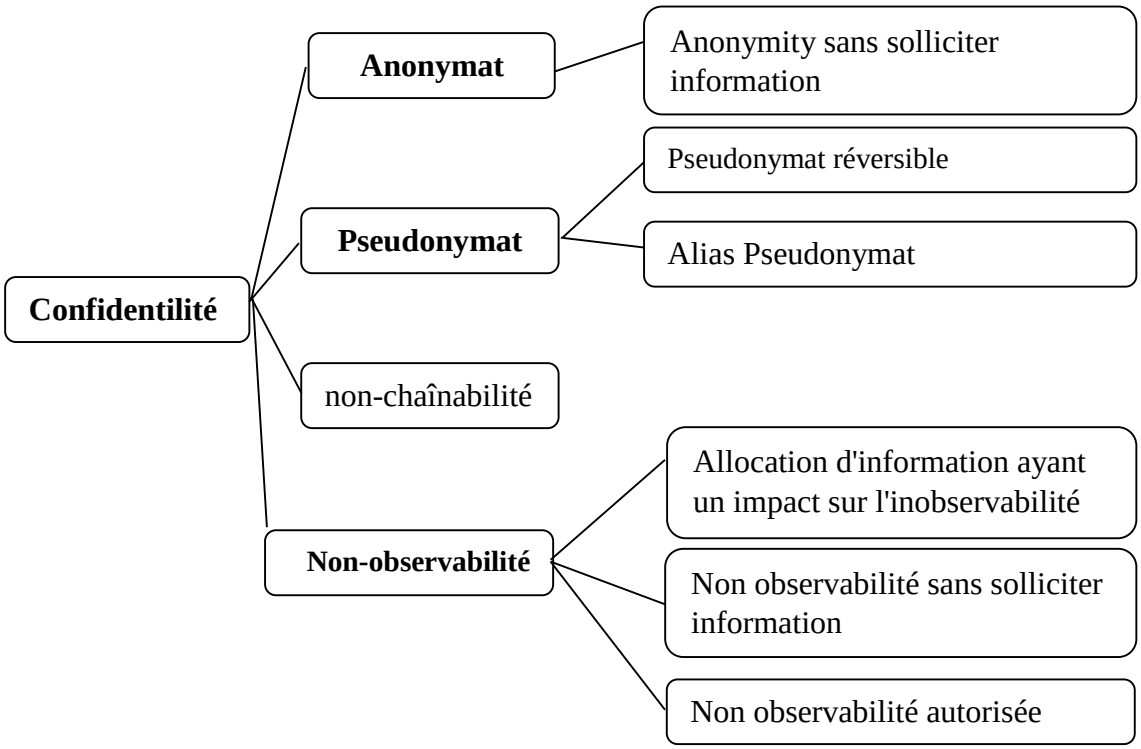
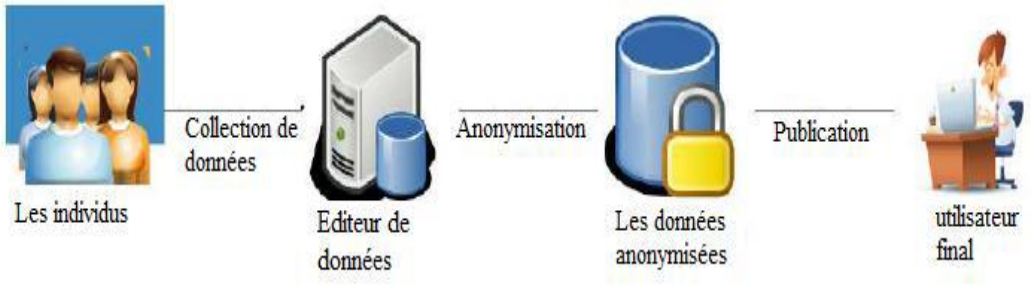


Figure I.1: Les aspects de la confidentialité. **[2]**

Préserver la confidentialité lors de la publication de données peut être représenté sous forme de phases comme le montre la **figure I.2**. la phase de « collecte de données » et la phase de « publication de données ». En phase de collecte, les données réelles sont rassemblées auprès des propriétaires par l'éditeur de données « data publisher » qui est un serveur de confiance. L'éditeur de données à son tour modifie les données en les anonymisant d'une manière qui garantit la confidentialité des informations personnelles, ensuite dans la phase de publication il les publie au destinataire des données qui peut être des data-miner et peut également être des adversaires. **[4]**



FigureI.2: Collection et publication de données. **[4]**

Il existe deux modèles d'éditeurs de données :

- **le modèle non confiant :** l'éditeur de données n'est pas confiant et peut identifier les informations sensibles des propriétaires de records. Plusieurs solutions des cryptographies,et des communications anonymes, sont proposées pour la collecte des records anonymes des propriétaires sans aucune révélation d'identités**. [1]**
- **Le modèle confiant :** l'éditeur de données est confiant et les propriétaires de records sont prêts à lui fournir leurs informations sensibles**. [1]**

3. Les attaques des données

De manière générale, une menace à la vie privée survient soit lorsqu'une identité est liée à un enregistrement ou lorsqu'une identité est liée à une valeur sur un attribut sensible. Ces menaces sont appelé couplage d'enregistrements et couplage d'attributs. Ci-dessous, nous supposons que l'attaquant connaît la quasi-identifier X d'un détenteur d'enregistrement cible. **[29]**

3.1. L'attaque par "Record linkage": Dans l'attaque par "Record linkage", une certaine valeur x sur un quasi-identificateur X identifie un petit nombre d'enregistrements dans le tableau validé T. Dans ce cas, le détenteur d'enregistrement ayant la valeur x est vulnérable à un lien avec un petit nombre d'enregistrements dans T. **[29]**

3.2. L'attaque par "Attribute linkage": Dans l'attaque par "Attribute linkage", l'attaquant peut ne pas identifier précisément l'enregistrement de la victime cible, mais pourrait déduire ses valeurs sensibles des données publiées T, sur la base de l'ensemble des valeurs sensibles associées au groupe auquel appartient la victime. Si certaines

valeurs sensibles prédominent dans un groupe, une inférence réussie devient relativement facile même si k-anonymat est satisfait**. [24]**

3.3. L'attaque par "Table linkage": Dans l'attaque par "Table linkage", l'attaquant est capable de lier le propriétaire de l'enregistrement cible au document publié ou la table publiée elle-même. Dans ce type d'attaque, l'attaquant trouve simplement les propriétaires des enregistrements présence absence. **[33]**

3.4. L'attaque probabiliste: Il ya une autre famille d'attaque sur la vie privée qui ne se concentre pas sur les enregistrements, les attributs ou bien les tables, mais l'attaquant dans ce modèle peut créer un lien vers une victime s'il peut changer son croyance probabiliste sur les informations sensibles de la victime après avoir accéder aux les données publiées. **[13]**

4. L'approche : Anonymisation

Cette recherche est basée sur la publication spécifique de données préservant la confidentialité (PCDP) approche connue sous le nom d'anonymisation qui vise à supprimer l'association entre les informations personnelles identifiables et la personne individuelle. L'anonymat approche de modification modifie les données afin de rendre impossible la mise en relation des leurs données. L'approche vise à protéger l'identité et / ou les données sensibles des personnes concernées lorsque les données sont partagées à des fins différentes**. [30]**

L'anonymisation s'appuie sur les quatre groupes distincts d'attributs que peut contenir une table relationnelle concernant un individu :

4.1. Les identifiants explicites (IE) : attribut ou ensemble d'attributs qui désignent directement l'individu (numéro de sécurité sociale, nom, prénom…). **[5]**

4.2. Les quasi-identifiants (QID) : Ce sont les attributs qui ne peuvent pas identifier directement un individu mais s'ils sont liés à des données accessibles au public, ils peuvent identifier facilement un individu. Par exemple code postal, âge, sexe, etc. Une classe d'équivalence est un ensemble d'enregistrements qui ont la même valeur pour tous les attributs quasi-identifiants. **[27]**

4.3. Les attributs sensibles (AS) : Ce sont les attributs qu'un individu veut cacher aux autres. Par exemple maladie. **[27]**

4.4. Les attributs non sensibles (ANS) : attribut n'appartenant à aucune des trois catégories précédentes**. [5]**

Une source de données comprend des attributs publics et privés. Nous devons révéler au public attributs au maximum sans révéler les attributs privés. Publication des données soit pseudonyme ou anonyme peut toujours être violé de la vie privée avec inférence. **[6]**

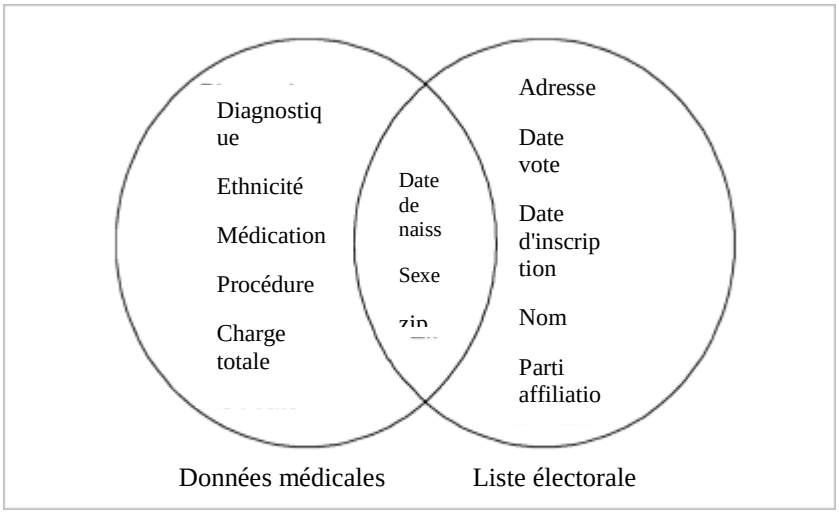


Figure I.3: La ré-identification des propriétaires par la liaison**. [6]**

Un exemple donné par Sweeney est reproduit à la figure I.3. Un le nom de la personne sur une liste électorale publique était lié à son dossier dans une base de données médicales grâce à la combinaison du code postal, de la date de naissance et du sexe. Chaque de ces attributs n'identifie pas de manière unique un propriétaire d'enregistrement, mais leur combinaison, appelé quasi-identificateur (QID), distingue souvent un unique ou un petit nombre de propriétaires de documents. Il est rapporté que 87% de la population américaine ressemblait à caractéristiques qui les ont probablement rendus uniques grâce à ce qui précède quasi-identifiants**. [6]**

5. L'architecteur de l'anonymisation

5.1. Anonymisation de connexion

Elle vise à protéger l'identité de la source et de la destination quand il y'a une communication. Dans ce terrain, la plupart des travaux dans l'anonymisation de connexion touchent la notion d'authentification anonyme. **[1]**

❖ L'authentification anonyme

L'Authentification Anonyme semble être un oxymore car l'Authentification représente le moyen de prouver l'identité de quelqu'un vis-à-vis d'une autre partie, tandis que le but de l'Anonymat est de cacher l'identité de quelqu'un .

Cependant, l'authentification anonyme réalise ces objectifs contradictoires en demandant par exemple aux utilisateurs de prouver uniquement l'appartenance à un

groupe de telle sorte que l'identité de l'utilisateur ne peut pas être déterminée par le Vérificateur : Partie qui réalise la vérification. **[28]**

❖ **Lien sémantique**

Il est la relation sémantique qui se trouve entre le fournisseur et le client. **[9]**

5.2. Anonymisation des données

5.2.1. Anonymisation des données statiques

Il y a plusieurs travaux d'anonymisation des données basées sur des ensembles de données statiques.

Centralisé : elle met l'accent à des bases des données centralisé

Décentralisé : concerne tous les travaux qui touchent les bases des données Décentralisé. **[1]**

5.2.2. Anonymisation des données dynamiques

La publication en série pour les bases de données dynamiques est nécessaire chaque fois qu'il y a des insertions, des suppressions et des mises à jour dans jeux de données.**[7]**

Mise à jour externe

Pour chaque entier i et j ($0 \leq i < j$), si l'enregistrement t ($t \neq \phi$) satisfit l'une des conditions suivantes :

1) $t_i \in T_i$ and $t_i \ni t_j$

2) $t_i \ni T_i$ and $t_i \in t_j$

Le mise à jour externe contient deux type spécifiques : l'insertion et la suppression qui correspondent aux conditions 1 et 2 respectivement .

donc t est une mise à jour externe de T_j contrairement à Ti , si t satisfit l'une des conditions. **[1]**

Mise à jour interne

Pour chaque entier i et j ($0 \leq i < j$), supposant pour un enregistrement t que $t_i \in T_i$ et $t_j \in T_j$. Si t_i et t_j satisfont au moins l'une des conditions suivantes:

1) $t_i[Q] \neq t_j[Q]$. { Q et S sont des moments différents }

2) $t_i[S] \neq t_j[S]$.

Alors, on dit qu'il y a une mise à jour interne sur t dans la période[i, j].**[1]**

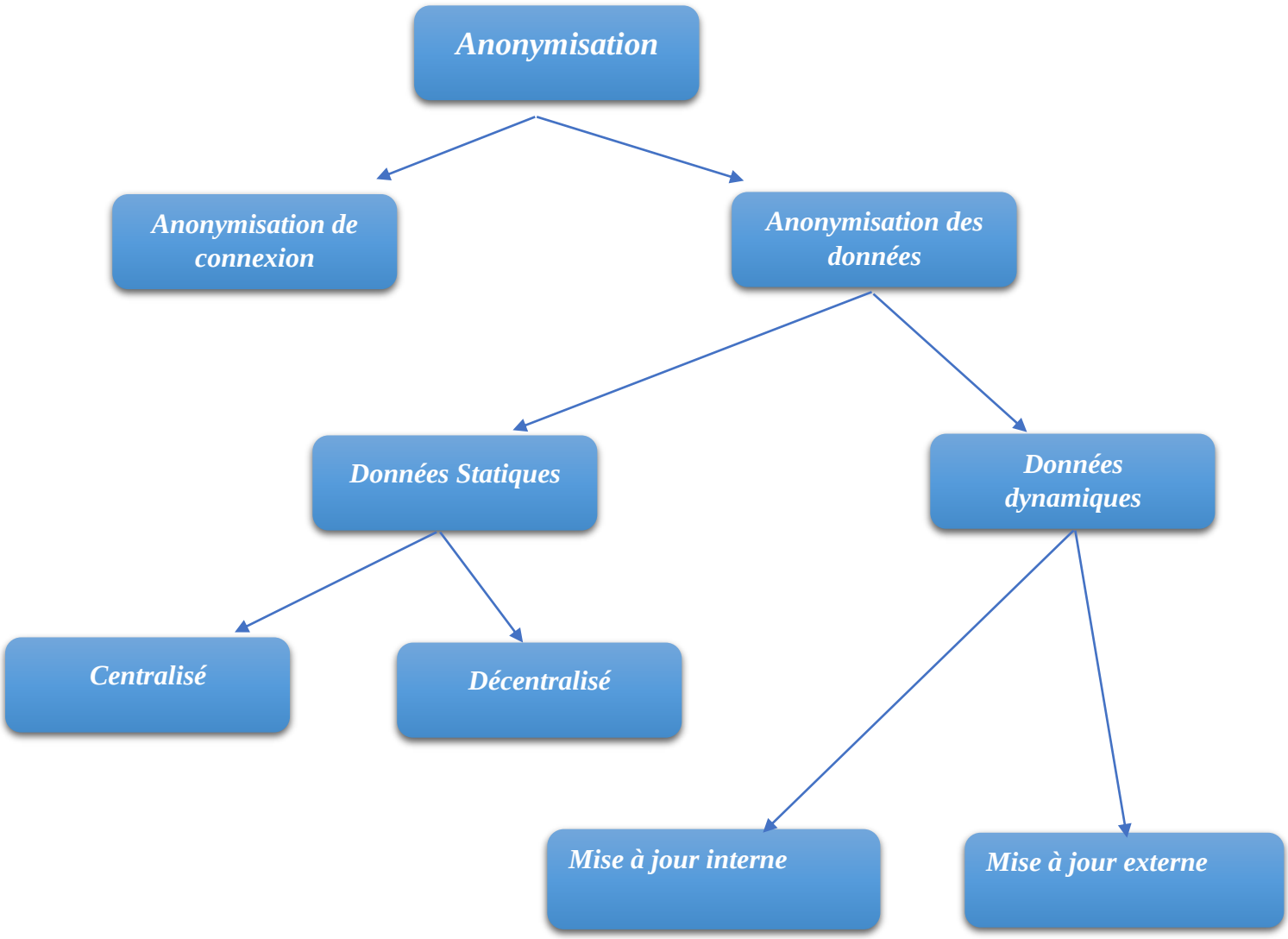


Figure I.4: L'architecteur de l'anonymisation

6. Les opérations d'anonymisation

Pour la préservation de confidentialité des données, il faut faire plusieurs modifications sur la table d'origine avant d'être publié, ces modification se fait à travers des séquences d'opérations d'anonymisation qui sont: la généralisation, la suppression et la permutation.

