Table des matières

introduction général

Avec l'augmentation de la puissance et de l'inter-connectivité des systèmes informatiques disponibles aujourd'hui qui offre la possibilité de stocker et de traiter de grandes quantités de données, ce qui permet d'accéder à des informations en réseau à partir de n'importe où à tout moment. À partir de ça, il est devenu possible d'exploitation des données personnelles des personnes a profondément modifier la stratégie d'attaques des données ainsi que leur mode de vol. Cela a soulevé des préoccupations universelles sur la protection de la vie privée des individus.

L'objectif de ce mémoire est d'augmenter la protection des Micro-données et améliorer la confidentialité des informations personnelles. L'idée de base est utiliser une approche qui se base sur la génération aléatoire des nouvelles données à partir des données originales. La génération aléatoire est faite en utilisant un classificateur automatique pour garder un maximum de sens entre les valeurs des attributs générés ,en plus ,cette génération est guidée par un ensemble de règles sémantiques qui rendent les données générées plus proches des données réelles. Les nouvelles données diffèrent totalement des données originales ce qui implique une grande protection

Afin d'aborder tous les aspects ayant une relation avec la protection de la vie privée et les micro-données, le travail est organisé comme suit :

le chapitre **1** explique de façon général une vue sur la vie privée, ses concept et les différents niveaux de protection et des risques relatifs à la vie privée sur Internet ainsi que quelques techniques d'attaques contre la vie privée .

Le chapitre **2** est consacré à un état de l'art sur la protection des micro-données .Pour cela nous présentons les différents risques de perte des micro-données et les techniques de protection de ses différents risques aussi on a terminé par quelques approches de protection.

le chapitre **3** introduit notre approche pour protéger les Micro-données. En fin nous terminons par une conclusion générale qui résume notre travail et présente quelques perspectives.

CHAPITRE I : la vie privée sur internet

I. Introduction

Avec un nombre d'utilisateurs qui ne cesse d'augmenter. Internet est devenu un moyen d'expression, de communication, d'information et de connaissance révolutionnaire.

Depuis son apparition, l'Internet a bouleversé les rôles et les structures sociales jusqu'alors bien établis. Alors que le géant Google a transformé l'accès à l'information de différentes façons (accessibilité, rapidité et réseautage), les réseaux sociaux sont devenus les principaux moyens de médiation et de relation entre les individus. Cela augmenté et favorisé la capacité des hommes à travailler ensemble de façon plus efficace et plus étendue.

Malheureusement tout cela n'a pas que des avantages, la grande masse d'information circulée sur Internet ouvre l'appétit de pas mal de gens et d'entreprises, qui utilisent ces informations hors de leur contexte légitime, touchant directement à la vie privée des personnes.

La vie privée sur l'internet relève des éléments propres à un individu qui sont considérés comme personnels, et dont l'accès au public n'est pas admis. Ces éléments correspondent à des informations permettraient l'identification directe ou indirecte de l'individu et sur lesquelles on veut garder le contrôle. Les données qui appartiennent aux domaines de l'identité (nom, prénom, âge, sexe, lieu de résidence, etc.), des activités (loisirs préférés, numéro de client, de carte bancaire, etc.), de la santé, de la vie sentimentale, conjugale ou familiale font partie des éléments de la vie privée. La notion de vie privée n'est pas uniquement un concept relationnel, mais un droit légal pour tous.

Dans ce chapitre on présente une vue générale sur la vie privée pour cela on commence par une définition avec les principes et les différents niveaux de la vie privée ainsi que quelques attaques, risque et enfin les technologies permettant la protection de la vie privée.

II. Définition de la vie privée :

La vie privée est la capacité, pour une personne ou pour un groupe de personnes, de s'isoler afin de protéger ses intérêts. Les limites de la vie privée ainsi que ce qui est considéré comme privé diffèrent selon les groupes, les cultures et les individus, selon les coutumes et les traditions bien qu'il existe toujours un certain tronc commun.

La vie privée peut parfois s'apparenter à l'anonymat et à la volonté de rester hors de la vie publique. Quand quelque chose est dit "privé" pour une personne, cela signifie que généralement qu'à cette chose sont rattachés des sentiments spéciaux et personnels. Le degré de privatisation de l'information dépend donc de la façon dont le public pourrait la recevoir, ce qui diffère selon les endroits et à travers le temps. La vie privée peut être vue sous un aspect sécuritaire[47].

III. le concept de la vie privée :

Le contenu de la vie privée est variable selon les circonstances, les personnes concernées et les valeurs d'une société ou d'une communauté. Généralement, la vie privée englobe la vie personnelle (identité, origine raciale, santé...) avec le secret professionnel, le secret médical, la protection de l'identité et de l'image et la protection de la correspondance et la réglementation des écoutes téléphoniques, la vie familiale, conjugale ou sentimentale, le domicile [1].

La vie privée sur internet est une notion plus importante que celle habituellement admise dans la vie de tous les jours. Il est primordial de bien comprendre que toute information non sécurisée mise en ligne peut être accessible par tout le monde.

IV. la caractérisation des données personnelles :

On appelle données personnelles les informations qui permettent, notamment sur Internet, d'identifier directement ou indirectement une personne physique [3].

Les données personnelles (ou nominatives) correspondent généralement aux nom, prénom, adresse électronique, numéro de téléphone, date de naissance, etc. qu'un individu peut transmettre par courrier électronique, inscrire sur un formulaire en ligne ou sur un site Web.

Le responsable du fichier ou du traitement de données personnelles doit informer les personnes concernées du but de ce traitement, de l'identité des destinataires de ces

informations. Les données de santé ne peuvent être collectées que dans certains cas bien précis, encadrés par la loi, par exemple pour le dossier médical informatisé d'un patient hospitalisé [3].

V. Les niveaux de protection de vie privée [55] :

La protection de la vie privée prend une importance de plus en plus grande. voici les quatre communs de la protection de la vie privée sur l'Internet:

V.1. Anonymat:

L'anonymat est le fait de ne pas mentionner le nom des personnes. Lors d'une étude, on rend anonyme les informations personnelles et les données. Par ex : Nom Prénom le nom du participant figure seulement sur un fichier de correspondance avec son code. Le code est ensuite utilisé lors de l'expérience et du traitement des données. En un mot il est impossible (pour d'autres utilisateurs) de déterminer le véritable nom de l'utilisateur associé à un sujet, une opération, un objet.

V.2. Pseudonymat:

La Pseudonymat consiste à supprimer les champs directement identifiants des enregistrements, et à rajouter à chaque enregistrement 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 problèmes.

V.3. Non-chaînabilité:

La non-chainabilité garantit qu'un utilisateur peut utiliser plusieurs fois des ressources ou des services sans que d'autres soient capables d'établir un lien entre ces utilisations. La non-chainabilité rend toute entité incapable de relier deux actions anonymes qui ont été menée par le même individu. Aussi elle rend impossible (pour d'autres utilisateurs) d'établir un lien entre différentes opérations faites par un même utilisateur.

V.4. Non-observabilité:

La Non-observabilité garantit qu'un utilisateur peut utiliser une ressource ou un service sans que d'autres, en particulier des tierces parties, soient capables d'observer que la ressource ou le service est en cours d'utilisation. Il est impossible (pour d'autres utilisateurs) de déterminer si une opération est en cours.

VI. Les risques relatifs à la vie privée sur Internet :

Les besoins de protection de la vie privée peuvent être d'ordre général, et surtout spécifiques au système étudie. Dans les systèmes de santé par exemple, une liste non exhaustive d'informations à rendre anonymes pourrait être : outre le nom, le prénom et le numéro de sécurité social, les données les plus sensibles sont la date de naissance (parfois, seulement l'année de naissance est nécessaire), l'adresse (parfois, seulement la région est intéressante à connaître), et parfois la nationalité.

Les données personnelles dont il faut garder secrètes correspondent non seulement aux données directement nominatives (comme le nom, le prénom, le numéro de sécurité sociale, le sexe et l'adresse), mais aussi aux données indirectement nominatives. En effet, il est souvent possible d'identifier un individu par un simple rapprochement de données personnelles de nature médicale ou sociale. Par exemple, l'âge, le sexe et le mois de sortie de l'hôpital, permettent d'isoler le patient dans une population restreinte et même dans certain cas l'identifier d'une manière précise.

VII. Les attaques sur la vie privée :

VII.1. Le Vol d'identité :

Il y a vol d'identité quand un individu s'empare de vos renseignements personnels et se fait passer pour vous à des fins frauduleuses.

Chaque année, des milliers de personnes sont victimes de ce type de vol. Vos nom, date de naissance, adresse, numéro de carte de crédit, numéro d'assurance sociale (NAS) et autres numéros d'identification personnels peuvent servir à obtenir une carte de crédit, à ouvrir un compte bancaire, à réacheminer le courrier, à établir un service de téléphone cellulaire, à louer un véhicule, de l'équipement ou une chambre ou même à obtenir un emploi. Ces brèches d'anonymat sont appelées : record linkage, le phishing, le profiling et l'attaque probabiliste.