6.1. Généralisation

Généralisation (également appelée recodage) consiste à remplacer les valeurs d'un attribut par des valeurs moins spécifiques mais cohérentes, employant souvent une hiérarchie de généralisation des valeurs , comme celles illustrées à la **figure I.5** . Les

valeurs au niveau le plus bas (à droite) sont dans le domaine fondamental de l'attribut, qui correspondent aux valeurs les plus spécifiques (valeurs d'origine). **[11]**

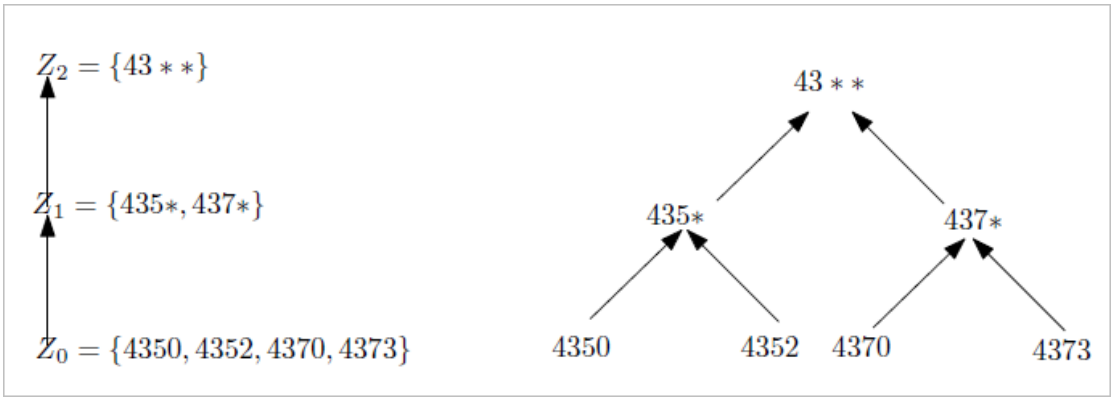


Figure I.5 : Hiérarchie de généralisation de l'attribut code zip. **[32]**

Une fois la hiérarchie définie, la technique de généralisation consiste à remplacer dans la table anonyme, une valeur d'un attribut du QID de la table originale par un de ses ancêtres dans la hiérarchie de généralisation. Le niveau de généralisation appliqué peut être différent pour chacun des attributs du QID. A titre d'exemple, si l'on considère que l'âge et la villes ont des attributs du QID, on pourrait, par le procédé de généralisation, obtenir le **Tableau I.2**.

Nim tuple	Sexe	Ville	Profession	Statut marital	Age
1	F	Paris	Ingénieur	Mariée	33
2	F	Pontoise	Ingénieur	Veuve	65
3	F	Paris	étudiant	Célibataire	21
4	M	Nice	étudiant	Célibataire	19
5	M	Paris	Statisticien	Marié	36
6	F	Cannes	étudiant	Mariée	28
7	F	Pontoise	Statisticien	Divorcé	46
8	M	Pontoise	Statisticien	Divorcé	81
9	M	Cannes	Statisticien	Marié	58
10	F	Nice	Professeur	Veuve	63
11	M	Paris	étudiant	Célibataire	19

Tableau I.1. Table originale avant anonymisation**. [9]**

Nim tuple	Sexe	Ville	Profession	Statut marital	Age
1	**F**	**Ile de France**	**Ingénieur**	**Mariée**	**[19,81]**
2	**F**	**Ile de France**	**Ingénieur**	**Veuve**	**[19,81]**
3	**F**	**Ile de France**	**étudiant**	**Célibataire**	**[19,81]**
4	**M**	**Région PACA**	**étudiant**	**Célibataire**	**[19,81]**
5	**M**	**Ile de France**	**Statisticien**	**Marié**	**[19,81]**
6	**F**	**Région PACA**	**étudiant**	**Mariée**	**[19,81]**
7	**F**	**Ile de France**	**Statisticien**	**Divorcée**	**[19,81]**
8	**M**	**Ile de France**	**Statisticien**	**Divorcé**	**[19,81]**
9	**M**	**Région PACA**	**Statisticien**	**Marié**	**[19,81]**
10	**F**	**Région PACA**	**Professeur**	**Veuve**	**[19,81]**
11	**M**	**Ile de France**	**étudiant**	**Célibataire**	**[19,81]**

Tableau I.2. Application de la technique de généralisation aux attributs ville et âge. **[9]**

6.2. La suppression

La suppression consiste à empêcher les données sensibles par l'enlever. La suppression peut être appliquée au niveau de cellule unique, tuple entier ou colonne entière, permet de réduire le degré de généralisation à imposer pour atteindre la k-anonymité. **[31]**

A titre d'exemple, supposons que, dans le **Tableau I.1**, les deux derniers enregistrements présentent un risque de ré-identification. Pour éviter cette dernière, la solution d'anonymisation proposée est la suppression des valeurs des attributs « sexe » et « profession » respectivement pour ces deux enregistrements (voir **Tableau I.3**). Le choix des valeurs à supprimer se fonde sur un calcul qui vise à diminuer le nombre de suppressions locales. Notons qu'une combinaison de ces deux modes de suppression est possible. **[9]**

Nim tuple	Sexe	Ville	Profession	Statut marital	Age
1	**F**	**Paris**	**Ingénieur**	**Mariée**	**33**
2	**F**	**Pontoise**	**Ingénieur**	**Veuve**	**65**
3	**F**	**Paris**	**étudiant**	**Célibataire**	**21**
4	**M**	**Nice**	**étudiant**	**Célibataire**	**19**
5	**M**	**Paris**	**Statisticien**	**Marié**	**36**
6	**F**	**Cannes**	**étudiant**	**Mariée**	**28**
7	**F**	**Pontoise**	**Statisticien**	**Divorcée**	**46**
8	**M**	**Pontoise**	**Statisticien**	**Divorcé**	**81**
9	**M**	**Cannes**	**Statisticien**	**Marié**	**58**
10	***	**Nice**	**Professeur**	**Veuve**	**63**
11	**M**	**Paris**	***	**Célibataire**	**19**

Tableau I.3. Application de la suppression locale au **Tableau I.1. [9]**

6.3. La permutation ou technique de “Swapping”

la permutation dissocie la corrélation entre quasi-identifiants et attributs sensibles en regroupant et en mélangeant les éléments sensibles valeurs dans un groupe quasi-identifiant. Contrairement à la généralisation et à la suppression, l'anatomisation ne modifie pas le quasi-identifiant ou l'attribut sensible, mais dissocie la relation entre les deux. L'idée de la permutation est de dissocier la relation entre un quasi-identifiant et un numérique attribut sensible en partitionnant un ensemble d'enregistrements de données en groupes et mélanger leurs valeurs sensibles au sein de chaque groupe. **[6]** A titre d’exemple, la permutation appliquée sur l’attribut Profession au sein du sous-ensemble constitué des tuples 3 et 5 donnerait la table anonyme suivante **(Tableau I.4). [9]**

Nim tuple	Sexe	Ville	Profession	Statut marital	Age
1	**F**	**Paris**	**Ingénieur**	**Mariée**	**[19,81]**
2	**F**	**Pontoise**	**Ingénieur**	**Veuve**	**[19,81]**
3	**F**	**Paris**	**Statisticien**	**Célibataire**	**[19,81]**
4	**M**	**Nice**	**étudiant**	**Célibataire**	**[19,81]**
5	**M**	**Paris**	**étudiant**	**Marié**	**[19,81]**
6	**F**	**Cannes**	**étudiant**	**Mariée**	**[19,81]**
7	**F**	**Pontoise**	**Statisticien**	**Divorcée**	**[19,81]**
8	**M**	**Pontoise**	**Statisticien**	**Divorcé**	**[19,81]**
9	**M**	**Cannes**	**Statisticien**	**Marié**	**[19,81]**
10	**F**	**Nice**	**Professeur**	**Veuve**	**[19,81]**
11	**M**	**Paris**	**étudiant**	**Célibataire**	**[19,81]**

Tableau I.4. Application du « data swapping » à l'attribut Profession**. [9]**

7. Comparaison entre le cryptage (chiffrement) et l'anonymisation

Le chiffrement, répond à un besoin de dissimulation d'informations, de données sensibles ou de données personnelles, aux utilisateurs qui ne sont pas habilité à les voir. Il permet de rendre les informations totalement incompréhensibles afin d'engarder la confidentialité. Le chiffrement est un processus réversible qui ne fait que masquer les données. Il est donc toujours possible de retrouver leur valeur initiale grâce à une clé. Cette clé, qui est un algorithme de déchiffrement, va permettre de verrouiller et déverrouiller le chiffrage des informations. **[26]**

L'anonymisation a pour but de changer les données tout en les gardant confidentielles. L'anonymisation permet de garderintacte la cohérence des informations pour qu'elles puissent être interprétées. Cependant, l'anonymisation est un processus irréversible. Une fois la donnée changée, il est impossible de la retrouver à son état d'origine. L'anonymisation permet de cacher les données stockées dans des environnements de production, test ou développement. **[26]**

8. Conclusion

Nous avons détaillé dans ce chapitre les caractéristiques de la préservation de la confidentialité pour les données publiées. En commençant par une définition des données à caractère personnelle. Ensuite, nous avons présenté l'approche d'anonymisation. Enfin, nous avons énuméré les principales opérations de l'anonymisation.

Dans le chapitre suivant nous allons présenté un état de l'art sur la préservation de la confidentialité pour les données publiées.

Chapitre II

Etat de l'art

1.Introduction

Dans ce chapitre, nous avons discuté de diverses techniques telles que ***k-anonymat, l-diversité, k-concealment et* δ-Présence** qui ont été suggérées pour trouver le bon équilibre entre la publication des données et la divulgation de données

Plusieurs algorithmes ont été proposés pour l'anonymisation des données personnelles, permettant de minimiser le risque de ré-identification tout en préservant l'utilité des données. Dans cet chapitre, nous décrivons une approche fondée sur les modèles qui guide l'éditeur de données dans son processus d'anonymisation. Le guidage, informatif ou suggestif, permet non seulement de choisir l'algorithme le plus pertinent, mais aussi de paramétrer cet algorithme en tenant compte des caractéristiques des données et du contexte. Dans cet article, nous nous intéressons aux algorithmes de généralisation de micro-données. Un processus de rétro-ingénierie des outils existants a permis d'extraire certaines connaissances. Nous les stockons avec toutes les connaissances liées à l'anonymisation, tant théoriques qu'expérimentales, dans une ontologie

2. La préservation de la confidentialité pour les bases de données centralis

La préservation de la vie privée est importante dans la publication de données. L'objectif est de publier une version anonymisée des données appartenant à une organisation comme un hôpital, une agence gouvernementale, ou une société d'assurance, sans violation de la vie privée ou divulgation des informations personnelles. **[4]**

Dans les années récentes, la préservation de la confidentialité des données publiées pour les bases des données centralisées a été largement étudiée, il y a des techniques telles que ***k-anonymat, l-diversité, k-concealment et* δ-Présence** qui ont été suggérées pour trouver le bon équilibre entre la publication des données et la divulgation de données. **[1]**

2.1. Anonymisation statique

2.1.1. K-anonymat

Le k-anonymat est une technique d'anonymisation qui utilise les opérations de généralisation et de suppression. Son objectif est de ne publier des informations que s'il y a au moins k individus dans chaque groupe de données généralisées. **[13]**

Le modèle k-anonymat tient compte d'une organisation des données en table. Chaque table étant composée de lignes d'information comportant des attributs dont les valeurs proviennent de différents domaines. La première opération à réaliser consiste à retirer tous les attributs tels que le nom ou le numéro de patient.

Le quasi-identificateur d'une table T, dénoté « QIT », est un ensemble d'attributs de T qui, si utilisés conjointement, peuvent mener à l'identification d'un individu avec une probabilité égale à 1. L'objectif principal de la méthode k-anonymat est de transformer une table de manière à ce que personne ne puisse établir de lien entre la table T et un individu avec une probabilité inférieure à 1/k. On dit qu'une table « T » est k-anonyme en rapport avec un quasi-identificateur « QIT » si et seulement si, pour tout enregistrement « r » dans « T », il existe au moins (k − 1) autres enregistrements dans T qui ne peuvent être distingués de « r » par rapport à « QIT ». **[14]**

Exemple

	Non-sensibles			sensible
	Zip	**âge**	**nationalité**	**état**
1	13053	28	Russie	Heart Disease
2	13068	29	Américaine	Heart Disease
3	13068	21	Japonais	Infection virale
4	13053	23	Américaine	Infection virale
5	14853	50	Indian	Cancer
6	14853	55	Russie	Heart Disease
7	14850	47	Américaine	Infection virale
8	14850	49	Américaine	Infection virale
9	13053	31	Américaine	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japonais	Cancer
12	13068	35	Américaine	Cancer

Table II.1. Tableau initial des données **[15]**

	Non-sensibles			sensible
	Zip	**âge**	**nationalité**	**état**
1	130**	<30	*	Heart Disease
2	130**	<30	*	Heart Disease
3	130**	<30	*	Infection virale
4	130**	<30	*	Infection virale
5	1485*	≥40	*	Cancer
6	1485*	≥40	*	Heart Disease
7	1485*	≥40	*	Infection virale
8	1485*	≥40	*	Infection virale
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Table II.2. Tableau de 4-anonyme **[15]**

La **Table II.1** présente les enregistrements médicaux d'un hôpital. On remarque que la table ne contient pas d'attributs identifiant comme le nom, le numéro de sécurité sociale, etc..

Dans cet exemple, on divise les attributs en deux groupes : les attributs sensibles (la situation médicale) et les attributs non sensibles (code postal, l'âge et la nationalité). Un attribut est marqués sensible si l'adversaire ne doit pas être autorisé à découvrir la valeur de cet attribut pour toute personne dans l'ensemble de données. Les attributs ne sont pas marqués sensibles sont non-sensitive. La collection d'attributs {code postal, l'âge, la nationalité} est un quasi-identifiant de cet ensemble de données.

La **Table II.2** montre un tableau 4 anonyme dérivé du la **Table II.4** ("*" représente une valeur supprimée, par exemple, "code zip = 1485 *" signifie que le code est dans l'intervalle [14850-14859] et «âge = 3 *" signifie que l'âge est compris entre [30 − 39]) **[16].**

2.1.2 L-diversité

Supposons que vous ayez un groupe de k enregistrements différents qui partagent tous un quasi-identificateur particulier. C'est bien, car un attaquant ne peut pas identifier l'individu sur la base du quasi-identifiant. Mais que se passe-t-il si la valeur qui les intéresse (par exemple, le diagnostic médical de l'individu) est la même pour chaque valeur du groupe. La distribution des valeurs cibles au sein d'un groupe est appelée «l-diversité». **[8]** Actuellement, il existe deux grandes catégories de techniques de l-diversité: la généralisation et la permutation. Une méthode de généralisation existante diviserait les données en groupes de transactions disjoints, de telle sorte que chaque groupe contienne suffisamment d'enregistrements avec des éléments sensibles bien distincts**.[17]**

- **Exemple**

Le tableau **II.3** est le tableau d'origine et le tableau **II.4** montre une version anonymisée satisfaisant distincte et entropique 3- la diversité. Il existe deux attributs sensibles: le salaire et la maladie.

Supposons que l'on sache que l'enregistrement de Bob correspond à l'un des trois premiers enregistrements, alors on sait que Bob est le salaire est dans la gamme [3K – 5K] et peut inférer que Bob est le salaire est relativement bas. Cette attaque ne concerne pas seulement des attributs numériques comme «Salaire», mais aussi des

attributs catégoriels comme «Maladie». Sachant que le record de Bob appartient à la première classe d'équivalence permet de conclure que Bob a des problèmes d'estomac, car les trois les maladies de la classe sont liées à l'estomac**. [18]**

	Code ZIP	Age	Salaire	Maladie
1	47677	29	3K	Ulcère gastrique
2	47602	22	4K	gastrite
3	47678	27	5K	Cancer de l'estomac
4	47905	43	6K	gastrite
5	47909	52	11K	grippe
6	47906	47	8K	bronchite
7	47605	30	7K	bronchite
8	47673	36	9K	pneumonie
9	47607	32	10K	Cancer de l'estomac

Tableau II.3. tableau d'origine **[18]**

	Code ZIP	Age	Salaire	Maladie
1	476**	2*	3K	Ulcère gastrique
2	476**	2*	4K	Gastrite
3	476**	2*	5K	Cancer de l'estomac
4	4790*	≥40	6K	Gastrite
5	4790*	≥40	11K	Grippe
6	4790*	≥40	8K	bronchite
7	476**	3*	7K	Bronchite
8	476**	3*	9K	Pneumonie
9	476**	3*	10K	Cancer de l'estomac

Tableau II.4. Une version 3 diversitiée du tableau II.3 **[18]**

2.1.3. t- closeness

La proximité t d'une classe d'équivalence est atteinte lorsque la distance d'attribut sensible dans cette classe n'est pas supérieure au seuil, t avec la distance d'attribut dans l'ensemble du tableau. La table est reconnue comme ayant une proximité t si toutes les classes d'équivalence ont une proximité t **[19].**

	Code ZIP	Age	Salaire	Maladie
1	4767*	≤40	3K	Ulcère gastrique
2	4767*	≤40	4K	Gastrite
3	4767*	≤40	5K	Cancer de l'estomac
4	4790*	≥40	6K	Gastrite
5	4790*	≥40	11K	Grippe
6	4790*	≥40	8K	bronchite
7	4760*	≤40	7K	Bronchite
8	4760*	≤40	9K	Pneumonie
9	4760*	≤40	10K	Cancer de l'estomac

Tableau II.5. Tableau qui a une proximité de 0,167 w.r.t. Salaire et proximité de 0,278 avec t. Maladie **[18]**

2.1.4.δ-Présence

C'est la probabilité d'inférer la présence de tout dossier de victime potentielle dans une plage spécifiée δ = (δ min, δ max). Formellement, étant donné une table publique externe E et une table privée T, où T ⊆ E, une table généralisée T satisfait (δ min, δ max) -présence si δ min ≤ P (t ∈ T | T) ≤ δ max pour tout t ∈ E. La présence de δ peut indirectement empêcher les couplages d'enregistrements et d'attributs parce que si l'adversaire a au plus δ% de confiance que le dossier de la victime cible est présent dans le tableau publié, alors la probabilité d'un couplage réussi avec son enregistrement et l'attribut sensible est au plus δ%. Bien que la δ-présence soit un modèle de confidentialité relativement «sûr», elle suppose que le détenteur de données a accès à la même table externe E que l'adversaire. Cela peut ne pas être une hypothèse pratique dans certaines situations **[20]**

❖ **Exemple [13]**

Les tableaux suivant représente un exemple sur le risque de la vie privée dans δ-présence où l'adversaire sait E et veut identifier les tuples dans les données privé T.
Les attributs dans le tableau privé (**Table II.7**) est un sous-ensemble de celui de l'ensemble de données dans le tableau publique. L'attribut "sen" ne fait pas partie du la table publique (**Table II.6**), mais précise qui tuples sont dans l'ensemble de données privées (**Table II.7**).