VII.2. L'attaque par « Phishing» :

Le phishing est une technique d'escroquerie relativement simple : il consiste à placer des liens piégés dans de faux e-mails imitant les messages d'organismes ou d'institutions diverses. Vous croyez vous rendre sur le site officiel d'un l'établissement, mais vous atterrissez en fait sur une copie, plus ou moins bien imitée. Toute information confidentielle que vous tapez alors – comme votre mot de passe – est immédiatement récupérée par les escrocs. Ces derniers n'ont alors plus qu'à se connecter sur le (vrai) site de votre banque ou de votre messagerie, pour avoir accès à votre compte [52].

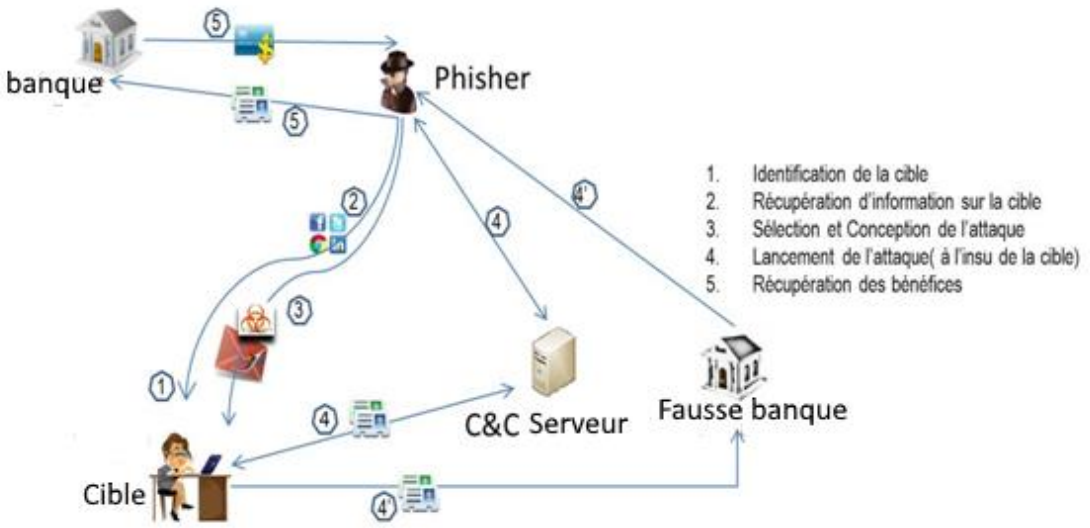

Figure I.1 le processus d'attaque par phishing**[53].**

VII.3. L'attaque par «Profiling » :

Chaque fois que l'utilisateur visite un site Web, quelqu'un, quelque part suit son activité en ligne. Ce profilage permet de recueillir des renseignements détaillés sur l'internaute et se pratique sur de nombreux sites, souvent à l'insu du visiteur ou sans son consentement, et cela présente un risque d'atteinte à la vie privée du fait qu'il permet d'analyser avec précision le comportement des consommateurs [4].

Le profilage est une technique de surveillance ou d'exploitation des données qui permet d'établis différentes actions, mesures ou décisions touchant les personnes concernées dans le cadre de finalités diverses. Les techniques de profilage représentent un intérêt important pour l'économie ou pour les administrations publiques; elles peuvent aussi avoir des effets bénéfiques pour les personnes concernées, par exemple dans le domaine de la santé.

Cependant, elles génèrent également des conséquences négatives sur le respect des droits et des libertés fondamentales, notamment le droit à la vie privée et à la protection des données [5].

VII.4. Les attaques sur micro-données :

VII.4.1. Principe d'attaque par « Attribute linkage » :

Si certaines valeurs de l'attribut sensible sont prédominantes dans une classe d'équivalence (exp : un groupe d'enregistrements ayant les mêmes valeurs de quasi identifiants), un adversaire n'aurait pas de difficultés à les relier aux individus en question dans ce groupe. De telles attaques sont appelées "attribute linkage". Vu cette vulnérabilité, plusieurs modèles ont été définis pour combattre les attaques par "attribute linkage". Parmi ces modèles, nous pouvons citer l-diversité (se trouve dans le chapitre 2 section VII.2). **[48].**

VII.4.2. Principe d'attaque par « Table linkage » :

Un Table Linkage se produit si un attaquant peut en toute confiance déduire la présence ou l'absence de l'enregistrement de la victime dans le tableau publié.

Le modèle δ-Présence a été proposé pour combattre les attaques par "Table linkage" **[49].**

VII.4.3. Principe d'attaque par « Record linkage » :

Cette attaque est possible quand certaines valeurs q de quasi-identifiants Q (attributs non identificateurs, mais si utilisés ensembles capable d'identifier un individu) identifient un petit nombre d'enregistrements dans T, le micro data à révéler. Dans ce cas, l'individu possédant la valeur q est susceptible d'être lié à un petit nombre de d'enregistrements dans T. Le modèle k-anonymat (voir le chapitre 2 section VII.1) a été proposé pour combattre les attaques par "record linkage".

La garantie obtenue avec celui-ci est qu'aucune information ne pourrait être liée à un groupe d'au moins k individus. Ainsi, le degré d'incertitude de l'attribut sensible est au moins égal à 1=k. Toutefois le principal inconvénient de ce modèle est sa vulnérabilité aux attaques de type "attribute linkage" **[48].**

VII.4.4. Principe d'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.

Le modèle « confidentialité différentielle (voir chapitre 2 section VII.4.)» a été proposé pour combattre les attaques "Probabilistes" **[49].**

VIII. Les techniques d'attaques :

VIII.1. TCP Session Hijacking :

Le vol de session (session hijacking en anglais) est une attaque qui consiste à s'introduire dans une session existante entre deux hôtes et à intervenir en cours de communication en se faisant passer pour l'un d'entre eux.

Le vol de session peut être effectué sur le réseau ou sur l'hôte directement. Ce vol déroule en trois étape :

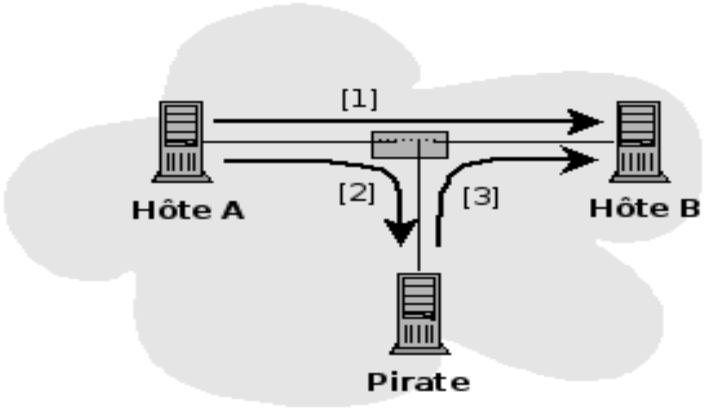

Figure I.2 : Tcp session hijacking technique [6].

1. Un hôte A établit une connexion (par exemple Telnet) avec un hôte B.
2. Un attaquant sur le même réseau sniffe la connexion
3. L'attaquant envoie à l'hôte B des paquets avec l'adresse source de l'hôte A et avec des numéros de séquence appropriés. Le pirate a volé la session de l'hôte A.[6]

VIII.2. L'attaque par « Injection de commandes SQL» :

De nombreux développeurs web ne sont pas conscients des possibilités de manipulation des requêtes SQL, et supposent que les requêtes SQL sont des commandes sûres. Cela signifie qu'une requête SQL est capable de contourner les contrôles et vérifications, comme les identifications, et parfois, les requêtes SQL ont accès aux commandes d'administration.

L'injection SQL directe est une technique où un pirate modifie une requête SQL existante pour afficher des données cachées, ou pour écraser des valeurs importantes, ou encore exécuter des commandes dangereuses pour la base. Cela se fait lorsque l'application prend les données envoyées par l'internaute, et l'utilise directement pour construire une requête SQL. Les exemples ci-dessous sont basés sur une histoire vraie, malheureusement.