	Nom	Zip	Age	Nationalité	Sen
A	Alice	47906	35	USA	0
B	Bob	47903	59	Canada	1
C	Christine	47906	42	USA	1
D	Dirk	47630	18	Brazil	0
E	Eunice	47630	22	Brazil	0
F	Frank	47633	63	Peru	1
G	Gail	48973	33	Spain	0
H	Harry	48972	47	Bulgaria	1
I	Iris	48970	52	France	1

TableII.6. Tableau publique E **[13]**

	Zip	Age	Nationalité
B	47903	59	Canada
C	47906	42	USA
F	47633	63	Peru
H	48972	47	Bulgaria
I	48970	52	France

TableII.7. Tableau privée T. **[13]**

E* est la généralisation de E (Table II.8) et T* est (0.5, 0. 66)- présence généralisation de T (Table II.9.**)** Les deux généralisations ont la même cartographie de généralisation.

	Zip	Age	Nationalité	Sen
A	47*	*	USA	0
B	47*	*	Canada	1
C	47*	*	USA	1
D	47*	*	Brazil	0
E	47*	*	Brazil	0
F	47*	*	Peru	1
G	48*	*	Spain	0
H	48*	*	Bulgaria	1
I	48*	*	France	1

TableII.8. Tableau publique E*. **[13]**

	Zip	Age	Nationalité
B	47*	*	Canada
C	47*	*	USA
F	47*	*	Peru
H	48*	*	Bulgaria
I	48*	*	France

TableII.9. Tableau privée T*. **[13]**

3. Les algorithmes de K-anonymisation

La généralisation de micro-données est mise en oeuvre *via* plusieurs algorithmes dont les plus connus sont : µ-Argus (Burton *et al.*, 1997), Datafly (Sweeney 1997), l'algorithme de Samarati (Samarati 2001), Incognito (LeFevre *et al*., 2005).

3.1.Algorithmes d'anonymisation optimaux

La première famille trouve une k anonymisation optimale, pour une métrique de données donnée, en se limitant à la généralisation du domaine complet et à la suppression des enregistrements. Étant donné que l'espace de recherche pour le schéma de généralisation de domaine complet est beaucoup plus petit que les autres schémas, il est possible de trouver une solution optimale pour les petits ensembles de données. Ce type de recherche exhaustive, cependant, n'est pas extensible à de grands ensembles de données, surtout si un schéma d'anonymisation plus flexible est utilisé. **[24]**

3.1.1.L'algorithme MinGen de Sweeney [2002b]

L'algorithme MinGen de Sweeney [2002b] examine de manière exhaustive toutes les généralisations potentielles dans le domaine complet pour identifier la généralisation optimale mesurée dans la DM. Sweeney a reconnu que cette recherche exhaustive n'est pas pratique même pour les ensembles de données de taille modeste, motivant la deuxième famille d'algorithmes d'anonymisation k pour plus tard discussion. Samarati [2001] a proposé un algorithme de recherche binaire qui identifie d'abord toutes les généralisations minimales, puis trouve la généralisation optimale mesurée en MD. L'énumération de toutes les généralisations minimales est une opération coûteuse, et donc non évolutive pour les grands ensembles de données. **[24]**

3.2.Algorithmes d'anonymisation minimale

La deuxième famille d'algorithmes produit une table k-anonyme minimale en utilisant une recherche gourmande guidée par une métrique de recherche. Étant de

nature heuristique, ces algorithmes trouvent une solution minimalement anonyme, mais sont plus évolutifs que la famille précédente. **[24]**

Afin d'illustrer les quatres algorithmes, nous utilisons une table originale de laquelle a été supprimé préalablement l'identifiant des individus. Hormis l'identifiant, cette table est constituée de trois attributs sexe, code postal et niveau d'étude formant le quasi-identifiant (QI) et d'un attribut sensible appelé Salaire **Tableau II.10**. Chaque attribut du QI possède une hiérarchie de généralisation. La **Figure II.1**, la **Figure II.2**et la **Figure II.3** représentent respectivement la hiérarchie de généralisation de l'attribut sexe, du code postal et du niveau d'étude. Dans ces figures, les niveaux de généralisation ont été identifiés par la première lettre de l'attribut correspondant à la généralisation et par un nombre mentionnant la position du niveau dans la hiérarchie. A titre d'exemple, pour la hiérarchie de la **Figure II.2**, la valeur « 1305* » est au niveau « Z1 » de cette hiérarchie. Aussi, la valeur « Seconde » se trouvant dans la hiérarchie de la **Figure II.3** est au niveau « E0 » de cette hiérarchie. Une description détaillée des déroulements des neuf algorithmes sur la table originale **Tableau II.7**

Quasi Identifiant			Attribut sensible
Sexe	**Code postal**	**Niveau d'étude**	**Salaire**
M	13050	$5^{ième}$	1200
F	13051	$3^{ième}$	1300
M	13050	Seconde	1200
M	13050	Seconde	1300
M	13051	1^{ier} et $2^{ième}$ cycle	1500
F	13050	1^{ier} et $2^{ième}$ cycle	1500
F	13061	1^{ier} et $2^{ième}$ cycle	1600
F	13061	Master	2000
F	13060	Master	2100
M	13061	Doctorat	3000
M	13060	Doctorat	4000
M	13061	Doctorat	4500

Tableau II.10. Table originale . **[9]**

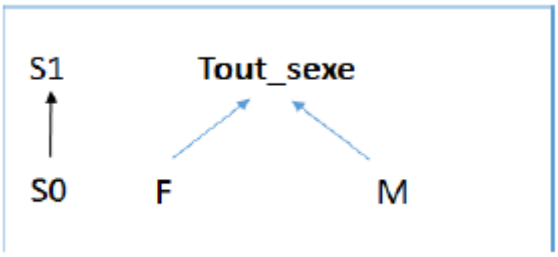


Figure II.1: La hiérarchie de généralisation de l'attribut sexe **[9]**

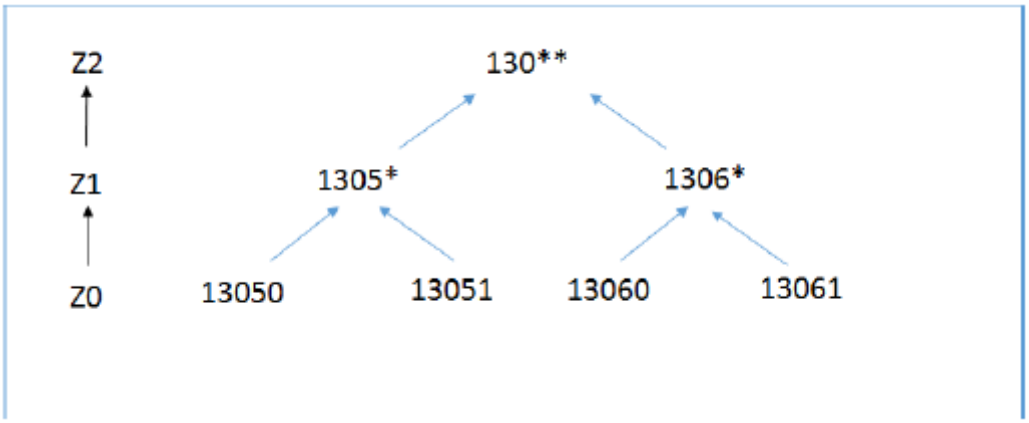


Figure II.2: La hiérarchie de généralisation de l'attribut code postal **[9]**

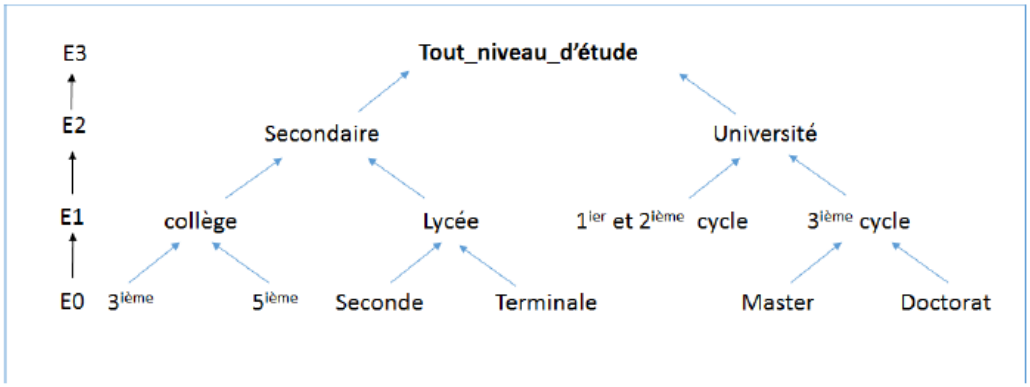


Figure II.3: La hiérarchie de généralisation de l'attribut niveau d'étude **[9]**

3.2.1. L'algorithme µ-Argus (Burton et al. 1997)

µ-Argus propose une exécution itérative du processus d'anonymisation. A chaque itération, (a) l'utilisateur choisit l'attribut à généraliser, (b) µ-Argus remplace chaque valeur de cet attribut par la valeur de son parent dans la hiérarchie de généralisation correspondante, (c) vérifie la satisfaction du k-anonymat et rend compte à l'utilisateur qui peut choisir entre la poursuite du processus ou encore la suppression locale de données (remplacement de valeurs par la valeur nulle).

La **Figure II.4** représente sa description dans (Sweeney 2002a). Dans cette figure, PT est la table à anonymiser, k est la contrainte de k-anonymat, DGHAi est la hiérarchie de généralisation de l'attribut Ai et MT la table anonyme résultante. µ-Argus construit, à partir de l'ensemble QI des attributs du quasi-identifiant, une partition en trois ensembles notés : Identifying, More et Most. L'utilisateur doit affecter, à chaque attribut, une valeur comprise entre 0 et 3. Celles-ci correspondent à "not identifying,", "identifying", "most identifying," et "more identifying" traduisant leur rôle dans la ré-identification d'individus. Les enregistrements qui ne satisfont pas le k-anonymat sont stockés dans une liste nommée "outliers" dans cette figure. Notons que µ-Argus ne teste pas toutes les combinaisons des attributs du quasi-identifiant. En effet, comme le montre l'étape 3 de cet algorithme, seules le sont les combinaisons de deux ou trois attributs. Ces combinaisons doivent contenir au moins un attribut "identifying".**[10]**

Input: Private Table **PT**; quasi-identifier QI = $(A_1, \ldots, A_n)$,
disjoint subsets of QI known as *Identifying*, *More*, and
Most where QI = *Identifying* ∪ *More* ∪ *Most*, *k* constraint;
domain generalization hierarchies DGH_{Ai}, where i=1,…,n.
Output: MT containing a generalization of PT[QI]
Assumes: |PT| ≥ *k*
Method:

1. freq ← a frequency list containing distinct sequences of values of PT[QI], along with the number of occurrences of each sequence.
2. Generalize each $A_i \in$ QI in freq until its assigned values satisfy *k*.
3. Test 2- and 3- combinations of *Identifying*, *More* and *Most* and **let** *outliers* store those cell combinations not having *k* occurrences.
4. Data holder decides whether to generalize an $A_j \in$ QI based on *outliers* and if so, identifies the A_j to generalize. freq contains the generalized result.
5. **Repeat** steps 3 and 4 until the data holder no longer elects to generalize.
6. Automatically suppress a value having a combination in *outliers*, where precedence is given to the value occurring in the most number of combinations of *outliers*.

Figure II.4:L'algorithme µ-argus**[12]**

❖ **Exemple [9]**

supposons que µ-argus est appliqué au **Tableau II.7** pour lequel l'utilisateur aura fourni les trois hiérarchies de généralisation représentées dans la **Figure II.1**, la **Figure II.2** et la **Figure II.3** relatives respectivement aux attributs du QI (sexe, code postal et niveau d'étude) et contraint le résultat par k = 2. Puis, lors de l'exécution du processus d'anonymisation, il aura opté, lors de la première itération, pour l'attribut niveau d'étude, puis pour l'attribut code postal et enfin pour l'attribut sexe. Enfin, il aura

terminé le processus par une suppression de la valeur de l'attribut « niveau d'étude » dans l'enregistrement qui ne satisfait pas le k-anonymat et fournira une table 2-anonyme **Tableau II.8**. Dans cette table, l'enregistrement touché par une suppression locale est marqué en rouge.**[9]**

Sexe	Code Postal	Niveau d'étude
Tout-sexe	1305*	Secondaire
Tout-sexe	1305*	Secondaire
Tout-sexe	1305*	Secondaire
Tout-sexe	1305*	Secondaire
Tout-sexe	1305*	1ier et 2iéme cycle
Tout-sexe	1305*	1ier et 2iéme cycle
Tout-sexe	1306*	*******
Tout-sexe	1306*	3iéme cycle
Tout-sexe	1306*	3iéme cycle
Tout-sexe	1306*	3iéme cycle
Tout-sexe	1306*	3iéme cycle
Tout-sexe	1306*	3iéme cycle
Tout-sexe	1306*	3iéme cycle

Tableau II.11. Résultat de l'application de µ-argus sur la table originale**[9]**

3.2.2. L'algorithme Datafly (Sweeney 1997)

est un algorithme heuristique gourmand qui effectue une généralisation unidimensionnelle de domaine complet. La figure 3 illustre les principales étapes de l'algorithme Datafly. Il compte la fréquence sur l'ensemble QID et si k-anonymat n'est pas encore satisfait, il généralise l'attribut ayant les valeurs les plus distinctes jusqu'à ce que k-anonymat soit satisfait. Alors que cet algorithme garantit une transformation k-anonyme, il ne fournit pas la généralisation minimale **[11]**

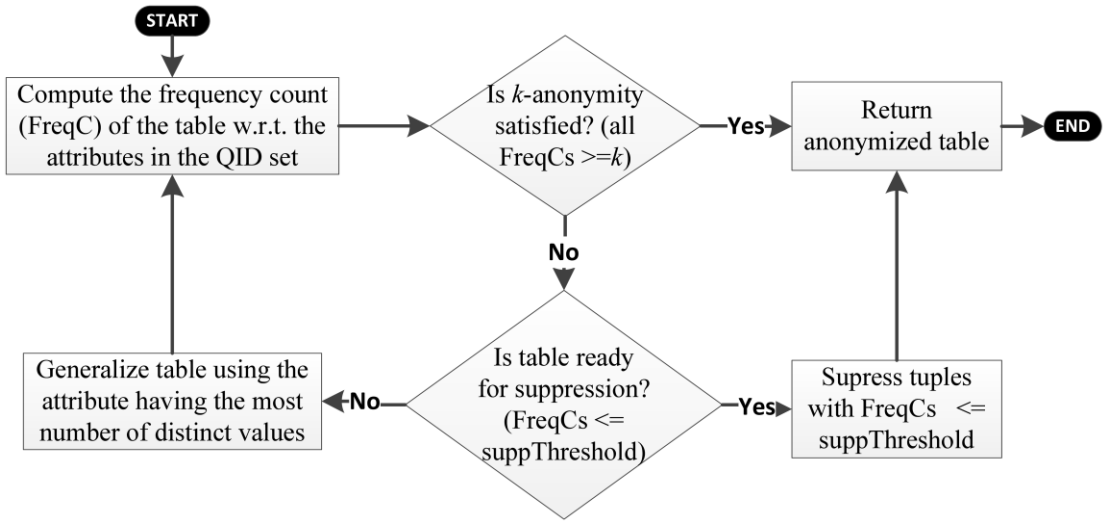


Figure II.5: Processus central de l'algorithme Datafly.**[11]**

suppose que l'utilisateur a fourni les trois hiérarchies de généralisation représentées dans la **Figure II.1**, la **Figure II.2** et la **Figure II.3**, relatives respectivement aux attributs du QI sexe, code postal et niveau d'étude et a contraint le résultat en fixant k = 2. Cet algorithme a fourni la table A titre d'exemple, à l'issue de la première itération de Datafly sur le **Tableau II.7**, on suivante :

Sexe	Code Postal	Niveau d'étude
M	13050	Collége
F	13051	Collége
M	13050	Lycée
M	13050	Lycée
M	13051	1er et 2éme cycle
F	13050	1er et 2éme cycle
F	13061	1er et 2éme cycle
F	13061	3éme cycle
F	13060	3éme cycle
M	13061	3éme cycle
M	13060	3éme cycle
M	13061	3éme cycle

Tableau II.12. Détection des enregistrements ne satisfaisant pas le k-anonymat **[9]**

Cette table a été obtenue après calcul du nombre de valeurs distinctes (dans la table originale) pour chaque attribut du QI sexe, code postal et niveau d'étude. Ce nombre

est respectivement de 2, 4 et 6. L'attribut niveau d'éducation, ayant le plus grand nombre, est généralisé. Etant donné que, dans la table résultant de la première itération, le nombre d'enregistrements (marqués en rouge) ne satisfaisant pas le k-anonymat (=8) est supérieur au seuil de suppressions autorisées (=2), alors l'algorithme poursuit la généralisation. La table 2-anonyme, résultat de son exécution totale, est la suivante :

Sexe	Code Postal	Niveau d'étude
M	1305*	Secondaire
M	1305*	Secondaire
M	1305*	Secondaire
M	1305*	Université
F	1305*	Université
F	1306*	Université
F	1306*	Université
F	1306*	Université
M	1306*	Université
M	1306*	Université
M	1306*	Université

Tableau II.13. Résultat de l'application de Datafly sur la table originale **[9]**

Les autres itérations de Datafly à partir du **Tableau II.9** afin d'atteindre le k-anonymat avec un minimum de généralisations et de suppressions (Sweeney 2002a).

3.2.3 L'algorithme « Bottom up generalization »

La généralisation ascendante est une méthode itérative du traitement des données pour généraliser les informations. Il est difficile d'établir un lien avec d'autres sources, même si les données généralisées restent utiles pour la classification. La maison de généralisation se résume à une structure de données de généralisations. Une clé est à chaque itération la meilleure généralisation se distingue pour gravir la hiérarchie. La généralisation ascendante convertit les données spécifiques en données moins spécifiques mais sémantiquement cohérentes pour la préservation de la confidentialité et s'est également concentrée sur deux problèmes principaux, l'évolutivité et la qualité. Le problème d'évolutivité a été résolu par une structure de données unique pour se concentrer sur de très bonnes généralisations. La même qualité est obtenue par le

système proposé mais une bien meilleure mesurabilité par rapport aux solutions existantes. Notre algorithme actuel a la probabilité de se retrouver coincé dans un optimum de voisinage en grimpant goulûment vers un état d'anonymat k .**[22]**

Une généralisation G, notée {di}€g, est la tâche consistant à remplacer toutes les valeurs filles di de l'ensemble {di} par leur valeur parente g. G est considérée comme candidate par rapport à une table si les descendants directs d1, d2, …, di, etc., notés {di} de g dans la hiérarchie de généralisation sont également dans la table. Elle est considérée comme bonne si elle renvoie le meilleur score calculé par application de la métrique de compromis IL/AG (Information Loss/Anonymity Gain) dont le rôle est de mesurer la perte d'information concernant la classification et le gain en sécurité liés à l'anonymisation.

La formule permettant de calculer le score d'une généralisation G, noté IL/AG(G) est la suivante :

$$IL/AG(G) = \begin{cases} \dfrac{InformationLoss(G)}{AnonymityGain(G)} & if \quad AnonymityGain(G) \neq 0 \\ InformationLoss(G) & otherwise \end{cases}.$$

Information Loss (G) correspond à la perte d'information suite à la réalisation de la généralisation G. Elle vise à garantir que le modèle de classification généré par l'ensemble de données anonymes a une efficacité approximativement équivalente au modèle de classification généré par les données originales. Elle se définit comme suit :

$$InformationLoss(G) = Entropy(R_g) - \sum_{di} \frac{|R_{di}|}{|R_g|} Entropy(R_{di})$$

où Rg (respectivement Rdi) représente l'ensemble des enregistrements contenant la valeur g (respectivement la valeur di).

Entropy (Rx) où x €{g, di} correspond à l'entropie de l'ensemble Rx.

Elle se calcule de la façon suivante :

$$Entropie(R_x) = -\sum_{cls} \frac{freq(R_x, cls)}{|R_x|} \times \log_2 \frac{freq(R_x, cls)}{|R_x|}$$

où freq (Rx , cls) représente le pourcentage d'individus de la classe labellisée cls dans Rx. Rappelons que la classification vise à classer les ensembles en catégories où chaque classe ou catégorie est labellisée.

Anonymity Gain(G) correspond au gain d'anonymat qui pourrait être engendré suite à la réalisation de la généralisation G. Intuitivement, cette mesure se calcule en comparant le degré d'anonymat d'une table avant et après application d'une généralisation G. De façon formelle, elle équivaut à :

Anonymat(T, après G) - Anonymat(T, avant G)

Où :

- Anonymat(T, après G) correspond à la taille de la plus petite classe d'équivalence (la classe d'équivalence qui renferme le plus petit nombre d'individus partageant le même QI) de T après application de G,

- et Anonymat(T, avant G) correspond à taille de la plus petite classe d'équivalence de T avant application de G. [9]

3.2.3.1.Algorithm 1 The bottom-up generalization **[23]**

1: **while** *R* does not satisfy the anonymity requirement **do**
2: **for all** generalization *G* **do**
3: compute *IP*(*G*);
4: **end for**;
5: find the best generalization *Gbest*;
6: generalize *R* by *Gbest*;
7: **end while**;
8: output *R*;

3.2.4.L'algorithme « Top down specialization » (B. C. Fung, Wang, et Yu 2005)

TDS est un processus répété qui commence à partir des valeurs de domaine les plus élevées dans les arborescences d'agencement des attributs.

Trouver la meilleure spécialisation, effectuer la spécialisation et mettre à jour les valeurs de la métrique de recherche. Un tel processus de TDS est répété jusqu'à ce que l'anonymat k soit violé, pour décrire le maximum de données qui seront utilisées. La justice d'une spécialisation est mesurée par une métrique de recherche. Les différentes autorisations d'application Android sont récupérées à partir d'applications Android. Ces autorisations sont utilisées comme ensemble de données pour le processus. En cela, vous acceptez le gain d'informations par perte de confidentialité (IGPL), une mesure de compromis qui prend en compte les exigences de confidentialité et d'informations. Une spécialisation avec la valeur IGPL maximale est considérée comme la meilleure et sélectionnée de chaque tour. Solution à tout moment pour la spécialisation descendante L'utilisateur peut parcourir chaque spécialisation pour déterminer le compromis

souhaité entre confidentialité et précision. L'utilisateur peut s'arrêter à tout moment et obtenir un tableau généralisé répondant à l'exigence d'anonymat.**[21]**

Soit S une spécialisation, notée aussi a→{si}. S est la tâche consistant à remplacer, dans la table à anonymiser, la valeur « a » par l'une des valeurs filles « si » de {si} se trouvant dans la hiérarchie de généralisation. Elle est considérée comme valide si elle respecte le k-anonymat et si elle renvoie le meilleur score par application de la métrique de compromis IG/AL (InformationGain/Anonymityloss). Cette métrique permet de réaliser le compromis entre le gain d'information et la perte d'anonymat dus à la spécialisation.

La formule permettant de calculer le score d'une spécialisation S, noté IG/AL(S), est la suivante :

$$IG/AL(S) = \begin{cases} \dfrac{InformationGain(S)}{AnonymityLoss(S)} & if \quad AnonymityLoss(S) \neq 0 \\ InformationGain(S) & otherwise \end{cases}$$

InformationGain(S) correspond au gain d'information suite à la réalisation de la spécialisation S. Elle se définit comme suit :

$$InformationGain(S) = Entropy(R_a) - \sum_{si} \frac{|R_{si}|}{|R_a|} Entropy(R_a)$$

où Ra (respectivement Rsi) représentent l'ensemble des enregistrements contenant la valeur a (respectivement la valeur si).

Entropy (Rx) où x € {a, si} correspond à l'entropie de l'ensemble Rx.

Elle se calcule de la façon suivante :

$$Entropie(R_x) = -\sum_{cls} \frac{freq(R_x, cls)}{|R_x|} \times \log_2 \frac{freq(R_x, cls)}{|R_x|}$$

où freq(R_x, cls) représente le pourcentage d'individus de la classe labellisée cls dans R_x. Rappelons que la classification vise à classifier les ensembles en catégories où chaque classe ou catégorie est labellisée.

AnonymityLoss (S) correspond à la perte d'anonymat qui pourrait être engendrée suite à la réalisation de la spécialisation S.

Intuitivement, cette mesure se calcule en comparant le degré d'anonymat d'une table avant et après application d'une spécialisation S. De façon formelle, elle équivaut à :

Anonymat(T, après S) - Anonymat(T, avant S)

Où :

- Anonymat(T, après S) correspond à la taille de la plus petite classe d'équivalence (la classe d'équivalence qui renferme le plus petit nombre d'individus partageant le même QI) de T après application de S,

- et Anonymat(T, avant S) correspond à la taille de la plus petite classe d'équivalence de T avant application de S. **[9]**

3.2.4.1Algorithm Top down specialization:**[25]**

Input: a table T, parameter k, weights of attributes,
hierarchies on categorical attributes;
Output: a k-anonymous table T′;
Method:
1: IF |T| ≤ k THEN RETURN;
2: ELSE {
3: partition T into two exclusive subsets T1 and T2 such
that T1 and T2 are more local than T, and either T1
or T2 have at least k tuples;
4: IF |T1| > k THEN recursively partition T1;
5: IF |T2| > k THEN recursively partition T2;
}
6: adjust the groups so that each group has at least k tuples;

4.Tableau comparative

Algorithmes	**Qalité**	**Métrique**	**Opération**	**Ascendant/ Descendant**	**Attack**
Min Gen	Optimal	MD	Généralisation/suppression		Cuplage d'enregistrement
µ-Argus	Minimal	DA	Généralisation/suppression	Ascendant	Cuplage d'enregistrement
Data fly	Minimal	MD	Généralisation /suppression	Ascendant	Cuplage d'enregistrement
Botton up	Minimal	ILPG	Généralisation	Ascendant	Cuplage d'enregistrement
Botton Dwn	Minimal	IGPL	Généralisation	Descendant	Cuplage d'enregistrement

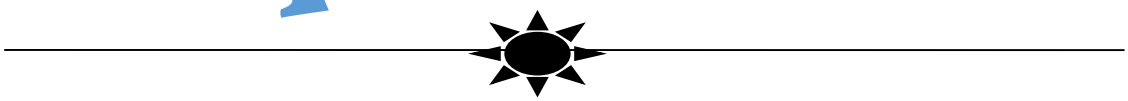
Tableau II.14. Tableau comparative

5 .Conclusion

L'analyse et la publication des données préservant la confidentialité deviennent des problèmes sérieux dans le monde actuel. C'est pourquoi différentes approches des

techniques d'anonymisation des données sont proposées. Il existe diverses techniques d'anonymisation et elles se concentrent principalement sur l'anonymat k, qui comprend à la fois la généralisation et la suppression. Les algorithmes de généralisation et leur implémentation pour protéger la confidentialité des données utilisés principalement pour l'analyse des données. En particulier, le document a présenté une généralisation ascendante pour transformer des données spécifiques en données moins spécifiques mais sémantiquement cohérentes pour la protection de la vie privée.

Chapitre III



l'approche d'anonymisation

1. Introduction

Avec le développement de la technologie Internet et ,technologie de traitement des données et un grand nombre de données associées avec des individus, tels que les données médicales du patient les données sont collectées et largement diffusées par le gouvernement départements et instituts de recherche. Cependant, ces données il peut contenir des informations privées pour les particuliers et un grand nombre de données entraîne une utilisation généralisée l'un des outils d'extraction de données, protégeant ainsi la vie privée l'information est une préoccupation majeure **[32].** Donc, pour cacher une identité les données et une variété de principes inconnus. Nous suggérons d'utiliser des algorithmes Bottom up generalization et Top down specialization.

2. L'algorithme « Bottom up generalization »

2.1. Les etapes

A chaque itération, la généralisation ascendante fonctionne comme suit :

1. Sélectionner les généralisations candidates et les généralisations critiques.

2. S'il existe des généralisations critiques, calculer le score de chaque. Si non calculer le score de chaque généralisation candidate.

3. Choisir la meilleure généralisation selon son score et l'effectuer.

4. Vérifier si la table satisfait le k-anonymat : si oui, l'algorithme s'arrête, si non, il passe à l'itération suivante.

2.2. Processus d'anonymisation

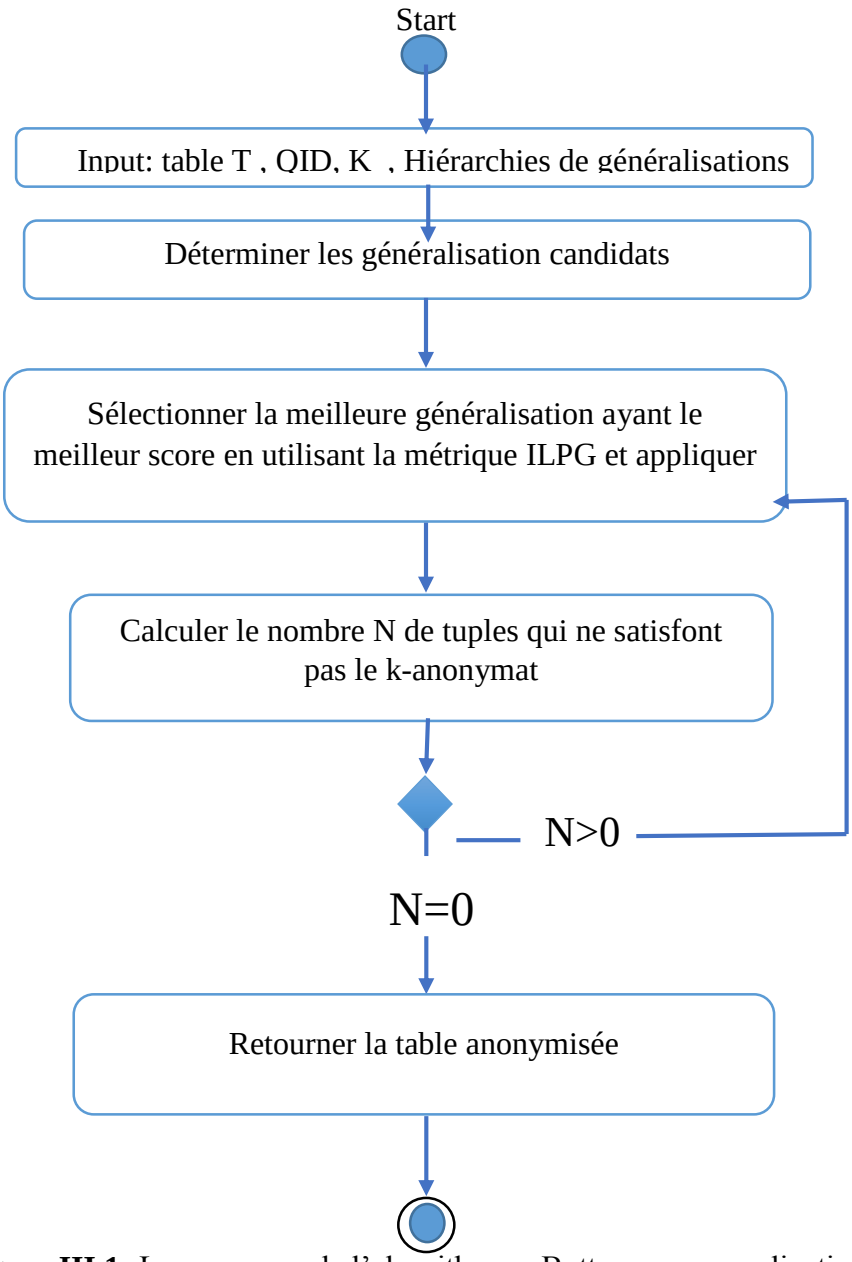


Figure III.1: Le processus de l'algorithme « Bottom up generalization »

K est le nombre d'enregistrements qui ont la même valeur de QID.