Avec le manque de vérification des données de l'internaute et la connexion au serveur avec des droits de super utilisateur, le pirate peut créer des utilisateurs, et créer un autre super utilisateur

Par exemple la Séparation des résultats en pages, et créer des administrateurs (PostgreSQL et MySQL)

```
<?php
$offset = $argv[0]; // Attention, aucune validation!
$query  = "SELECT id, nom FROM produits ORDER BY name LIMIT 20 OFF
SET $offset;";
$result = pg_query($conn, $query);
?>
```

Un utilisateur normal clique sur les boutons 'suivant' et 'précédent', qui sont alors placés dans la variable **$offset**, encodée dans l'URL. Le script s'attend à ce que la variable **$offset** soit alors un nombre décimal. Cependant, il est possible de modifier l'URL en ajoutant une nouvelle valeur, au format URL, comme ceci : Exemple d'injection SQL

```
0;

insert into pg_shadow(usename,usesysid,usesuper,usecatupd,passwd)

    select 'crack', usesysid, 't','t','crack'

    from pg_shadow where usename='postgres';

-- 0;
```

Si cela arrive, le script va créer un nouveau super utilisateur. Notez que la valeur 0; sert à terminer la requête originale et la terminer correctement. [7]

VIII.3. L'attaque par « Botnet» :

Les botnets, ou réseaux d'ordinateurs zombies, constituent l'un des premiers outils de la délinquance sur Internet aujourd'hui. La création d'un botnet consiste à prendre le contrôle d'un maximum de systèmes informatiques connectés à Internet, par la diffusion d'un logiciel malveillant qui se connecte à un système de commande, placé sous le contrôle du malfaiteur.

Ces systèmes de commande peuvent être de nature différente, mais le plus souvent il s'agira de l'utilisation du protocole HTTP (celui du Web) ou IRC (Internet relay chat, protocole permettant la discussion ou l'échange de fichiers). Celui qui contrôle un tel réseau est traditionnellement appelé « pasteur ». Par la suite, l'ensemble de ces machines, lorsqu'elles sont connectées à Internet, répondent aux directives de

leur « pasteur » à l'insu de leur utilisateur légitime et peuvent être utilisées pour conduire

des attaques en déni de service distribué, la distribution de contenus illicites ou malveillants, la diffusion de courriers électroniques non sollicités (ou « spam »), la collecte des données personnelles des usagers de ces machines, du calcul distribué ou toute autre activité qu'il souhaitera.[8]

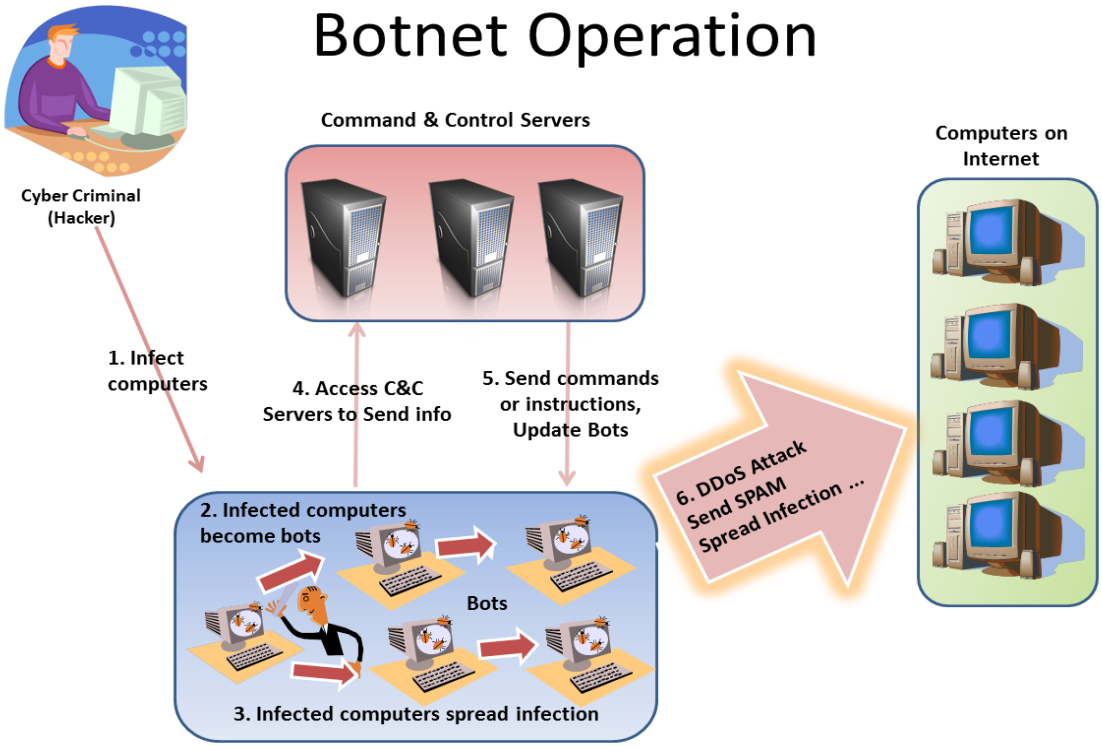

Figure I.3 : botnet technique d'attaque**[54].**

VIII.4. Les cookies (biscuits empoisonnées) :

Un cookie (ou témoin de connexion)est défini par le protocole de communication HTTP comme étant une suite d'informations envoyée par un serveur HTTP à un client HTTP, que ce dernier retourne lors de chaque interrogation du même serveur HTTP sous certaines conditions. Il est envoyé en tant qu'en-tête HTTP par le serveur web au navigateur web qui le renvoie inchangé à chaque fois qu'il accède au serveur.

Un cookie peut être utilisé pour une authentification, une session (maintenance d'état), et pour stocker une information spécifique sur l'utilisateur, comme les

préférences d'un site ou le contenu d'un panier d'achat électronique. Le terme cookie est dérivé de magic cookie, un concept bien connu dans l'informatique d'UNIX, qui a inspiré l'idée et le nom des cookies de navigation. Quelques alternatives aux cookies existent, chacune à ses propres utilisations, avantages et inconvénients. Étant de simples fichiers de texte, les cookies ne sont pas exécutables. Ils ne sont ni des logiciels espions ni des virus, bien que des cookies provenant de certains sites soient détectés par plusieurs logiciels antivirus parce qu'ils permettent aux utilisateurs d'être suivis quand ils ont visité plusieurs sites. La plupart des navigateurs récents permettent aux utilisateurs de décider s'ils acceptent ou rejettent les cookies. Les utilisateurs peuvent aussi choisir la durée de stockage des cookies.

Toutefois, le rejet complet des cookies rend certains sites inutilisables. Par exemple, les paniers d'achat de magasins ou les sites qui exigent une connexion à l'aide d'identifiants (utilisateur et mot de passe) [5].

VIII.5. Chevaux de Troie :

Un cheval de Troie (appelés également « trojan ») est un type de virus qui prétend être quelque chose d'utile, d'agréable ou d'amusant alors qu'en réalité il provoque des dommages ou vole des données.

Les chevaux de Troie sont souvent diffusés au travers d'une pièce jointe infectée ou d'un téléchargement qui se cache dans des jeux, applications, films ou cartes de vœux gratuits.[9] Lorsqu'il est exécuté, il peut effacer des répertoires ou ouvrir une « porte arrière » à l'ordinateur, permettant ainsi à quelqu'un de s'y introduire et de prendre le contrôle de système. Ces intrus peuvent alors copier et effacer les dossiers, utiliser l'ordinateur comme point de départ pour pirater d'autres compagnies

IX. Conclusion :

Nous avons présenté dans ce chapitre des notions théoriques sur la vie privée sur Internet, en commençant par une définition avec les principes et les différents risques relatifs à la vie privée ainsi que quelques attaques et les techniques associés.

Le chapitre suivant sera consacré à un état de l'art sur la protection des Micro-Données.

CHAPITRE II : La Protection des Micro-donnée

I. Introduction :

L'augmentation de la puissance et de l'inter-connectivité des systèmes informatiques disponibles aujourd'hui offre la possibilité de stocker et de traiter de grandes quantités de données, ce qui permet d'accéder à des informations en réseau à partir de n'importe où à tout moment. Ce processus de partage et de diffusion de l'information est clairement sélectif. En effet, si, d'une part, il est nécessaire de diffuser des données, il existe d'autre part un besoin aussi important de protéger ces données qui, pour diverses raisons, ne devraient pas être divulguées. Considérons, par exemple, le cas d'une organisation privée mettant à disposition diverses données concernant ses activités (produits, ventes, etc.), mais en même temps voulant protéger des informations plus sensibles, comme l'identité de ses clients ou des plans pour Produits futurs. À titre d'exemple, les organismes gouvernementaux, lors de la publication des données historiques, Peut exiger un processus de désinfection pour "annuler" les informations jugées sensibles, directement ou en raison des informations sensibles qu'il permettrait au destinataire de déduire. Le partage et la diffusion efficaces de l'information peuvent avoir lieu uniquement si le détenteur de données a l'assurance que tout en libérant l'information, la divulgation d'informations sensibles n'est pas un risque.

II. Définition sur les Micro-données :

Les micro-données sont définies comme étant l'ensemble des données contenant des informations sur des répondants individuels à une enquête. Les individus peuvent être des ménages ou bien appartiennent à des organisations telles que des écoles, des hôpitaux ou des entreprises.

Les réponses à une enquête nationale sur la santé de la population est un bon exemple des micro-données. L'utilisation et le partage de ce type de données est devenu une nécessité dans plusieurs domaines telles que la santé, l'administration, l'économie ainsi que la recherche et l'enseignement universitaire[46].

III. Les risques de perte des Micro-Données :

La performance de toute technique de protection est généralement mesurée en termes de perte d'information et de risque de divulgation , la perte d'information est la quantité d'informations qui existe dans les micro-données d'origine et en raison de la technique de protection ne se produit pas dans les micro-données protégées , le risque de divulgation est le risque qu'une divulgation se produise si des micro-données protégées sont diffusées , dans ce qui suit, nous décrivons certaines des méthodes les plus importantes utilisées pour quantifier le risque de divulgation et la perte d'information[46].

III.1. Risque de divulgation :

En général, il existe deux types de divulgation: la divulgation d'identité et la divulgation d'attributs [11]. La divulgation d'identité signifie qu'une identité spécifique peut être liée à un tuple dans la table des micro-données. La divulgation d'attribut signifie que des informations ont été divulguées sur un attribut d'un individu. En général, deux facteurs peuvent avoir une incidence sur la divulgation d'identité:

- **L'unicité de population signifie :** Que la probabilité d'identifier un répondant qui est le répondant unique avec une combinaison spécifique d'attributs est élevée si ces attributs sont présents dans la table des micro-données
- **Ré identification :** Signifie que les micro-données publiées sont liées à un autre tableau publié, où les identifiants n'ont pas été supprimés.

Différentes méthodes ont été proposées pour mesurer le risque de divulgation des micro-données publiées, par exemple, la combinaison minimale non sécurisée d'attributs [49] , renvoie le nombre d'attributs avec une combinaison unique dans un tuple de micro-données spécifique. Cette méthode ne peut être adoptée qu'avec des techniques de masquage non perturbatrices et, plus cette valeur est élevée, plus le risque de divulgation est faible. Nous concentrons sur les principales méthodes de mesure du risque de divulgation d'identité, qui sont l'unicité et la liaison des enregistrements, et la méthode principale pour mesurer la divulgation d'attribut qui est la divulgation d'intervalle [12, 13].

a) Record Linkage :

Record linkage Consiste à trouver une correspondance entre un tuple dans la table de micro-données protégées et un tuple dans une source d'information publique et non anonyme externe (par exemple, une liste d'électeurs qui contient le registre de tous les électeurs d'une région ou d'une ville). Comme il n'est pas possible de connaître a priori toutes les sources externes d'informations pouvant être utilisées par un éventuel utilisateur malveillant, une vérification probabiliste des micro-données protégées est effectuée , différent Record linkage méthodes doivent être adoptées selon que la table des micro-données et les informations externes présentent ou non des attributs communs. S'il existe des attributs communs, il est nécessaire d'adopter une représentation unique pour les attributs communs. Par exemple, les abréviations différentes au nom d'une personne conduiraient à la conclusion que deux tuples ne sont pas liés, alors qu'ils se réfèrent réellement au même répondant. Il est alors possible d'adopter une stratégie pour Record linkage [13, 14, 15]. Record linkage méthodes peuvent être divisées en trois grandes catégories: déterministes, probabilistes et à distance.

- **Déterministe.** Il recherche une correspondance exacte sur un ou plusieurs attributs entre les tuples dans différents ensembles de données. L'inconvénient principal de cette méthode Est-ce qu'il ne prend pas en compte la pertinence des attributs pour trouver un lien.
- **Probabiliste** Donné Deux Ensembles de données $D1$ et $D2$, chaque tuple d 1i $\in$ D1 adapté au tuple le plus proche $d2j \in D2$. Cette méthode nécessite la définition d'une fonction de distance $\mathcal{F}$ entre couples de tuples. Par exemple, la définition de distance $\mathcal{F}$ peut exploiter les fonctions de distance définies sur les attributs et peut affecter différents poids à chaque attribut, selon son importance dans le processus de liaison. Un exemple de fonction de distance est la distance euclidienne qui considère chaque tuple comme vecteur et attribue le même poids à chaque attribut. Ce Record linkage méthode n'est pas adaptée aux attributs catégoriels, car il est difficile de définir la distance entre deux catégories, en particulier si leur domaine n'est pas commandé.

Notez bien que Record linkage est considéré comme une menace, il existe de nombreuses situations où il peut être utile. Record linkage Peut être utilisé dans la gestion de grandes bases de données pour extraire des informations importantes sur le même sujet. Ceci est particulièrement utile lorsque les données sont distribuées sur différents serveurs (par exemple, l'information médicale de la population est généralement distribuée sur différents systèmes et Record linkage technique peut être exploitée pour reconstruire les informations associées à un individu donné [16].

b) Divulgation Intervalle :

La mesure d'information d'intervalle est calculée de différentes manières, en fonction du type de données de l'attribut (continu ou catégorique). Dans le cas d'un attribut catégorique, pour chaque tuple dans la table des micro-données, les intervalles classés sont construits comme suit. Chaque attribut est classé indépendamment et un intervalle de rang est déduit autour de la valeur assumée par l'attribut dans chaque tuple t . Les rangs des valeurs dans l'intervalle construit autour du tuple t devraient être inférieurs à P%, du nombre total de tuples aussi, le rang dans le centre de l'intervalle devrait correspondre à la valeur assumée par l'attribut considéré dans le tuple t.

Le risque de divulgation est alors la proportion des valeurs initiales qui se situent dans l'intervalle centré autour de la valeur protégée correspondante. Si une telle proportion est égale à 100%, un attaquant potentiel est sûr que la valeur d'origine réside dans l'intervalle autour de la valeur protégée. En cas de données continues, la méthode est similaire à la précédente.

La différence principale est la façon dont les intervalles classés sont construits: il n'est pas possible d'exploiter le classement et la construction est basée sur l'écart type de l'attribut[46].

III.2. Perte d'information :

La mesure de perte d'information est strictement liée à l'objet pour lequel les informations seront utilisées. Étant donné que les objectifs peuvent être différents et ne sont pas connus a priori, il n'est pas possible d'établir une mesure générale de perte d'information fondée sur l'objet. Les méthodes utilisées sont donc basées sur les concepts analytiquement valides et analytiquement intéressants, qui sont déduits comme suit [17].

- Une table de micro-données protégées est **analytiquement valide** si elle préserve approximativement l'analyse statistique (par exemple, la moyenne et la covariance) qui peut être produite avec les micro-données d'origine;
- Une table de micro-données protégées est **analytiquement intéressante** si elle contient un nombre suffisant d'attributs pouvant être analysés de manière valide.

En général, il existe deux stratégies pour calculer la perte d'information:
i) comparer directement les tuples des micro-données protégées avec les tuples dans les micro-données d'origine;
ii) comparer les statistiques calculées sur les micro-données protégées avec les mêmes statistiques évaluées sur les micro-données d'origine.

Nous décrivons maintenant l'idée fondamentale de certaines des mesures de perte d'information les plus courantes qui sont divisées en deux catégories en fonction du type de données des attributs. D'autres méthodes ont été proposées, à la fois pour des techniques spécifiques de protection des micro-données et pour les cas génériques [18, 19].

a) Données continues :

Pour mesurer la perte d'information, la statistique d'intérêt (par exemple, les matrices de co-variance, les matrices de corrélation ou leurs variantes) est évaluée à la fois sur les données originales et protégées, la différence entre les deux valeurs est calculée. Les écarts entre les deux statistiques peuvent être évalués de trois façons différentes: **erreur carrée moyenne**, **erreur absolue moyenne** et **variation moyenne**. En plus des mesures statistiques, les données peuvent être comparées,

avant et après l'application d'une technique de protection des micro-données, en calculant à nouveau la différence en utilisant l'une des trois méthodes susmentionnées. Il est important de noter que la valeur de la perte d'information devrait avoir une valeur maximale (par exemple, 100 si une notation de pourcentage est utilisée) pour comparer différentes méthodes ayant la même échelle pour le calcul de perte d'information [12, 20, 21, 22].

b) Données catégoriques :

Les mesures de perte d'information introduites brièvement pour les attributs continus ne sont pas directement applicables aux attributs catégoriques. Dans ce cas, il existe trois mesures principales [20]: **comparaison directe**, **comparaison des tableaux de contingence** et **mesure d'entropie**.

La comparaison directe des valeurs des attributs catégoriques nécessite la définition d'une fonction de distance entre les catégories. Dans le cas de catégories non commandées, la distance entre la catégorie c_1 Dans les micro-données d'origine et la catégorie correspondante c_2 Dans les micro-données protégées est Égal à 0, si les deux catégories sont identiques.1, sinon. Par contre, S'il y a un ordre entre les catégories, la distance entre les catégories c_1 et c_2 Est égal au nombre de catégories entre c_1 et c_2 Divisé par le nombre total de catégories.