2.3. Etude d'un Exemple

Sexe	Code Postal	Niveau d'étude	Classe
M	13050	5iéme	2Y0N
F	13051	3iéme	3Y0N
M	13050	Seconde	2Y0N
F	13051	Seconde	3Y0N
F	13050	1er et 2éme cycle	4Y0N
F	13061	1er et 2éme cycle	1Y0N
F	13061	1er et 2éme cycle	0Y1N
F	13060	Master	0Y4N
M	13061	Master	0Y1N
M	13060	Doctorat	0Y2N

Tableau III.1. table original. **[9]**

la colonne classe contient la fréquence de la classe des enregistrements. La classe Y pour ceux qui ont un salaire < 2000 et la classe N pour ceux ayant un salaire>= 2000. Par exemple, 2Y1N signifie qu'il existe deux enregistrements appartenant à la classe Y et un seul appartient à la classe N. **[9]**

Itération 1

Etape1: Les généralisations candidates sont: "{F,M}→tout-sexe", "{13050,13051}→1305*", "{13060,13061}→1306*", "{3ième, 5ième }→ collège", "{seconde,terminal}→lycée", "{master,doctorat}→3ième cycle"

Etape 2:

Calcul du score de la généralisation « {F,M}→tout-sexe »

Sexe	**Classe**	**occurrence**
F	11Y 5N	16
M	4Y 3N	7

Tableau III.2. fréquence de la classe de généralisation sexe

I(F)= -(11/16 *log2 (11/16))- (5/16* log2 (5/16))= 0.25+0.36=0.61

I(M)= - (4/7 *log2 (4/7)) - (3/7 *log2 (3/7)) = 0,55 + 0,36 = 0,91

I(tout-sex)= - (15/23* log2 (15/23à) - (8/23* log2 (8/23)) = 0,27 + 0,36 = 0,63

InfoLoss (tout-sex) = I(tout-sex) – ((16/23 * I (F)) + (7/23 * I (M)))= 0,63 – (0,7*0,61 + 0,3*0,91) = 0,46

Calcul du score de la généralisation « {13050,13051}→1305*»

Code Postal	Classe	Occurrence
13050	8Y 0N	8
13051	6Y 0N	6
1305*	14Y 0N	14

Tableau III.3. fréquence de la classe de généralisation code postal"1305*"

I (13050) = 0

I (13051) = 0

I (1305*) = 0

InfoLoss (1305*) = 0

Calcul du score de la généralisation « {13060,13061}→ 1306*»

Code Postal	Classe	Occurrence
13060	0Y 5N	5
13061	1Y 2N	3
1306*	1Y 7N	8

Tableau III.4. fréquence de la classe de généralisation code postal"1306*"

I (13060) = -(0/5 * log2 (0/5)) – (5/5 * log2 (5/5))= 0

I (13061) = - (1/3 * log2 (1/3)) – (2/3 * log2 (2/3))= 0.52+ 0,39= 0,62

I (1306*) = - (1/8 * log2 (1/8)) – (7/8 * log2 (7/8)) = 0,38 + 0.17 = 0,36

InfoLoss (1306*) = I (1306*) – ((5/8 * I (13060))+ (3/8 * I (13061)) = 0,36 – (0,37 * 0,62) = 0,13

Calcul du score de la généralisation «{3ième, 5ième }→ collège»

Niveau d'etude	Classe	Occurrence
3iéme	2Y 0N	2
5iéme	3Y 0N	3
collége	5Y 0N	5

Tableau III.5. fréquence de la classe de généralisation niveau d'etude"collège"

I (3ième) = 0

I (5ième) = 0

I (collège) = 0

InfoLoss (collège) = 0

Calcul du score de la généralisation «{seconde, terminal}→lycée»

Niveau d'etude	**Classe**	**Occurrence**
seconde	5Y 0N	5
terminal	0Y 0N	0
lycée	5Y 0N	5

Tableau III.6. fréquence de la classe de généralisation niveau d'etude"lycée"

I (seconde) = 0

I (terminal) = 0

I (lycée) = 0

InfoLoss (lycée) = 0

Calcul du score de la généralisation «{master, doctorat}→ 3ième cycle»

Niveau d'etude	**Classe**	**Occurrence**
master	0Y 5N	5
doctorat	0Y 3N	3
3iéme cycle	0Y 8N	8

Tableau III.7. fréquence de la classe de généralisation niveau d'etude"3ième cycle "

I (master) = 0

I (doctorat) = 0

I (3ième cycle) = 0

InfoLoss (3ième cycle) = 0

Etape 3 : Les généralisations ayant le minimum de score sont:

“{13050,13051}→1305*”, “{3ième, 5ième }→ collège”,

“{seconde,terminal}→lycée”, “{master,doctorat}→ 3ième cycle”. Nous choisissons aléatoirement la généralisation “{13050,13051}→1305*”.

Sexe	Code Postal	Niveau d'étude	Classe
M	1305*	5iéme	2Y0N
F	1305*	3iéme	3Y0N
M	1305*	Seconde	2Y0N
F	1305*	Seconde	3Y0N
F	1305*	1er et 2éme cycle	4Y0N
F	13061	1er et 2éme cycle	1Y0N
F	13061	1er et 2éme cycle	0Y1N
F	13060	Master	0Y4N
M	13061	Master	0Y1N
M	13060	Doctorat	0Y2N

Tableau III.8.table de résultat de la première itération

Etape 4: Cette table ne satisfait pas le k- anonymat. L'algorithme passe à l'itération suivante.

Itération 2

Etape 1: Les généralisations candidates sont: "{F,M}→tout-sexe", "{13060,13061}→ 1306*","{3ième,5ième}→collège","{seconde,terminal}→lycée", "{master,doctorat}→ 3ième cycle"

Etape 2:

Score de "{F,M}→tout-sexe" = 0.46

Score de "{13060,13061}→1306*" =0.13

Score de "{3ième, 5ième }→ collège" = 0

Score de "{seconde,terminal}→lycée" = 0

Score de "{master,doctorat}→ 3ième cycle" = 0

Etape 3 : Les généralisations ayant le minimum de score sont: "{3ième, 5ième}→ collège", "{seconde,terminal}→lycée", "{master,doctorat}→ 3ième cycle". Nous choisissons aléatoirement la généralisation "{3ième, 5ième }→ collège".

Sexe	Code Postal	Niveau d'étude	Classe
M	1305*	Collége	2Y0N
F	1305*	Collége	3Y0N
M	1305*	Seconde	2Y0N
F	1305*	Seconde	3Y0N
F	1305*	1er et 2éme cycle	4Y0N
F	13061	1er et 2éme cycle	1Y0N
F	13061	1er et 2éme cycle	0Y1N
F	13060	Master	0Y4N
M	13061	Master	0Y1N
M	13060	Doctorat	0Y2N

Tableau III.9. table de résultat de la 2éme itération

Etape 4 : Cette table ne satisfait pas le k- anonymat. L'algorithme passe à l'itération suivante.

Itération 3

Etape1: Les généralisations candidates sont: "{F,M}→tout-sexe", "{13060,13061}→1306*", "{seconde,terminal}→lycée", "{master,doctorat}→3ième cycle"

Etape 2:

Score de "**{**F,M}→tout-sexe" = 0,46

Score de "{13060,13061}→1306*" =0.13

Score de "{seconde,terminal}→lycée" = 0

Score de "{master,doctorat}→3ième cycle" = 0

Etape 3: Les généralisations ayant le minimum de score sont: "{seconde,terminal}→lycée", "{master,doctorat}→ 3ième cycle". Nous choisissons aléatoirement la généralisation "{master,doctorat}→ 3ième cycle".

Sexe	Code Postal	Niveau d'étude	Classe
M	1305*	Collége	2Y0N
F	1305*	Collége	3Y0N
M	1305*	Seconde	2Y0N
F	1305*	Seconde	3Y0N
F	1305*	1er et 2éme cycle	4Y0N
F	13061	1er et 2éme cycle	1Y0N
F	13061	1er et 2éme cycle	0Y1N
F	13060	3ième cycle	0Y4N
M	13061	3ième cycle	0Y1N
M	13060	3ième cycle	0Y2N

Tableau III.10. table de résultat de la 3ième itération

Etape 4 : Cette table ne satisfait pas le k- anonymat. L'algorithme passe à l'itération suivante.

Itération 4

Etape1: Les généralisations candidates sont: "{F,M}→tout-sexe", "{13060,13061}→1306*", "{seconde,terminal}→lycée".

Etape 2:

Score de "{F,M}→tout-sexe" = 0,46

Score de "{13060,13061}→ 1306*" =0.13

Score de "{seconde,terminal}→lycée" = 0

Etape3 : La généralisation ayant le minimum de score est "{seconde,terminal}→lycée"

Sexe	Code Postal	Niveau d'étude	Classe
M	1305*	Collége	2Y0N
F	1305*	Collége	3Y0N
M	1305*	lycée	2Y0N
F	1305*	lycée	3Y0N
F	1305*	1er et 2éme cycle	4Y0N
F	13061	1er et 2éme cycle	1Y0N
F	13061	1er et 2éme cycle	0Y1N
F	13060	3ième cycle	0Y4N
M	13061	3ième cycle	0Y1N
M	13060	3ième cycle	0Y2N

Tableau III.11. table de résultat de la 4ème itération

Etape 4 : Cette table ne satisfait pas le k- anonymat. L'algorithme passe à l'itération suivante.

Itération 5

Etape1: Les généralisations candidates sont: "**{**F,M}→tout-sexe", "{13060,13061}→1306*", "{collège, lycée}→ secondaire", "{1ier et 2ième cycle, 3ième cycle }→ université".

Etape 2:

Score de "**{**F,M}→tout-sexe" = 0,46

Score de "{13060,13061}→ 1306*" =0.13

Score de "{collège, lycée}→ secondaire"= 0

Score de "{1ier et 2ième cycle, 3ième cycle }→ université" = 0,68

Etape3: La généralisation ayant le minimum de score est "{collège, lycée}→secondaire"

Sexe	Code Postal	Niveau d'étude	Classe
M	1305*	secondaire	2Y0N
F	1305*	secondaire	3Y0N
M	1305*	secondaire	2Y0N
F	1305*	secondaire	3Y0N
F	1305*	1er et 2éme cycle	4Y0N
F	13061	1er et 2éme cycle	1Y0N
F	13061	1er et 2éme cycle	0Y1N
F	13060	3ième cycle	0Y4N
M	13061	3ième cycle	0Y1N
M	13060	3ième cycle	0Y2N

Tableau III.12. table de résultat de la 5ième itération

Etape 4 : Cette table ne satisfait pas le k- anonymat. L'algorithme passe à l'itération suivante.

Itération 6

Etape1: Les généralisations candidates sont: "**{**F,M}→tout-sexe", "{13060,13061}→1306*", "{1ier et 2ième cycle, 3ième cycle }→ université".

Etape 2:

Score de "**{**F,M}→tout-sexe" = 0,46

Score de "{13060,13061}→ 1306*" =0.13

Score de "{1ier et 2ième cycle, 3ième cycle }→ université" = 0,68

Etape 3 : La généralisation ayant le minimum de score est "{13060,13061}→ 1306*"

Sexe	Code Postal	Niveau d'étude	Classe
M	1305*	secondaire	2Y0N
F	1305*	secondaire	3Y0N
M	1305*	secondaire	2Y0N
F	1305*	secondaire	3Y0N
F	1305*	1er et 2éme cycle	4Y0N
F	1306*	1er et 2éme cycle	1Y0N
F	1306*	1er et 2éme cycle	0Y1N
F	1306*	3ième cycle	0Y4N
M	1306*	3ième cycle	0Y1N
M	1306*	3ième cycle	0Y2N

Tableau III.13. table de résultat de la 6 ème itération

Etape 4 : Cette table ne satisfait pas le k- anonymat. L'algorithme passe à l'itération suivante.

Itération 7

Etape1: Les généralisations candidates sont: "**{**F,M}→tout-sexe",

{1305*,1306*}→130** et "{1ier et 2ième cycle, 3ième cycle }→ université".

Etape 2:

La seule généralisation critique est {1305*,1306*}→130**.

Etape 3 : La généralisation ayant le minimum de score est "{1305*,1306*}→130**"

Sexe	Code Postal	Niveau d'étude	Classe
M	130**	secondaire	2Y0N
F	130**	secondaire	3Y0N
M	130**	secondaire	2Y0N
F	130**	secondaire	3Y0N
F	130**	1er et 2éme cycle	4Y0N
F	130**	1er et 2éme cycle	1Y0N
F	130**	1er et 2éme cycle	0Y1N
F	130**	3ième cycle	0Y4N
M	130**	3ième cycle	0Y1N
M	130**	3ième cycle	0Y2N

Tableau III.14. table de résultat final

Etape 4 : Cette table satisfait le k- anonymat. L'algorithme s'arrête.

3. L'algorithme Spécialisation descendante «Top down specialization»

3.1. Processus d'anonymisation

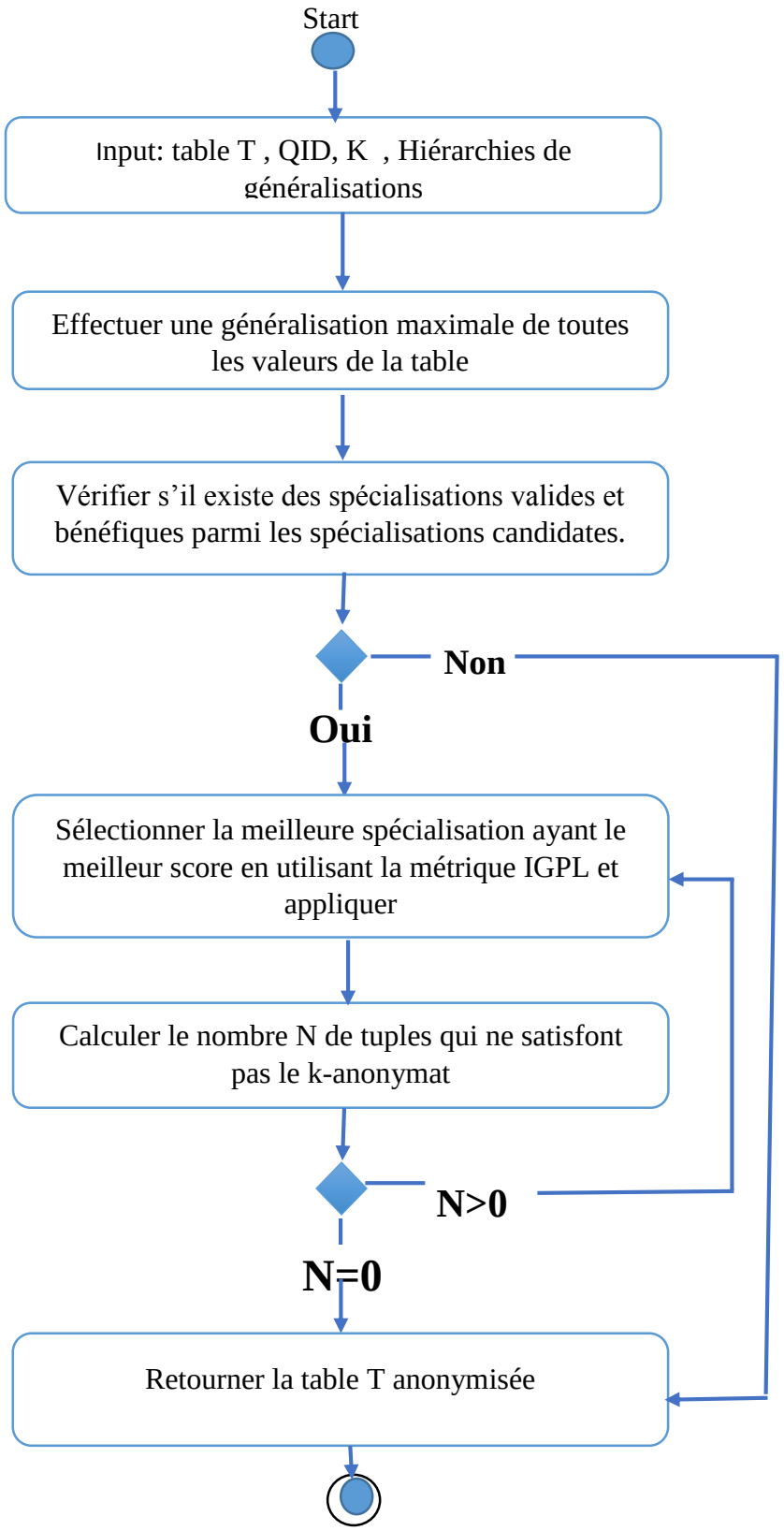


Figure III.2: Le processus de l'algorithme « Top down specialization»

3.2. Les étapes

La première partie de TDS permet d'effectuer une généralisation maximale de toutes les valeurs de la table originale. Ensuite, chaque itération se fonctionne comme suit :

1. Vérifier s'il existe des spécialisations valides et bénéfiques parmi les spécialisations candidates. Si oui, l'algorithme passe à l'étape suivante, si non, il s'arrête.