La mesure de comparaison des tableaux de contingence consiste à comparer les tableaux de contingence correspondants. Une mesure basée sur l'entropie [23, 24] peut être utilisée chaque fois qu'une table de micro-données a été protégée en appliquant la suppression locale, le recodage global, Ou PRAM techniques. L'idée est que la perte d'information peut être mesurée à l'aide de Shannon Entropie car le processus de masquage est modélisé comme le bruit ajouté aux micro-données d'origine lorsqu'ils sont transmis par un canal bruyant. La mesure de perte d'information utilise la probabilité conditionnelle (la probabilité d'une valeur dans les micro-données d'origine, une fois que la valeur dans les micro-données protégées est donnée).

IV. Combinaison de risque de divulgation et de perte d'information :

Dans ce chapitre ont un impact différent sur l'utilité des données et le risque de divulgation. Pour pouvoir évaluer les techniques alternatives de protection des micro-données, nous avons d'abord besoin d'un cadre pour évaluer la qualité d'une technique de protection.

Le risque de divulgation et la perte d'information doivent donc être combinés. Une méthode simple consiste à calculer la moyenne des 2 valeurs et à choisir la technique (et le paramètre) qui a la valeur de score la plus élevée [21]. Une autre méthode est la carte de confidentialité R-U [25], qui est un graphique où la mesure de l'utilité de données (l'inverse de la perte d'information) est signalée sur l'axe x, et le risque de divulgation est signalé sur l'axe y.

Pour chaque technique de protection des micro-données, une ligne est dessinée sur le plan cartésien avec un point pour chaque paramètre. Sur la base du graphique obtenu, il est possible de comparer les différentes techniques de protection et de choisir le mieux adapté. Une fois qu'une technique de protection a été choisie, les cartes de confidentialité R-U peuvent également être utilisées pour sélectionner les paramètres. Il est important de noter qu'une carte R-U n'est qu'une méthode pour corréler le risque de divulgation et la perte d'information et ces mesures doivent être calculées en utilisant l'une des méthodes mentionné

V. Les facteurs de risque de divulgation :

En général, les facteurs principaux suivants contribuent aux risques de divulgation [20].

- L'existence de tuples à haute visibilité (c.-à-d., des tuples possédant des caractéristiques uniques, comme un revenu élevé).
- L'existence d'un nombre élevé d'attributs communs entre la table des micro-données et les sources externes, ce qui peut augmenter la possibilité de la relier ou la rendre plus précise. En revanche, les principaux facteurs qui diminuent les risques de divulgation peuvent être résumés comme suit.

- Un tableau de micro-données contient souvent un sous-ensemble de l'ensemble de la population. Cela implique que l'information d'un répondant spécifique, qu'un utilisateur malveillant voudra peut-être savoir, peut ne pas être incluse dans la table de micro-données.
    - Les informations spécifiées dans les tableaux de micro-données publiées au public ne sont pas toujours à jour (souvent au moins un ou deux ans). Cela signifie que les valeurs des attributs des répondants correspondants ont peut-être été modifiées entre-temps. En outre, l'âge des sources d'information externes utilisées pour la liaison peut être différent de l'âge de l'information contenue dans le tableau des micro-données.
    - Une table de micro-données et les sources externes d'informations contiennent naturellement du bruit qui diminue la capacité de lier les informations.
    - Une table de micro-données et les sources externes d'informations peuvent contenir des données exprimées sous différentes formes, ce qui réduit la capacité de lier des informations.

VI. Classification des techniques de protection de micro-données :

Le contrôle de la divulgation des micro-données est un problème pratique important tant dans le secteur privé que dans les secteurs public et gouvernemental. Les techniques de protection des micro-données ont deux objectifs apparemment contrastés. D'un côté, ils devraient éviter la ré-identification qui se produit chaque fois que l'information d'un répondant apparaissant dans un tableau de micro-données est identifiée, c'est-à-dire être associé à l'identité du répondant correspondant. D'autre part, l'application de ces techniques devrait conserver les propriétés statistiques clés des données d'origine que les destinataires de données ont indiquées comme importantes. C'est-à-dire être associé à l'identité du répondant correspondant. D'une autre manière, plus précisément, compte tenu d'une table de micro-données T, une technique de protection des données devrait transformer cette table d'origine en une autre table de micro-données T′ de manière à ce que:

i) Le risque qu'un utilisateur malveillant puisse utiliser T′ pour déterminer des informations confidentielles ou pour identifier un répondant devrait être faible.

ii) L'analyse statistique sur T et sur T′ devrait produire des résultats similaires.

Plusieurs techniques de protection de divulgation de micro-données ont été proposées dans la littérature. Fondamentalement, ces techniques sont basées sur le principe selon lequel la ré-identification peut être contrariée en réduisant la quantité d'informations diffusées, en masquant les données (par exemple en ne libérant pas ou en perturbant leurs valeurs), ou en libérant des valeurs plausibles mais confondues au lieu de la réalité Ceux. Selon ce principe, les techniques de protection des micro-données peuvent être classées en deux catégories principales: les techniques de masquage et les techniques de génération de données synthétiques (voir la figure II.1).

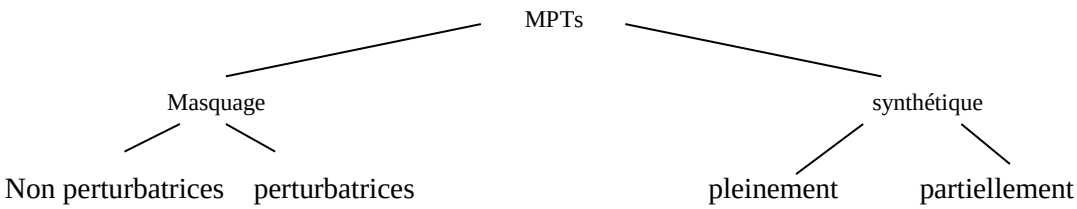

Figure. II.1. Classification des techniques de protection des micro-données (MPTs) [46].

Dans ce qui suit, nous décrivons les principales techniques de protection des micro-données

VI.1. Techniques de masquage :

Nous présentons certaines des techniques de masquage non perturbatrices et perturbatrices les plus populaires. La Table II.2. et la Table. II.3. répertorient les techniques indiquant si elles sont applicables (oui) ou non (non) aux types de données continues ou catégoriques.

VI.1.1. Techniques non perturbatrices :

Les techniques non perturbatrices produisent des micro-données protégées en éliminant les détails des micro-données initiales. Nous discutons de ces techniques en illustrant à titre d'exemples leur application à la protection de la table. II.1. Le résultat de l'application des techniques est illustré à la table. II.4.

SSN	nom	Race	D N	Sexe	ZIP	E C	Maladie	DH	Chol	Temp
		asiatique	64/09/27	F	94139	Divorcé	Hypertension	3	260	35.2
		asiatique	64/09/30	F	94139	Divorcé	Obésité	1	170	37.7
		asiatique	64/04/18	M	94139	Marié	Douleur de poitrine	*40*	200	38.1
		asiatique	64/04/15	M	94139	Marié	Obésité	7	280	37.4
		Noir	63/03/13	M	94138	Marié	Hypertension	2	*190*	35.3
		Noir	63/03/18	M	94138	Marié	Courte respiration	3	*185*	38.2
		Noir	64/09/13	F	94141	Marié	Courte respiration	5	200	36.5
		Noir	64/09/07	F	94141	Marié	Obésité	*60*	290	39.8
		blanc	61/05/14	M	94138	célibataire	Douleur de poitrine	7	*170*	37.6
		blanc	61/05/08	M	94138	célibataire	Obésité	10	300	40.1
		blanc	61/09/15	F	94142	Veuve	Courte respiration	5	200	36.9

Table. II.1 Un exemple de tableau de micro-données médicales [46].

SSN	nom	Race	D N	Sexe	ZIP	E C	Maladie	DH	Chol	Temp
		asiatique	64/09/27	F	9413*	Divorcé	Hypertension	3	260	nf
		asiatique	64/09/30	F	9413*	Divorcé	Obésité	1	<195	f
		asiatique	64/04/18	M	9413*	Marié	Douleur de poitrine	>30	200	f
		asiatique	64/04/15	M	9413*	Marié	Obésité	7	280	f
		Noir	63/03/13	M	9413*	Marié	Hypertension	2	<195	nf
		Noir	63/03/18	M	9413*	Marié	Courte respiration	3	<195	f
		Noir	64/09/13	F	9414*	Marié	Courte respiration	5	200	nf
		Noir	64/09/07	F	9414*	Marié	Obésité	>30	290	hf
		blanc	61/05/14	M	9413*	célibataire	Douleur de poitrine	7	<195	f
		blanc	61/05/08	M	9413*	célibataire	Obésité	10	300	hf
		blanc	61/09/15	F			Courte respiration	5	200	nf

Table. II.4. Tableau des micro-données de la Table. II.1 obtenu en appliquant les techniques non perturbatrices énumérées à la table. II.2 [46].