2. Calculer le score de chaque spécialisation.

3. Choisir la spécialisation ayant le meilleur score.

3.3. Etude d'un Exemple

Partie1

Sexe	Code postal	Niveau d'étude	Class	Occurence
Tout-sexe	130**	Tout-education	15Y8N	23

Tableau III.15. spécialisation de table origine

Partie2

Itération1

Etape 1: les spécialisations valides et bénéfiques sont : "tout-sexe→ {F,M}", "130**→{1306*,1305*}", "tout-education → {université, secondaire}"

Etape 2 : • Calcul du score de la spécialisation "tout-sexe→{F,M}"

Si on applique "tout-sexe→ {F,M}", le résultat sera la table suivante:

Sexe	Code postal	Niveau d'étude	Class	Occurence
F	130**	Tout-éducation	11Y5N	16
M	130**	Tout-éducation	4Y3N	7

Tableau III.16. la résultat de la spécialisation tout-sexe

I(F)= - (11/16 *log2 (11/16)) - (5/16* log2(5/16)) = 0,25 + 0,36 = 0,61

I(M)= - (4/7* log2 (4/7)) - (3/7* log2(3/7)) = 0,55 + 0,36 = 0,91

I(tout-sex)= - (15/23 *log2 (15/23)) - (8/23* log2(8/23)) = 0,27 + 0,36 = 0,63

MC gain (tout-sex) = I(tout-sex) – ((16/23 * I (F)) + (7/23 * I (M)))= 0,63 – (0,7*0,61 + 0,3*0,91) = 0,46

Anonymity Loss (tout-sex) = 23 – 7 = 16

Score (tout-sexe→{F,M}) =0,46/16 = 0,028

• **Calcul du score de la spécialisation** ”130**→ {1305*,1306*}”

Si on applique la spécialisation ”130**→ {1305*,1306*}”, le résultat sera la table suivante:

Sexe	Code postal	Niveau d'éducation	Classe	Occurrence
Tout-sexe	1305*	Tout-education	14Y0N	14
Tout-sexe	1306*	Tout-education	1Y8N	9

Tableau III.17. la résultat de la spécialisation130**

I(1305*) = - (14/14* log2 (14/14)) = 0

I(1306*)= - (1/9 *log2(1/9)) - (8/9*log2(8/9)) = 0,24 + 0,1 = 0,34

I(130**)= - (15/23* log2 (15/23)) - (8/23 *log2(8/23)) = 0,27 + 0,36 = 0,63

MC gain (130**) = I(130**) – ((14/23 * I (1305*)) + (9/23 * I (1306*)))= 0,63 – (0,39*0,34) = 0,63- 0,13 = 0,49

Anonymity Loss (130**) = 23 – 10 = 13

Score (130**→ {1305*,1306*}) =0,49/13 = 0,03

• **Calcul du score de la spécialisation** ”tout-éducation→ { Secondary, Université }”

Si on applique la spécialisation ”tout-éducation→ { Secondary, Université }”, le résultat sera la table suivante:

Sexe	Code postal	Niveau d'éducation	Classe	Occurrence
Tout-sexe	1305*	secondaire	7Y0N	7
Tout-sexe	1306*	Université	8Y8N	16

Tableau III.18. la résultat de la spécialisation tout-éducation **' 1'**

I(secondaire)= - (7/7* log2 (7/7)) = 0

I(Université)= - (8/16* log2 (8/16)) - (8/16* log2 (8/16)) = 0,35 + 0,35 = 0,7

I(tout-education)= - (15/23* log2 (15/23)) - (8/23* log2(8/23)) = 0,27 + 0,36 = 0,63

MCgain (tout-education)= I(tout-education) – ((7/23 * I (Secondaire)) + (16/23 * I (université))) = 0,63 - AnonymityLoss(tout-education)= 23 – 8 = 15

Score (tout-éducation→{ Secondaire, Université }) =0,63/15 = 0,042

Etape 3 : la spécialisation qui a le meilleur score est tout-education →{secondaire, université}

Sexe	Code postal	Niveau d'éducation	Classe	Occurrence
Tout-sexe	1305*	secondaire	7Y0N	7
Tout-sexe	1306*	Université	8Y8N	16

Tableau III.19. la résultat de la spécialisation tout-éducation **' 2'**

Itération 2

Etape 1: les spécialisations valides et bénéfiques sont : ”tout-sexe→ {F,M}”, “130**→ {1306*,1305*}”, université→ {1 ier et 2ième cycle, 1 ier et 3ième cycle }.

Etape 2 :

- **Score** (université→ {1 ier et 2ième cycle, 1 ier et 3ième cycle }) =0,7
- **Score** (130**→ {1306*,1305*}) =0,49
- **Score** (tout-sexe→ {F,M}) = 0,115

Etape 3 : la spécialisation qui a le meilleur score est université→{1 ier et 2ième cycle, 3 ième cycle }

Sexe	Code postal	Niveau d'éducation	Classe	Occurrence
Tout-sexe	130**	secondaire	7Y0N	7
Tout-sexe	130**	1 ier et 2ième cycle	8Y0N	8
Tout-sexe	130**	3 ième cycle	0Y8N	8

Tableau III.20. la résultat de la spécialisation tout-éducation **' 3'**

Itération 3

Etape 1: la spécialisation valide et bénéfique est ”tout-sexe→{F,M}”.

Etape 2: Score (tout-sexe→ {F,M}) = 0,115

Etape 3 : la spécialisation qui a le meilleur score est tout-sexe→ {F,M}

Sexe	Code postal	Niveau d'éducation	Classe	Occurrence
F	130**	secondaire	3Y0N	3
M	130**	secondaire	4Y0N	4
F	130**	1 ier et 2ième cycle	8Y0N	8
F	130**	3 ième cycle	0Y5N	5
M	130**	3 ième cycle	0Y3N	3

Tableau III.21. la résultat de la spécialisation tout-sexe

Itération 4

Etape 1: Il n'y a aucune spécialisation valide et bénéfique. La table proposée au data publisher est la suivante **[9]**:

Sexe	Code postal	Niveau d'éducation	Classe	Occurrence
F	130**	secondaire	3Y0N	3
M	130**	secondaire	4Y0N	4
F	130**	1 ier et 2ième cycle	8Y0N	8
F	130**	3 ième cycle	0Y5N	5
M	130**	3 ième cycle	0Y3N	3

Tableau III.22. table de résultat final

4. Conclusion

Dans ce chapitre, nous avons discuté des algorithmes Bottom up generalization et Top down specialization. En détail en termes d'étapes de chaque algorithme et d'un exemple détaillé pour chaque unité.

Chapitre IV

Implémentation et discussion

1. Introduction

Dans le présent chapitre, on va présenter la partie implémentation de l'algorithme " Bottom up généralisation " pour la préservation de la confidentialité.

Nous choisissions l'algorithme "Bottom up généralisaion" pour les raisons suivantes:

- Minimise la perte d'informations et maximise le gain de confidentialité
- Elle préserve la «véracité» d'informations, ce qui rend les données publiées significatives au niveau record.
- Discrimination et purification: un avantage particulier de l'approche ascendante repose sur le fait qu'elle est susceptible de capturer des modèles comparativement plus purs.

2. Les outils de développement

Pour réaliser notre travail, on a eu besoin d'un ensemble d'outils et de moyens de développement .On a choisi dans notre cas et pour des raisons d'efficacité et de fiabilité les moyens suivants :

- ❖ L'environnement de simulation
- ❖ Langage de programmation : Java
- ❖ L'environnement de développement (IDE) : Netbeans 8.2

2.1. L'environnement de simulation

Nous avons développé notre application à l'aide du langage Java sur Windows 7 Professional 32 bit, de RAM de 4.00 Go, et de processus Intel(R) Core(TM) i3-5005U CPU @ 2.20GHz, 2.00GHz.

2.2. Java comme langage de programmation

Le langage **Java** est un langage de programmation informatique orienté objet, La particularité et l'objectif central de Java est que les logiciels écrits dans ce langage doivent être très facilement portables sur plusieurs systèmes d'exploitation tels que UNIX, Windows, Mac OS ou GNU/Linux, avec peu ou pas de modifications. Pour cela, divers plate-formes et frameworks associés visent à guider, sinon garantir, cette portabilité des applications développées en Java. [13]

2.3. Neatbeans comme environnement de développement

NetBeans est un environnement de développement intégré (EDI), NetBeans constitue par ailleurs une plateforme qui permet le développement d'applications spécifiques (bibliothèque Swing (Java)). **[35]**

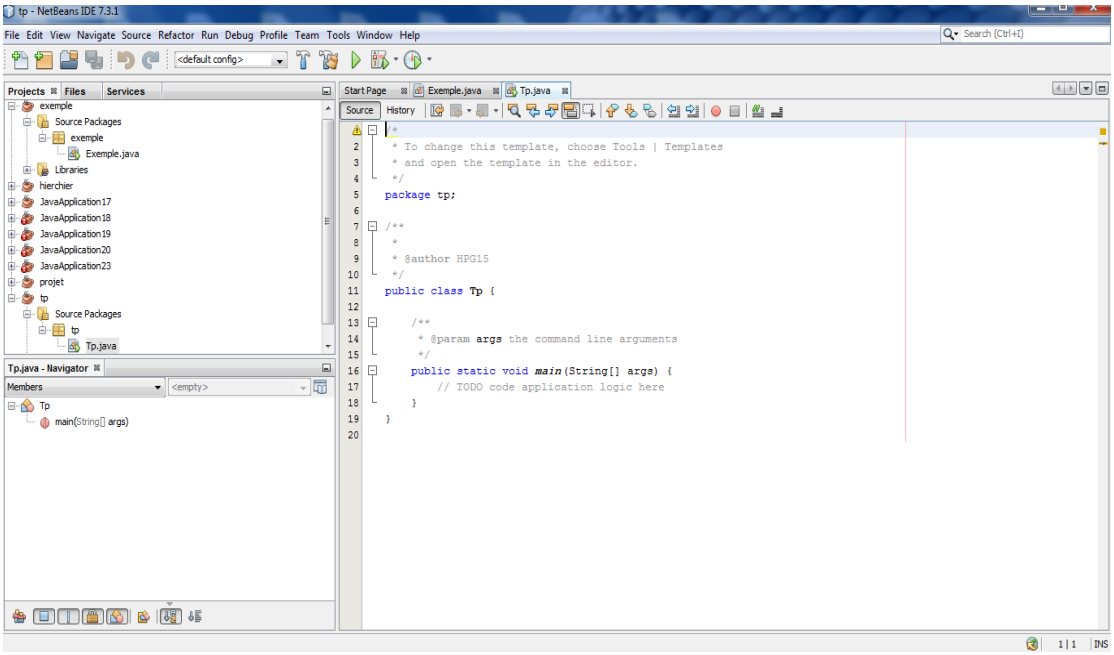


Figure IV.1: L'interface de NetBeans

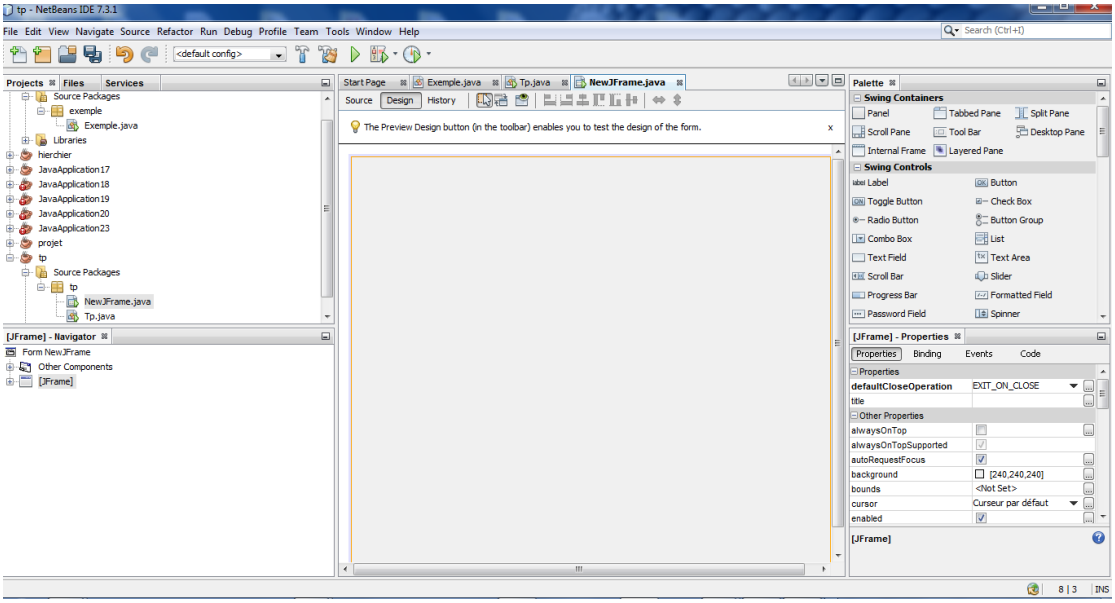


Figure IV.2: Fenetre JFrame Form en NetBeans

3. Le dataset utilisés

3.1. Dataset Adult

Nous avons testé notre algorithme sur la base de données du Bureau du recensement des États-Unis ADULT qui contient les caractéristiques démographiques d'un petit échantillon de la population américaine. ADULT contient 48842 enregistrements avec 14 attributs publics (Age, Work CLass, etc.). L'information privée signifie que les gains individuels sont plus ou moins de 50 mille dollars par an. Nous avons utilisé cet ensemble de données vu que la majorité des travaux qui touchent l'aspect anonymisation l'utilise. **[1]**

3.1.1.Les attributs

Age:

est le nombre d'années que quelque chose a vécu ou a existé. Un exemple d'âge est d'avoir 16 ans.

Sexe:

Le mot "sexe"se réfère davantage aux caractéristiques biologiques et physiologiques qui différencient les hommes des femmes**. [39]**

Class de travail:"Work class"

le groupe social composé de personnes qui effectuent habituellement un travail physique et qui n'ont généralement pas beaucoup d'argent ou un niveau d'éducation très élevé.

Education

les connaissances et le développement résultant du processus d'éducation.

Course:"Race"

l'un des groupes dans lesquels les humains sont souvent divisés en fonction de leurs traits physiques.

"Marital-Status " : L'état civil

L'état civil, ou l'état matrimonial, sont les options distinctes qui décrivent la relation d'une personne avec un autre significatif. Marié, célibataire, divorcé et veuf sont des exemples d'état civil. **[40]**

Pays d'origine :"Native-Country"

le pays dans lequel une personne est née ou native.

4. L'objectif de notre application

On a réalisé l'algorithme Bottom up généralisation qui convertit les données spécifiques à des données moins spécifiques mais sémantiquement cohérentes pour la préservation de la confidentialité.

Cette algorithme sélectionne la meilleure généralisation qui minimise la perte d'informations et maximise le gain de confidentialité**,** afin d'évaluer cette algorithme découvrir les paramètres qui lui influent pour obtenir les meilleures résultats.

5. Le choix de l'algorithme

Nous choisissions l'algorithme "Bottom up généralisaion" pour les raisons suivantes:

- Minimise la perte d'informations et maximise le gain de confidentialité
- Elle préserve la «véracité» d'informations, ce qui rend les données publiées significatives au niveau record.
- Discrimination et purification: un avantage particulier de l'approche ascendante repose sur le fait qu'elle est susceptible de capturer des modèles comparativement plus purs

6. L'organisation de l'application

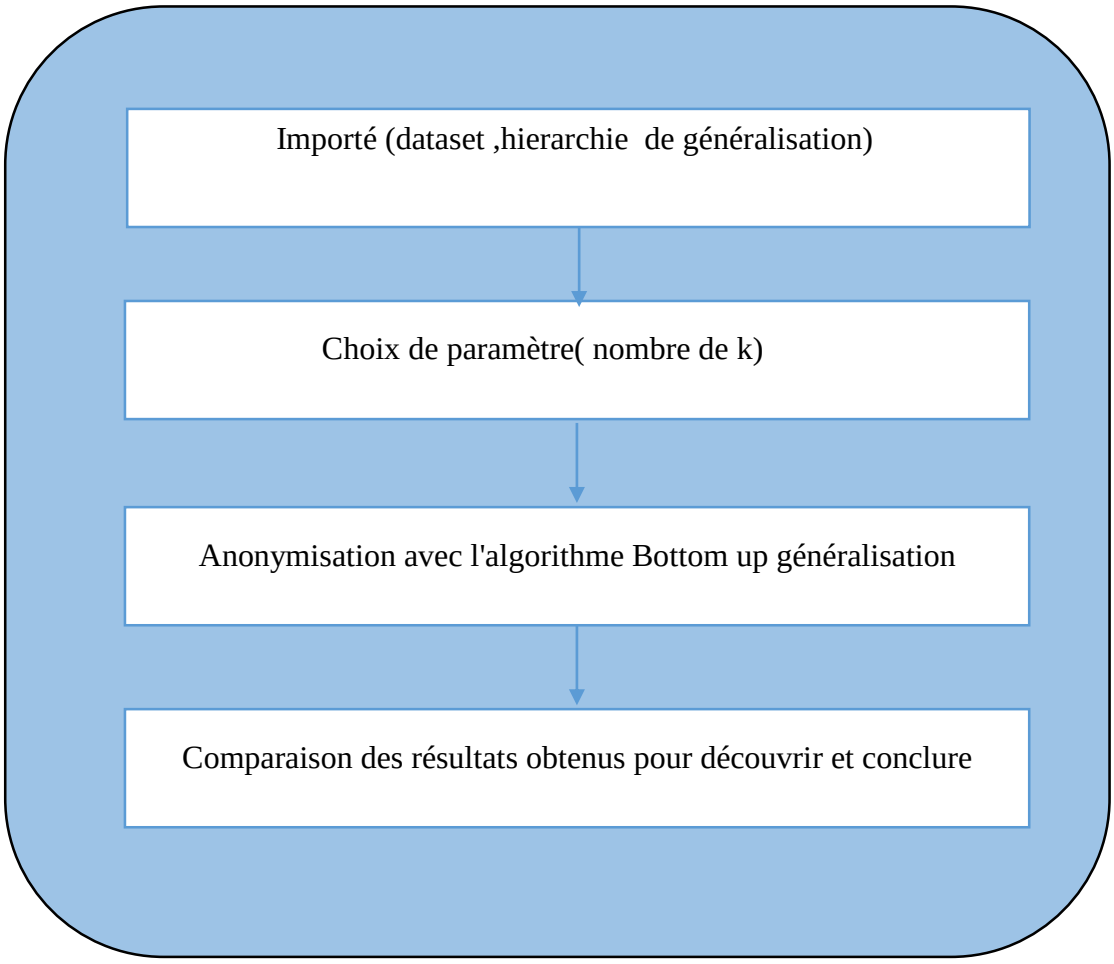


Figure IV.3 Les étapes de simulation

7. Description des étapes de simulation

L' interface graphique de cette application est comme suite:

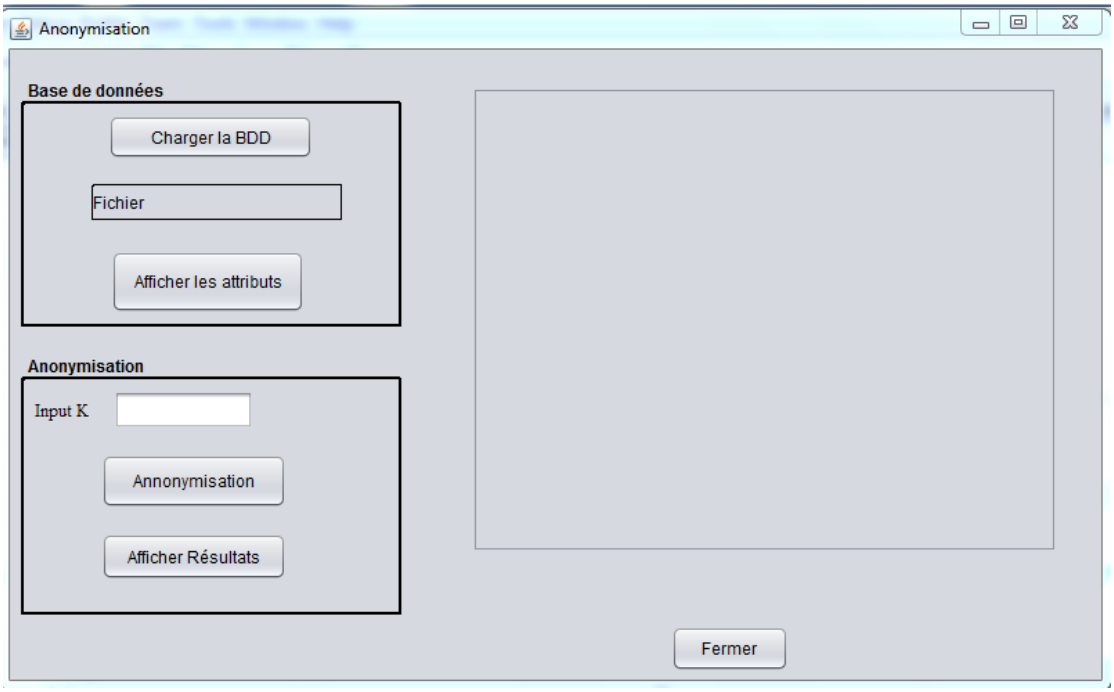


Figure IV.4: L'interface de simulation.

7.1. Importé (dataset ,hierarchie de généralisation)

Avant de simuler l'algorithme de Bottom up généralisation, on fait importation dataset et hierarchie de généralisation (table.xls)

Importé BDD

Lorsque vous cliquez sur le bouton Télécharger BDD, une nouvelle fenêtre apparaîtra contenant les fichiers de votre ordinateur, nous choisissons donc le fichier qui contient le fichier excel, puis cliquez sur Ouvrir.

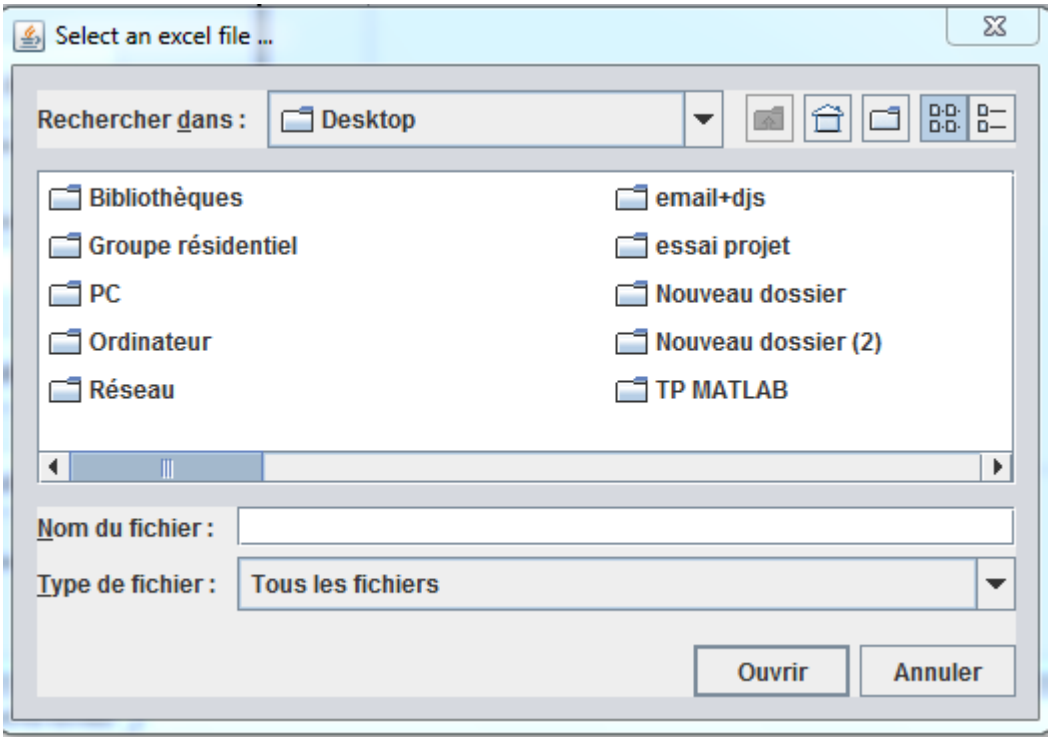


Figure IV.5: L'interface qui contenant les fichiers de votre ordinateur

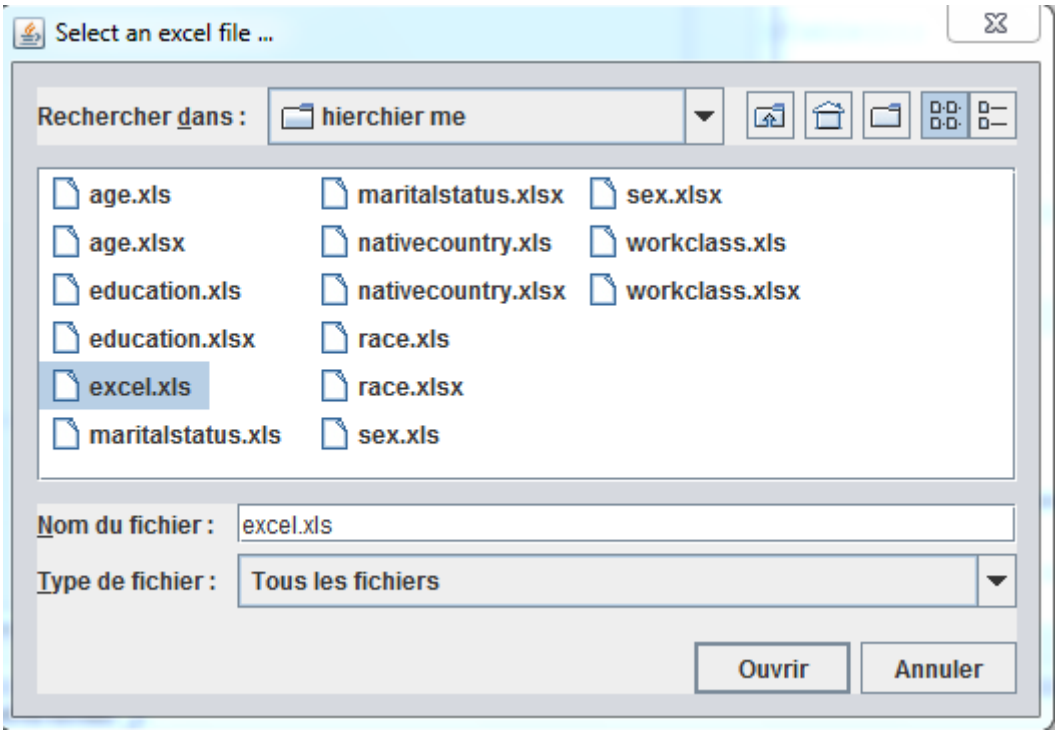


Figure IV.6: L'interface qui contenant les fichiers Excel

Après avoir sélectionné le fichier, le tableau de BDD apparaît

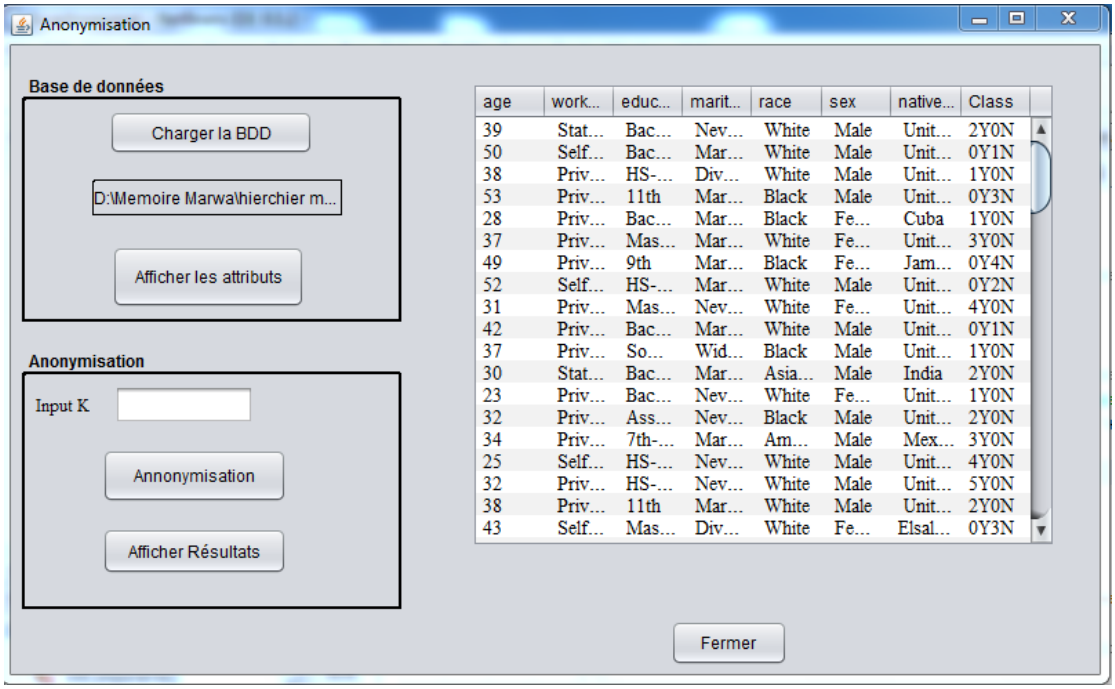


Figure IV.7: L'interface qui contenant le tableau de BDD

Afficher les hierarchies de généralisation

Après avoir sélectionné le tableau, nous cliquons sur bouton afficher les attributs

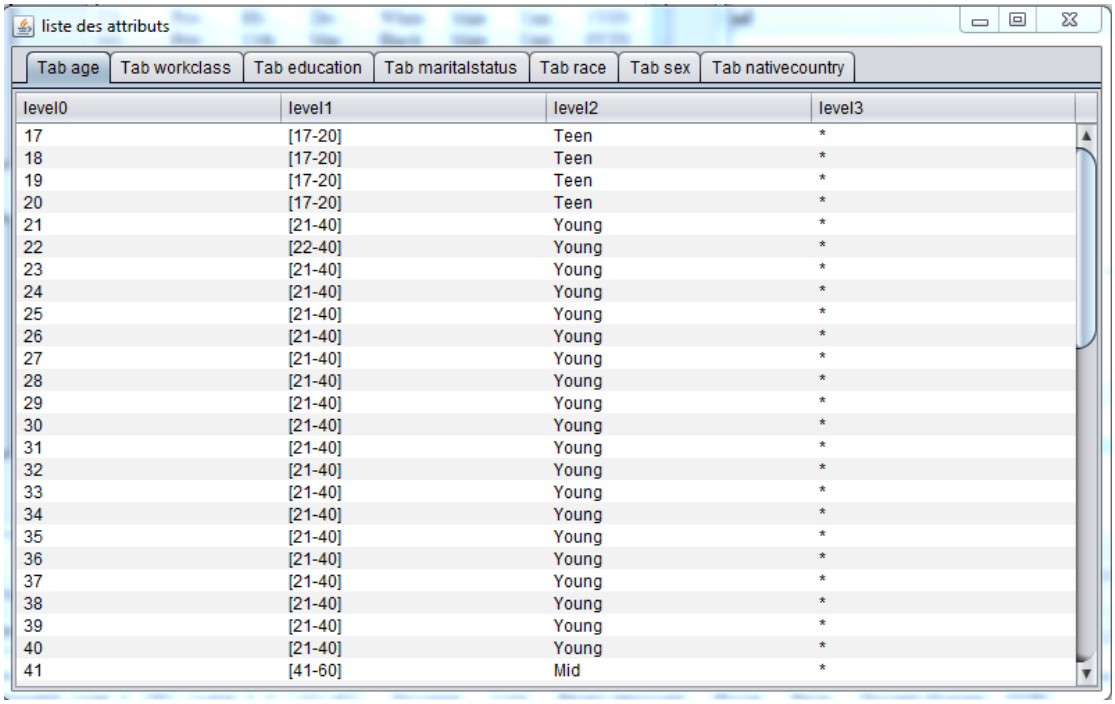


Figure IV.8: L'interface qui contenant la liste des attributs (hierarchie l'attribut age)

7.2. Choix de paremètre "K"

Après l'importation dataset et hiérarchie de généralisation, il faut d'abord saisir le paramètre "K" tq le k est le nombre d'enregistrements qui ont la même valeur de QID

7.3. Anonymisation avec l'algorithme

Une fois dataset et hierarchie de généralisation sont importées, l' anonymisation par l'algorithme Bottom up généralisation doit être effectuée.

7.3.1. Généralisation

La technique de généralisation consiste à remplacer les valeurs de données par des valeurs plus générales. Par conséquent, les données sont vraies mais moins précises. La généralisation est appliquée sur un quasi-identifiant. Il nécessite la définition d'une hiérarchie pour chaque attribut du QI. Chaque hiérarchie contient au moin deux niveaux. La racine est la valeur la plus générale. Il représente le niveau le plus élevé. Les feuilles correspondent aux valeurs de données d'origine et constituent le niveau le plus bas noté 0.

Exemple, l'arbre de la **Figure IV.9** a représente une hiérarchie de généralisation de l'attribut" éducation". Le nœud «Junior» est au niveau 1 de la hiérarchie. **Figure IV.10** est un exemple de hiérarchie pour l'attribut «âge» où ce dernier est généralisé par intervalles

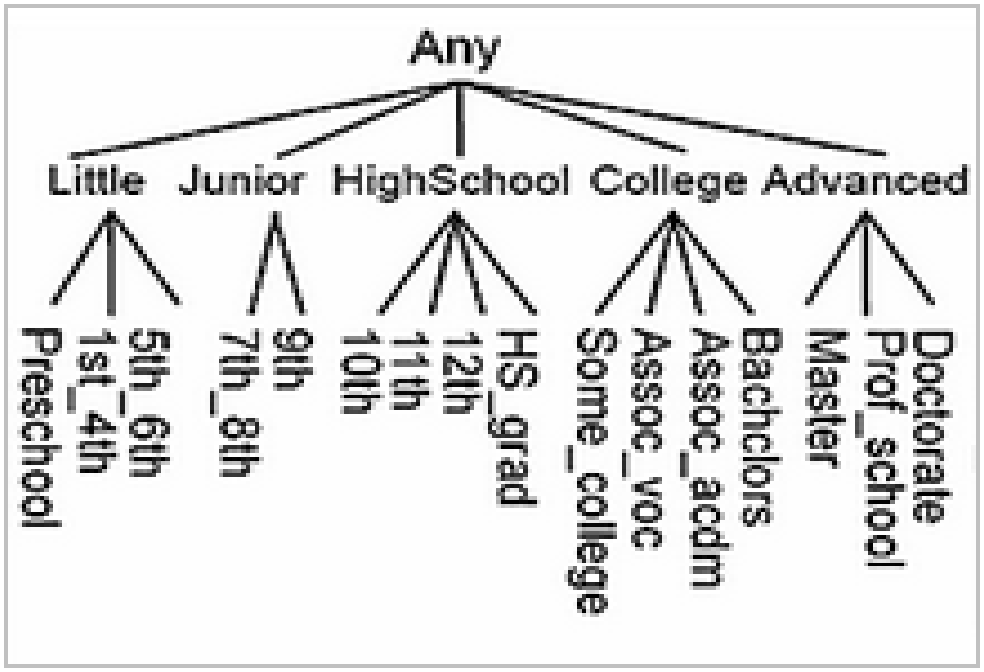


Figure IV.9: Hiérarchies de généralisation pour l'éducation

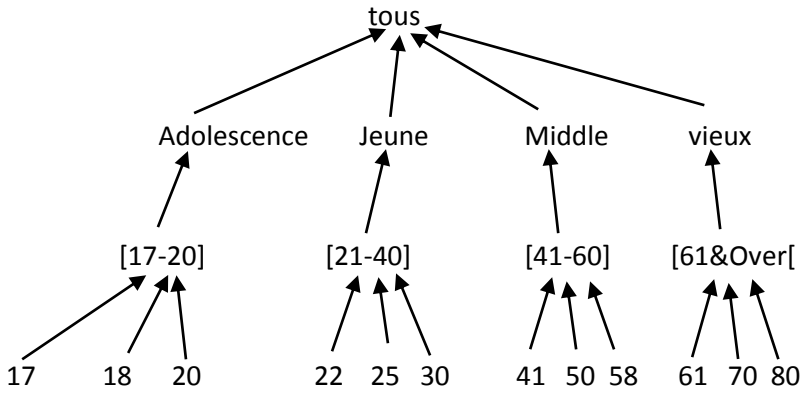


Figure IV.10: Hiérarchies de généralisation pour l'age

7.3.2. Métriques pour la généralisation

Nous considérons une métrique pour une seule généralisation, qui sert à guider la recherche d'une séquence de généralisation dans la section suivante. Une «bonne» généralisation doit préserver informations pour la classification et se concentrer sur l'objectif de atteindre le K-anonymat.