Technique	Continu	Catégorique
Échantillonnage	oui	Oui
Suppression locale	oui	Oui
Recodage globale	oui	Oui
Top-codage	oui	Oui
Bas Codage	oui	Oui
Généralisation	oui	Oui

Table. II.2. Applicabilité des techniques de masquage non perturbatrices aux différents types de données [46].

Technique	Continu	Catégorique
Ré-échantillonnage	Oui	Non
compression avec pertes	Oui	Non
Arrondi	Oui	Non
PRAM	Non	Oui
MASSC	Non	Oui
Bruit aléatoire	Oui	Oui
Swapping	Oui	Oui
Changement de classement	Oui	Oui
Micro-agrégation	Oui	Oui

Table. II.3. Applicabilité des techniques de masquage perturbatrices aux différents types de données [46].

VI.1.1.1. Échantillonnage [10] :

La table de micro-données protégées est obtenue en tant que témoin de la table de micro-données d'origine. En d'autres termes, la table de micro-données protégées comprend uniquement les données (tuples) d'un échantillon de la population entière. Puisqu'il y a une incertitude quant à savoir si oui ou non un répondant spécifique est dans l'échantillon, le risque de ré-identification dans les micro-données publiées diminue. Par exemple, nous pouvons décider de publier uniquement les tuples pairs de la table de micro-données d'origine. Cette technique ne fonctionne que sur des attributs catégoriques.

VI.1.1.2. Suppression locale [26,27]

Cette technique consiste à supprimer la valeur d'un attribut (c'est-à-dire le remplacer par une valeur manquante) limitant ainsi les possibilités d'analyse. Fondamentalement, cette technique met en évidence certaines valeurs d'attributs (cellules sensibles) susceptibles de contribuer de manière significative au risque de divulgation du tuple impliqué. Par exemple, nous pouvons supprimer les attributs ZIP et MarStat dans le dernier tuple.

VI.1.1.3. Recodage global (ou recodage dans les intervalles) [28, 14, 29] :

Le domaine d'un attribut est divisé en intervalles disjoints, habituellement de la même largeur, et chaque intervalle est associé à une étiquette. La protection du tableau des micro-données est obtenu en remplaçant les valeurs de l'attribut par l'étiquette associée à l'intervalle correspondant. Intuitivement, le recodage global diminue les détails dans la table des micro-données et, par conséquent, il devrait réduire le risque de ré-identification. Par exemple, supposons que les valeurs de l'attribut Temp dans la table II.1. soient partitionnées en trois intervalles: [35.0,36.9] avec étiquette pas de fièvre (nf); [37,0,38.9] avec la fièvre de l'étiquette (f).et [39.0,40.9] avec une forte fièvre d'étiquette (hf). La valeur dans le premier tuple est alors remplacée par l'étiquette "f".La deuxième, la troisième et la quatrième valeur sont remplacées par l'étiquette "f".. etc. Notez que si le domaine original de l'attribut considéré est continu, il devient discret après l'application de cette technique.

Deux techniques de recodage global Sont le Top-codage et le Bas codage décrit dans ce qui suit.

VI.1.1.4. Top-codage [13, 30] :

Il est basé sur la définition d'une limite supérieure, appelée top-code, pour chaque atout à protéger. Toute valeur supérieure à cette valeur est remplacée par le code top. Par exemple, considérez l'attribut DH dans la table II.1.et supposons que le top-code soit 30. Dans ce cas, plutôt que de publier le troisième et le huitième tuple montrant un nombre de jours dans un hôpital égal à 40 et 60, respectivement, ces deux tuples peuvent seulement montrer que le nombre de jours est \> 30 ". L'idée est que de longues périodes dans l'hôpital peuvent être facilement associées à des répondants spécifiques.

Le codage supérieur peut être appliqué à des attributs catégoriques qui peuvent également être ordonnés linéairement Quant aux attributs continus.

VI.1.1.5. Bas-Codage [13, 30]

Il est similaire au codage supérieur. Il consiste à définir une limite inférieure, appelée Bas-Codage, pour chaque attribut à protéger. Par conséquent, toute valeur inférieure à cette limite n'est pas publiée et remplacée par le Bas-Codage. Par exemple, considérez l'attribut Chol dans la table II.1. et supposez que le code inférieur est 195. Les deuxième, cinquième, sixième et neuvième tuples sont modifiés de telle sorte que la valeur publiée pour l'attribut Chol soit \ <195 ". Au fond Étant donné que les valeurs faibles en cholestérol chez les personnes ayant des problèmes d'obésité ou d'hypertension sont rares, elles doivent être obscures pour éviter une éventuelle ré-identification. Comme pour le codage supérieur, cette technique peut être appliquée aux attributs catégoriels qui peuvent être classés linéairement ainsi qu'à des attributs continus.

VI.1.1.6. Généralisation [31] :

Il consiste à représenter les valeurs d'un attribut donné en utilisant des valeurs plus générales. Cette technique est basée sur la définition d'une hiérarchie de généralisation, où la valeur la plus générale est à la racine de la hiérarchie et les feuilles correspondent aux valeurs les plus spécifiques. Un processus de généralisation procède donc en remplaçant les valeurs représentées par les nœuds foliaires par l'un de leurs nœuds ancêtres à un niveau supérieur. Différentes tables de micro-données généralisées peuvent être construites, selon le nombre d'étapes de généralisation appliquées sur l'attribut considéré. Par exemple, considérez l'attribut ZIP et le correspondant

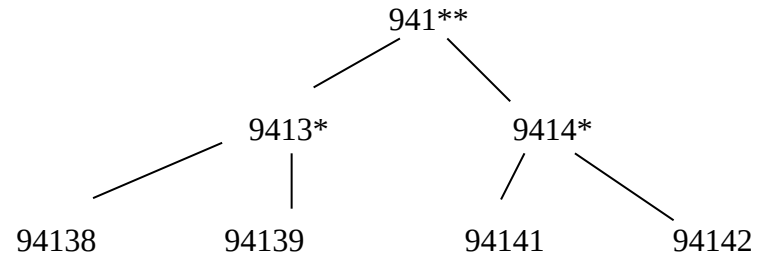

Figure. II.2. Hiérarchie de généralisation pour l'attribut ZIP[31] .

SSN	nom	Race	D N	Sexe	ZIP	E C	Maladie	DH	Chol	Temp
		asiatique	64/09/27	F	9413*	Divorcé	Hypertension	3	260	nf
		asiatique	64/09/30	F	9413*	Divorcé	Obésité	1	<195	f
		asiatique	64/04/18	M	9413*	Marié	Douleur de poitrine	>30	200	f
		asiatique	64/04/15	M	9413*	Marié	Obésité	7	280	f
		Noir	63/03/13	M	9413*	Marié	Hypertension	2	<195	nf
		Noir	63/03/18	M	9413*	Marié	Courte respiration	3	<195	f
		Noir	64/09/13	F	9414*	Marié	Courte respiration	5	200	nf
		Noir	64/09/07	F	9414*	Marié	Obésité	>30	290	hf
		blanc	61/05/14	M	9413*	célibataire	Douleur de poitrine	7	<195	f
		blanc	61/05/08	M	9413*	célibataire	Obésité	10	300	hf
		blanc	61/09/15	F			Courte respiration	5	200	nf

Table. II.4. Tableau des micro-données de la Table. II.1 obtenu en appliquant les techniques non perturbatrices énumérées à la table. II.2[46].

Hiérarchisation de la généralisation à la figure. II.2. Chaque étape de généralisation consiste à supprimer le chiffre le moins significatif du code postal. Dans ce cas, si nous choisissons d'appliquer une étape de généralisation, les valeurs 94138, 94139, 94141 et 94142 sont généralisées à 9413 * et 9414 *. Cette technique s'applique aux attributs continus et catégoriques. Notez également que la technique de recodage globale peut être considérée comme un cas particulier de généralisation.

La table II.4.contient le tableau de micro-données protégé obtenu à partir des micro-données de la table. II.1 en appliquant, comme on l'a vu, la technique de codage supérieur sur l'attribut DH, la technique de codage de fond sur l'attribut Chol, la technique de recodage globale sur l'attribut Temp, la technique de suppression locale sur le dernier tuple et une étape de généralisation sur l'attribut ZIP.

VI.1.2. Techniques perturbatrices :

Avec les techniques perturbatrices, la table des micro-données est modifiée pour publication. Les modifications peuvent faire disparaître des combinaisons de valeurs uniques dans la table d'origine et introduire de nouvelles combinaisons.

S1	S2	S3	S4
260	220	170	210
170	280	290	190
200	210	220	230
280	310	270	200
190	290	185	185
185	180	300	260
200	285	250	220
290	265	260	290
170	150	190	230
300	270	270	310
200	298	200	170

(A) échantillons initiaux

S1	S2	S3	S4	avarie
170	150	170	170	165
170	180	185	185	180
185	190	190	190	188.75
190	210	200	200	200
200	220	220	210	212.5
200	265	250	220	233.75
200	270	260	230	240
260	280	270	230	260
280	285	270	260	273.75
290	290	290	290	290
300	310	300	310	305

(B) échantillons commandés

Les Valeur d'origine (S1)	Valeur dégagée
260	260
170	165
200	212.5
280	273.75
190	200
185	188.75
200	233.75
290	290
170	180
300	305
200	240

(C) Données publiées

Table. II.5. Un exemple de ré-échantillonnage sur l'attribut Chol[32].

VI.1.2.1. Ré-échantillonnage [32, 33] :

Cette technique consiste à remplacer les valeurs d'un attribut continu sensible par la valeur moyenne calculée sur un nombre donné d'échantillons prélevés sur la population d'origine. Plus précisément, soit N le nombre de tuples dans une table de micro-données et $S_1, \ldots, S_t$ être t échantillons de taille N.

Les moyennes obtenues sont ensuite réordonnées en prenant en considération l'ordre des valeurs d'origine. La première valeur moyenne remplace la première valeur d'origine, la seconde moyenne remplace la seconde valeur d'origine, etc. Par exemple, supposons que l'attribut Chol soit protégé en appliquant cette technique et que nous choisissons $t = 4$ échantillons.

La table II.5. illustre les différentes étapes de la protection de l'attribut Chol. Notez que le premier échantillon (colonne S_1) correspond aux valeurs Chol dans la table de micro-données d'origine.

VI.1.2.2. Compression à perte [13, 34] :

C'est une technique récente qui exploite des algorithmes de compression d'image. Une table de micro-données continue est interprétée comme une image, et un algorithme de compression à perte (par exemple, jpeg) est appliqué sur elle. Le résultat est la table de micro-données protégées. En fonction de l'algorithme de compression à perte utilisé, il est nécessaire de détecter une correspondance appropriée entre les plages d'attributs et les échelles de couleurs.

Cette technique ne peut être appliquée que sur des données continues et le taux de compression coïncide avec le paramètre d'offuscation: plus le taux de compression est élevé, plus les données sont protégées.

VI.1.2.3. PRAM (Poser la méthode randomisée) [14, 36, 23] :

Il consiste à remplacer la valeur catégorielle pour un ou plusieurs attributs dans chaque tuple par une autre valeur catégorielle basée sur un mécanisme de probabilité. Par exemple, une matrice de Markov $P = [P_{ij}]$ (C.-à-d. Un réel $n \times n$ matrice, où tous les éléments p_{ij} sont supérieurs ou égaux à 0 et $\sum_{j=1}^{n} P_{ij} = 1, i = 1, \ldots\ldots\ldots, n$) Peut contenir la probabilité de remplacer les catégories dans la table de micro-données d'origine par d'autres catégories. En d'autres termes, P_{ij} est la probabilité que la catégorie c_i dans les micro-données d'origine soit remplacée par la catégorie C_j dans les micro-données protégées.

VI.1.2.4. MASSC (Micro-Agglomération, Substitution, Sous-échantillonnage et Calibration) [38] :

C'est une technique qui comprend quatre étapes qui fonctionnent comme suit.

- **Micro-agglomération**. Les Tuples dans la table de micro-données d'origine sont répartis en différents groupes caractérisés par un risque similaire de divulgation. Chaque groupe est formé sur la base de leur quasi-identificateur. Intuitivement, les tuples avec des combinaisons de valeurs rarissimes pour les attributs quasi identitaires présentent un risque plus élevé et devraient être dans le même groupe.
- **Substitution**. Les données originales sont perturbées en suivant une stratégie probabiliste optimale.
- **Sous-échantillonnage**. Certaines cellules ou des tuples entiers sont supprimés selon une stratégie de sous-échantillonnage probabiliste optimale.
- **Calibrage optimal**. Les poids d'échantillonnage, utilisés dans l'étape précédente, sont étalonnés pour préserver une certaine propriété statistique. En particulier, cet étalonnage implique des attributs qui doivent être utilisés par les destinataires de données pour les enquêtes.

Cette technique a été initialement proposée pour réduire la divulgation risque dû au lien entre les attributs catégoriques et les sources externes. C'est donc ne convient pas aux tables qui contiennent des attributs continus.

VI.1.2.5. Swapping(Échange de données) [32, 21, 37]:

Il consiste à modifier un sous-ensemble des tuples dans une table de micro-données en échangeant les valeurs d'un ensemble d'attributs sensibles, appelés attributs échangés, entre des paires de tuples sélectionnées (les paires sont sélectionnées selon un critère bien défini). Intuitivement, cette technique réduit le risque de ré-identification car elle introduit une incertitude quant à la valeur réelle des données d'un répondant.

À titre d'exemple, supposons que les attributs échangés sont Maladie, DH, Chol et Temp et que les paires de tuples sélectionnées doivent avoir une correspondance sur les attributs Sexe and E C . La table II.6 illustre le tableau obtenu en échangeant le tuple t3 avec t5, t7 avec t8 et t9 avec t10

SSN	nom	Race	D N	Sexe	ZIP	E C	Maladie	DH	Chol	Temp
		asiatique	64/09/27	F	94139	Divorcé	Hypertension	3	260	35.2
		asiatique	64/09/30	F	94139	Divorcé	Obésité	1	*170*	37.7
		asiatique	64/04/18	M	94139	Marié	*Hypertension*	*2*	190	*35.3*
		asiatique	64/04/15	M	94139	Marié	Obésité	7	280	37.4
		Noir	63/03/13	M	94138	Marié	*Douleur de poitrine*	*40*	*200*	*38.1*
		Noir	63/03/18	M	94138	Marié	Courte respiration	3	185	38.2
		Noir	64/09/13	F	94141	Marié	*Obésité*	60	*290*	*39.8*
		Noir	64/09/07	F	94141	Marié	*Courte respiration*	*5*	*200*	*36.5*
		blanc	61/05/14	M	94138	célibataire	*Obésité*	10	*300*	*40.1*
		blanc	61/05/08	M	94138	célibataire	*Douleur de poitrine*	7	170	*37.6*
		blanc	61/09/15	F	94142	Veuve	Courte respiration	6	200	36.9

Table. II.6. Table des micro-données de la table. II.1 protégés par l'échange sur les attributs Maladie, DH, Chol et Temp [21].

(les valeurs échangées sont signalées dans italique). Bien que cette technique soit facile à appliquer, en général elle a l'inconvénient de ne pas préserver les propriétés statistiques sur les sous-domaines. La technique originale n'a été présentée que pour les attributs catégoriques. Cependant, dans [39] l'échange de données a été étendu à des données continues.

VI.2. Techniques de génération de données synthétiques :

La génération de données synthétiques est une option alternative pour la protection des micro-données. Le principe de base sur lequel ces techniques sont basées est que, étant donné que le contenu statistique des données n'est pas lié à l'information fournie par chaque répondant, un modèle qui représente les données pourrait en principe remplacer les données elles-mêmes [18]. Une exigence importante pour la génération de données synthétiques, ce qui rend le processus de génération un problème compliqué, c'est que les données synthétiques et originales devraient présenter la même qualité d'analyse statistique. Le principal avantage de cette classe de techniques est que les données synthétiques publiées ne sont pas référées à un répondant et, par conséquent, leur publication ne peut pas conduire à une ré-identification. Ces techniques permettent aux détenteurs de données de porter leur attention sur la qualité des données diffusées au lieu d'attirer l'attention sur le problème de ré-identification. Dans le reste de cette section, nous décrivons les principales techniques de génération de données synthétiques. La table II.7. et la table. II.8. répertorie les techniques indiquant si elles sont applicables (oui) ou non (non) aux types de données continues ou catégoriques.

Technique	***Continu***	***Catégorique***
Bootstrap	oui	Non
Décomposition de Cholesky	oui	Non
Imputation multiple	oui	oui
Entropie maximale	oui	oui

Table. II.7. Applicabilité de techniques entièrement synthétiques aux différents types de données [18].

Technique	***Continu***	***Catégorique***
IPSO	oui	Non
Masquage hybride	oui	Non
Réponse aléatoire	Non	oui
SMIKe	oui	oui

Table. II.8. Applicabilité des techniques partiellement synthétiques aux différents types de données [18].

Les techniques sont divisées en deux catégories: ***techniques entièrement synthétiques*** et ***Techniques partiellement synthétiques***. La première catégorie contient des techniques qui génèrent un tout nouveau jeu de données, tandis que les techniques de la deuxième catégorie fusionner les données d'origine avec des données synthétiques.

VI.2.1. Techniques entièrement synthétiques :

Nous décrivons certaines techniques significatives de génération entièrement synthétiques qui ne publient que des données synthétiques.