La formule permettant de calculer le score d'une généralisation G, noté IL/AG(G) est la suivante :(explication de cette formule est dans le **chapitre II**).

IL / AG (G) = { InformationLoss (G) / AnonymityGain (G) InformationLoss (G)

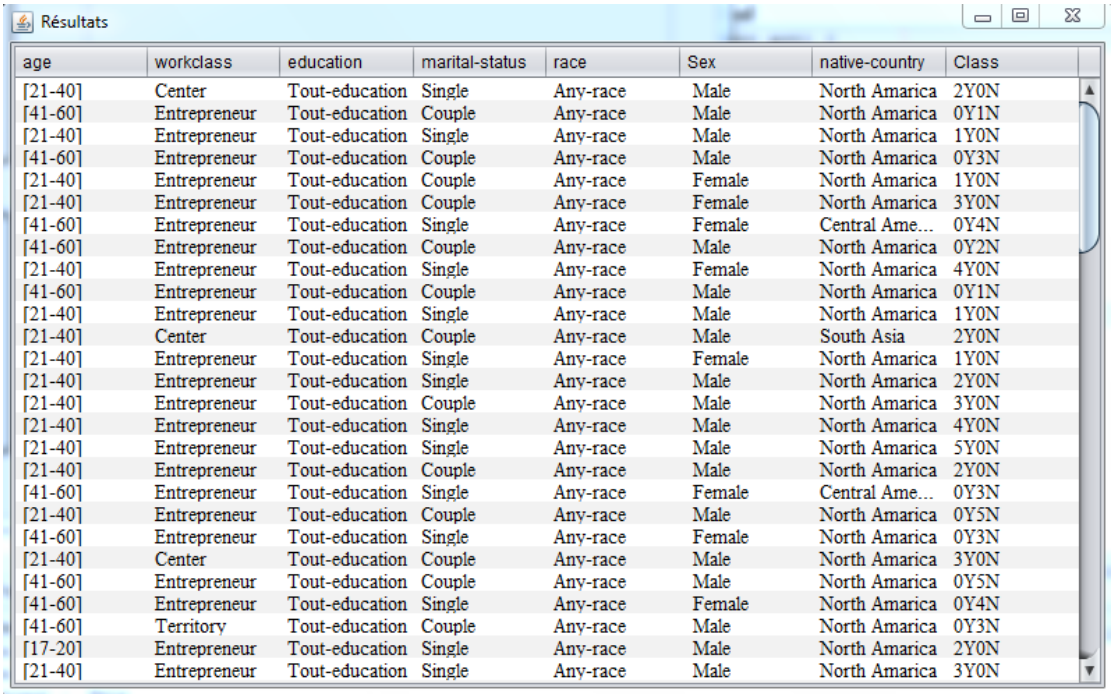
If AnonymityGain(G) ≠ 0

7.3.3. Trouver la meilleure généralisation

La généralisation est considérée comme bonne si elle renvoie le meilleur score (Min) calculé par application de la métrique de compromis IL/AG dont le rôle est de mesurer la perte d'information concernant la classification et le gain en sécurité liés à l'anonymisation.

8. Comparaison des résultats

Après avoir choisi le nombre de k , nous cliquons sur le bouton anonymisation cela fonctionne sur l'algorithme, et après son achèvement, un message apparaît, et à la fin nous cliquons sur le bouton afficher le résultat pour nous montrer le résultat final


Résultats

age	workclass	education	marital-status	race	Sex	native-country	Class
[21-40]	Center	Tout-education	Single	Any-race	Male	North America	2Y0N
[41-60]	Entrepreneur	Tout-education	Couple	Any-race	Male	North America	0Y1N
[21-40]	Entrepreneur	Tout-education	Single	Any-race	Male	North America	1Y0N
[41-60]	Entrepreneur	Tout-education	Couple	Any-race	Male	North America	0Y3N
[21-40]	Entrepreneur	Tout-education	Couple	Any-race	Female	North America	1Y0N
[21-40]	Entrepreneur	Tout-education	Couple	Any-race	Female	North America	3Y0N
[41-60]	Entrepreneur	Tout-education	Single	Any-race	Female	Central Ame...	0Y4N
[41-60]	Entrepreneur	Tout-education	Couple	Any-race	Male	North America	0Y2N
[21-40]	Entrepreneur	Tout-education	Single	Any-race	Female	North America	4Y0N
[41-60]	Entrepreneur	Tout-education	Couple	Any-race	Male	North America	0Y1N
[21-40]	Entrepreneur	Tout-education	Single	Any-race	Male	North America	1Y0N
[21-40]	Center	Tout-education	Couple	Any-race	Male	South Asia	2Y0N
[21-40]	Entrepreneur	Tout-education	Single	Any-race	Female	North America	1Y0N
[21-40]	Entrepreneur	Tout-education	Single	Any-race	Male	North America	2Y0N
[21-40]	Entrepreneur	Tout-education	Couple	Any-race	Male	North America	3Y0N
[21-40]	Entrepreneur	Tout-education	Single	Any-race	Male	North America	4Y0N
[21-40]	Entrepreneur	Tout-education	Single	Any-race	Male	North America	5Y0N
[21-40]	Entrepreneur	Tout-education	Couple	Any-race	Male	North America	2Y0N
[41-60]	Entrepreneur	Tout-education	Single	Any-race	Female	Central Ame...	0Y3N
[21-40]	Entrepreneur	Tout-education	Couple	Any-race	Male	North America	0Y5N
[41-60]	Entrepreneur	Tout-education	Single	Any-race	Female	North America	0Y3N
[21-40]	Center	Tout-education	Couple	Any-race	Male	North America	3Y0N
[41-60]	Entrepreneur	Tout-education	Couple	Any-race	Male	North America	0Y5N
[41-60]	Entrepreneur	Tout-education	Single	Any-race	Female	North America	0Y4N
[41-60]	Territory	Tout-education	Couple	Any-race	Male	North America	0Y3N
[17-20]	Entrepreneur	Tout-education	Single	Any-race	Male	North America	2Y0N
[21-40]	Entrepreneur	Tout-education	Single	Any-race	Male	North America	3Y0N

Figure IV.10: Résultat final de l'algorithme 'table qui satisfait 2 anonymat'

Après l'exécution de cette algorithme on a trouvé un tableau moins spécifique a partie de la table origine pour maximiser le gain de la confidentialité

9. Conclusion

Dans ce chapitre nous avons présenté la partie pratique de notre travail de ce fait nous avons présenté les détails l'approche Bottom up généralisation pour la préservation de la confidentialité des données publiées avec quelques résultats.

On peut dire que cette approche est efficace du fait que les performances issu des tables originales et générées sont très proches en termes confidentialité pour la table générée.

Conclusion générale

La préservation de la confidentialité pour les données est un droit fondamental dont la définition doit être adaptée à l'ère numérique. Les propriétaires de données se trouvent alors dans une situation qui exige de satisfaire deux buts contradictoires : respecter la confidentialité des données et, en même temps, préserver leur utilité.
Plusieurs techniques et algorithmes d'anonymisation des données sont proposés, lesquels modifient les données originales afin de minimiser les risques de ré-identification tout en sauvegardant autant que possible l'utilité de ces données. De plus, il est impossible de proposer un seul algorithme qui s'adapte à tous les contextes et qui donne le meilleur résultat à chaque fois .

Le choix du « bon » algorithme dépend d'un certain nombre de paramètres de contexte tels que les caractéristiques de la base de données, le besoin de l'anonymisation, etc. Plusieurs outils d'anonymisation existent afin d'offrir à l'éditeur de données la possibilité d'appliquer ces techniques d'anonymisation.

Bibliographie

[1] Salah Eddine KABOU ‹‹ La gestion de la confidentialité dans le Cloud Computing ››Thése de Doctorat LMD Université Djilali liabes-Sidi bel abbes 2017.

[2] Anas Abou El Kalam, Yves Deswarte, Gilles Trouessin, Emmanuel Cordonnier‹‹Une démarche méthodologique pour l'anonymisation de données personnelles sensibles››.

[3] http://www.benjamin-nguyen.fr/ consulté 27/7/2020.

[4] BENSIMESSAOUD Sihem ‹‹ Préservation de la confidentialité des informations personnelles dans la publication des réseaux sociaux ›› mémoire pour l'obtontion du diplôme de Magister Univesité A.Mira-Bejaia 29/09/2016.

[5] Asst.Prof.Ms. Apeksha Sakhare**,** Ms. Swati Ganar ‹‹ Anonymization: A Method To Protect Sensitive Data In Cloud ››International Journal of Scientific & Engineering Research, Department of Computer Science and Engineering G.H.Raisoni College of Engineering, Nagpur**.** May-2013.

[6] Korra Sathya Babu ‹‹Utility-Based Privacy Preserving Data Publishing ›› Une these Soumis en partie aux exigences du diplôme de doctorat en philosophie septembre 2013.

[7] Bhushan Mahajan, Swati Ganar ‹‹Review Pa,per on Preserving Confidentiality of Data in Cloud Using Dynamic Anonymization ›› Revue internationale d'informatique et de réseau, Volume 1, Issue 6, October 2012.

[8] Anas Abou El Kalam ‹‹ Modèles et politiques de securite pou les domaines de la sante et des affaires sociales ›› Thése de Doctorat , 04 décembre 2003 .

[9] Feten Ben Fredj ‹‹ Méthode et outil d'anonymisation des données sensibles ››.

[10] Feten Ben Fredj, Nadira Lammari, Isabelle Comyn-Wattiau ‹‹ Anonymisation de données par généralisation Méthode avec guidage ›› CEDRIC-CNAM, 2 rue Conté, 75003 Paris, ESSEC Business School, 1 av. Bernard Hirsch, 95021 Cergy.

[11] Vanessa Ayala-Rivera, Patrick McDonagh_, Thomas Cerqueus, Liam Murphy ‹‹ A Systematic Comparison and Evaluation of k-Anonymization Algorithms for Practitioners ›› École d'informatique et d'informatique, University College Dublin, Irlande., École d'ingénierie électronique, Dublin City University, Irlande, 2014.

[12] LATANYA SWEENEY ‹‹ Achieving *k*-anonymity privacy protection using generalization and suppression. ›› École d'informatique, Université Carnegie Mellon, Pittsburgh, Pennsylvanie,USA May 2002.

[13] Mlle.CHARIF Ismahan ‹‹ La protection de la vie privée sur Internet: Application sur les données personnelles ›› mémoire pour l'obtontion du diplôme de master, Université Abou Bakr Belkaid– Tlemcen, 01 Juillet 2013.

[14] LEMAY, Alain ‹‹ Infrastructure logicielle visant à protéger la confidentialité du patient dans les images médicales utilisées en recherche ›› mémoire présenté à l'école de technologie supérieure de Québec université comme exigence partielle àl'obtention de la matrise génie M.Ing, 17Mai 2009.

[15] Ashwin Machanavajjhala Johannes Gehrke Daniel Kifer, Muthuramakrishnan Venkitasubramaniam, ‹‹ ℓ-Diversity: Privacy Beyond k-Anonymity ››, Department of Computer Science, Cornell University.

[16] Ousseynou Sané,Fodé Camara, Samba Ndiaye, Yahya Slimani,Blocage ‹‹ des canaux d'inférences dans les données k-anonymes ››, Département mathématiques-informatique, Faculté des Sciences et Techniques,Université Cheikh Anta Diop de Dakar SENEGAL,Département d'informatique, Faculté des Sciences Université Tunis,TUNISIE, 2003.

[17] Amar Paul Singh ‹‹A Review of Privacy Preserving Data Publishing Technique ›› Article recherché, School of CSE Bahra University Shimla Hills, Inde, juin 2013.

[18] Ninghui Li Tiancheng Li Suresh Venkatasubramanian ‹‹*t*-Closeness: Privacy Beyond *k*-Anonymity and _-Diversity›› Département d'informatique, Université Purdue.

[19] Keerthana Rajendran , Manoj Jayabalan, Muhammad Ehsan Rana ‹‹A Study on k-anonymity, l-diversity, and t-closeness Techniques focusing Medical Data ›› Article, École d'informatique et de technologie Université de technologie et d'innovation Asie-Pacifique, Technology Park Malaysia 57000 Bukit Jalil, Kuala Lumpur, Malaisie décembre 2017.

[20] Feten Ben Fredj, Nadira Lammari, Isabelle Comyn-Wattiau ‹‹Characterizing Generalization Algorithms First Guidelines for Data Publishers ›› ESSEC Business School, 1 Av B. Hirsch, 95000 Cergy, France.

[21] Shweta S. Bhand, Prof. J .L. Chaudhari ‹‹The Wrapper Top-Down Specialization and Bottom-Up Generalization Approach for Data Anonymization Using Mapreduce on Hadoop›› P. G. Étudiant, Département de génie informatique, BSIOTR, Wagholi, Pune, Inde Professeur assistant Mai 2015.

[22] Priyanka Gawali, Dhananjay Gawali ‹‹Big data privacy preservation using K Anonymization and l-Diversity›› Department de CSE Dr. DYPSOET, Lohegaon Pune, India. New Art Commerce & Science College, Shevgaon,India,Novembre 2016.

[23] Ke Wang, Philip S. Yu, Sourav Chakraborty‹‹ Bottom-Up Generalization: A Data Mining Solution to Privacy Protection›› Simon Fraser University, IBM T. J. Watson Research Center .

[24] BENJAMIN C. M. FUNG, KE WANG, RUI CHEN, PHILIP S. YU ‹‹Privacy-Preserving Data Publishing: A Survey of Recent Developments ›› Université Concordia, Montréal Simon Fraser University, Burnaby University of Illinois at Chicago ACM Computing Surveys, Vol. 42, n ° 4, article 14, date de publication: juin 2010.

[25] Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang,Baile Shi, Ada Wai-Chee Fu, ‹‹Utility-Based Anonymization Using Local Recoding››1Université de Fudan.

[26]http://esbd.eu/fr/blog/cryptage-chiffrement-anonymisation-et-pseudonymisation-quelles-sont-les-differences, consulté le 11/13/2019 20:20.

[27] Divya Sadhwani, Dr. Sanjay Silakari, Mr. Uday Chourasia ‹‹ Preserving Privacy during Big Data Publishing using K-Anonymity Model – A Survey ›› rapport d'enquete département d'informatique & Engg. UIT-RGPV Bhopal, Volume 8, No. 5, May-June 2017.

[28] DJELLALBIA Amina‹‹Authentfication anonyme dans un envirennement clud ››mémoire pour l'obtontion du diplôme de Magister Univesité A.Mira-Bejaia 29/09/2016.

[29] Benjamin C. M. Fung ‹‹Privacy-Preserving Data Publishing ›› une thèse soumise dans le respect partiel des exigences pour le diplôme de docteur en philosophie à la School of Computing Science 2007 .

[30] Tinabo.R‹‹A Modified Anonymisation Algorithm Towards Reducing Information Loss›› Thése de Doctorat Université Technologie Dublin 2013.

[31] Kavitha S, Sivaraman E, Raja Vadhana P ‹‹ A Survey on k-Anonymity Generalization Algorithms ›› Revue internationale de recherche avancée en génie informatique et de la communication Vol. 3, Issue 11, November 2014.

[32] Xiaoxun Sun ,Min Li, Hua Wang, Ashley Plank ‹‹An effcient hash-based algorithm for minimal *k*-anonymity››Toowoomba, Queensland 4350, Australia Université du sud du Queensland.

[33] Rajeswari C, M.Asha Jerlin, Jayakumar Sadhasivam, Nithya.S ‹‹A case study on attack models andprivacy models in mining medical DATASETS››ArticleID:IJMET_08_11.

[34] X.Xiao Y.Tao, Anatomy: Simple and Effective Privacy Preservation[C],Proc.Intl Conf.Very Large Data Bases(VLDB), 2006

[35] http://netbeans.developpez.com/faq/?page=Introduction,consulté le 4/10/2020 .

[36] Seethal K S1, Siddana Gowda**. "**A Secure and Efficient Way of Accessing Encrypted Cloud Data bases Using Adaptive Encryption Scheme". International Journal of Science and Research, 2013.

[37] Warin, S. (2011, Février 07). Un livre blanc sur le cloud computing. Consulté le Novembre 01, 2011.

[38] Vaidya, J., Clifton, C, "Privacy-preserving data mining": Why, how, and when. IEEE Security & Privacy, 2004.

[39] www.who-int/gender/what is gender/fr/ consulté 7/10/2020.

[40] www.wikipedia.com consulté 7/10/2020.