VI.2.1.1. Bootstrap [33] :

Compte tenu d'une table de micro-données avec des attributs $\mathcal{P}$, cette technique calcule d'abord la fonction $\mathcal{F}$, une distribution cumulative de $\mathcal{P}$-variable. Une fonction de distribution cumulée $\mathcal{P}$ -variable est une fonction qui décrit complètement la distribution de probabilité d'un ensemble de variables aléatoires à valeur réelle (Par exemple, la fonction gaussienne). Les paramètres qui caractérisent $\mathcal{F}$ peuvent être déterminés en utilisant la technique bootstrap. Fondamentalement, bootstrap estime chaque paramètre de la population en utilisant un ensemble d'échantillons synthétiques, obtenus à partir de l'échantillon d'origine par un ré-échantillonnage avec remplacement. Une fois que les paramètres ont été estimés, la fonction correspondante $\mathcal{F}$ sur la population est modifiée pour obtenir une fonction similaire $\mathcal{F}'$. Cette nouvelle fonction est ensuite échantillonnée pour obtenir un ensemble de données synthétiques. Les modifications apportées à la fonction $\mathcal{F}$ devraient néanmoins préserver les propriétés statistiques des données d'origine. Notez que cette technique ne peut être appliquée que sur des attributs continus car il n'est pas possible de calculer la fonction $\mathcal{F}$ sur les données catégorielles.

VI.2.1.2. Décomposition Cholesky [40] :

Cette technique, qui ne fonctionne que sur des attributs continus et dans le temps linéaire dans la taille de l'échantillon, conserve la moyenne, la variance et la co-variance des données d'origine et est basée sur la méthode de décomposition de la matrice de Cholesky. Compte tenu d'une table de micro-données T, Qui peut être représenté

comme une matrice d'éléments N × M, où les lignes sont des tuples et les colonnes sont des attributs, il est nécessaire de calculer la matrice de co-variance C sur T.

L'étape suivante consiste à générer une matrice aléatoire, notée comme R, de taille N × M, De sorte que la matrice d'identité I est la matrice de co-variance. Ensuite, la décomposition de Cholesky U de C est déterminée, où C = U^t× U. La matrice de micro-données synthétiques est ensuite calculée comme R . U, et elle a exactement la même matrice de co-variance que T.

VI.2.2. Techniques partiellement synthétiques :

Étant donné qu'il peut être difficile de générer des données synthétiques plausibles pour tous les attributs, des techniques qui génèrent des ensembles de données partiellement synthétiques ont également été prises en considération. Fondamentalement, ces techniques produisent un mélange de valeurs synthétiques et originales. Nous décrivons maintenant les principales techniques partiellement synthétiques.

VI.2.2.1. IPSO (Information Preserving Statistical Obfuscation) [40] :

Cette technique est basé sur la distinction de deux catégories d'attributs: les données publiques Y et les données d'enquête spécifiques X. elle libère un sous-ensemble de l'échantillon d'origine après une opération de perturbation réalisée uniquement par rapport aux attributs publics, obtenant ainsi un nouvel ensemble de valeurs Y′ . Étant donné que le but principal de cette technique est de publier autant de valeurs que possible collectées dans l'enquête spécifique, empêchant la ré-identification, seules des informations dans Y sont publiées pour préserver les statistiques S les plus importantes sur ces données. Plus précisément, le set Y′ est généré de manière à conserver X inchangé et à maintenir l'ensemble S des statistiques sur Y. À la fin, le nouvel échantillon (X; Y 0) est relâché. Cette technique fonctionne uniquement sur des attributs continus.

VI.2.2.2. Masquage hybride [34] :

Cette classe de techniques combine les données d'origine avec les données synthétiques. En particulier, après la génération d'un échantillon simulé, chaque tuple dans la table de micro-données d'origine est assorti d'un tuple dans l'image simulée. Ensuite, tous les tuples appariés sont combinés linéairement, en ajoutant ou en multipliant leurs valeurs, et les valeurs obtenues sont publiées. Ces techniques ont l'avantage de préserver la validité analytique des données. Ils fonctionnent uniquement sur des attributs continus.

VII. Quelques approches de protection des micro-données:

Dans cette partie nous allons présenter les approches les plus répondues de protection des données à caractère personnel, à savoir : les modèles k-anonymat,l-diversité, *t*-proximité , et differential Privacy.

VII.1. Le 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 **[41].**

Le k-anonymat fait partie de ce que nous appelons les méthodes syntaxiques (par oppositions aux méthodes sémantiques). Ces méthodes reposent sur une hiérarchisation des données, du plus général au plus spécifique, couramment appelée taxonomie. Par exemple, une hiérarchisation très simple peut être représentée par la figure II.3. : nous voyons que le Québec (élément général) peut être vu comme un ensemble de villes (Montréal, Québec, Rimouski), qui sont constituées de quartiers ou arrondissements. À la différence de cet exemple, les hiérarchies dans les méthodes d'anonymisation doivent être totales, c'est à dire que toutes les valeurs individuelles doivent s'y retrouver.

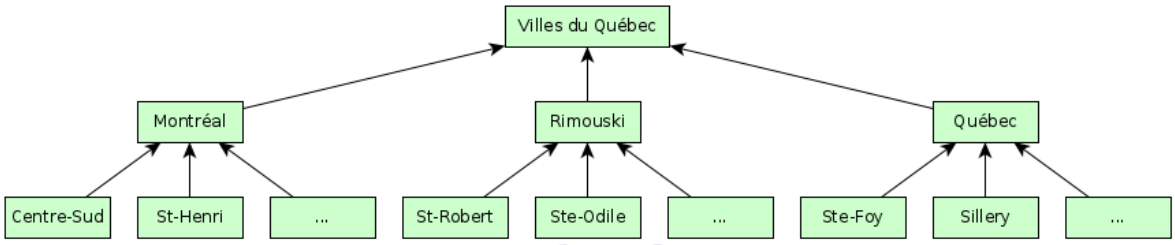

Figure II.3. Exemple de taxonomie : certaines villes et quartiers du Québec[41].

Plus spécifiquement, cette méthodologie [42] repose sur deux concepts permettant de fournir des garanties prouvables sur la confidentialité de l'ensemble de données résultant :

- **la généralisation** des paramètres, soit de remplacer une valeur spécifique par une valeur générale sans la falsifier
- **la suppression** des valeurs des attributs qui ne pouvant pas survivre à une généralisation satisfaisante

Le but du k-anonymat est de contraindre l'information relâchée par le propriétaire de sorte que toute tentative d'identification par quasi-identificateurs retourne au moins k tuples (d'où le nom).

VII.1.1. Le k-anonymat : un exemple d'application :

Le k-anonymat ne prescrit aucun algorithme particulier. Le respect des conditions incombe à l'implémentation. Il est toutefois assez aisé de montrer un exemple sur un petit ensemble de données afin de voir les avantages et contraintes d'une telle méthode.

Date de naissance	Sexe	Lieu de résidence	Statut marital	Cause de sélection
1970-10-14	F	Montréal	Divorcé	IMC trop élevé
1970-10-16	M	Montréal	Divorcé	Hypertension
1970-11-14	F	Montréal	Célibataire	IMC trop bas
1964-07-11	M	Montréal	Célibataire	Hypertension
1964-03-24	F	Québec	Marié	Séropositif
1964-04-03	F	Québec	Marié	Diabète de type 2
1964-03-25	F	Québec	Marié	IMC trop élevé
1964-04-27	M	Sherbrooke	Veuf	AVC
1964-03-25	M	Sherbrooke	Célibataire	Séropositif

Table II.9. Ensemble de données [41].

En reprenant le Table II.9. et en supposant que toutes les combinaisons des champs Date de naissance, sexe, lieu de résidence et Statut marital soient à notre disposition, posons ici k = 2. Nous pourrons alors obtenir l'ensemble de données 2-anonymisé suivant.

Date de naissance	Sexe	Lieu de résidence	Statut marital	Cause de sélection
≤1970-12-31	F	*	*	IMC trop élevé
≤1970-12-31	M	*	*	Hypertension
≤1970-12-31	F	*	*	IMC trop bas
≤1970-12-31	M	*	*	Hypertension
≤1970-12-31	F	*	*	Autre
≤1970-12-31	F	*	*	Autre
≤1970-12-31	F	*	*	IMC trop élevé
≤1970-12-31	M	*	*	Autre
≤1970-12-31	M	*	*	Autre

Table II.10. Ensemble de données-type 2-anonymisé

Les champs dénotés par un astérisque représentent une suppression du champ. Étant donné le faible nombre de tuples et la diversité des champs, les résultats de l'exemple ne sont pas très intéressants. Le but était ici d'illustrer succinctement le processus de k-anonymisation. Avec un nombre plus élevé d'observations, nous pouvons augmenter le nombre de paramètres et leur granularité et possiblement augmenter le paramètre k.

VII.2. La l-diversité [44]:

Comme on l'a vu à la Figure II.4.,

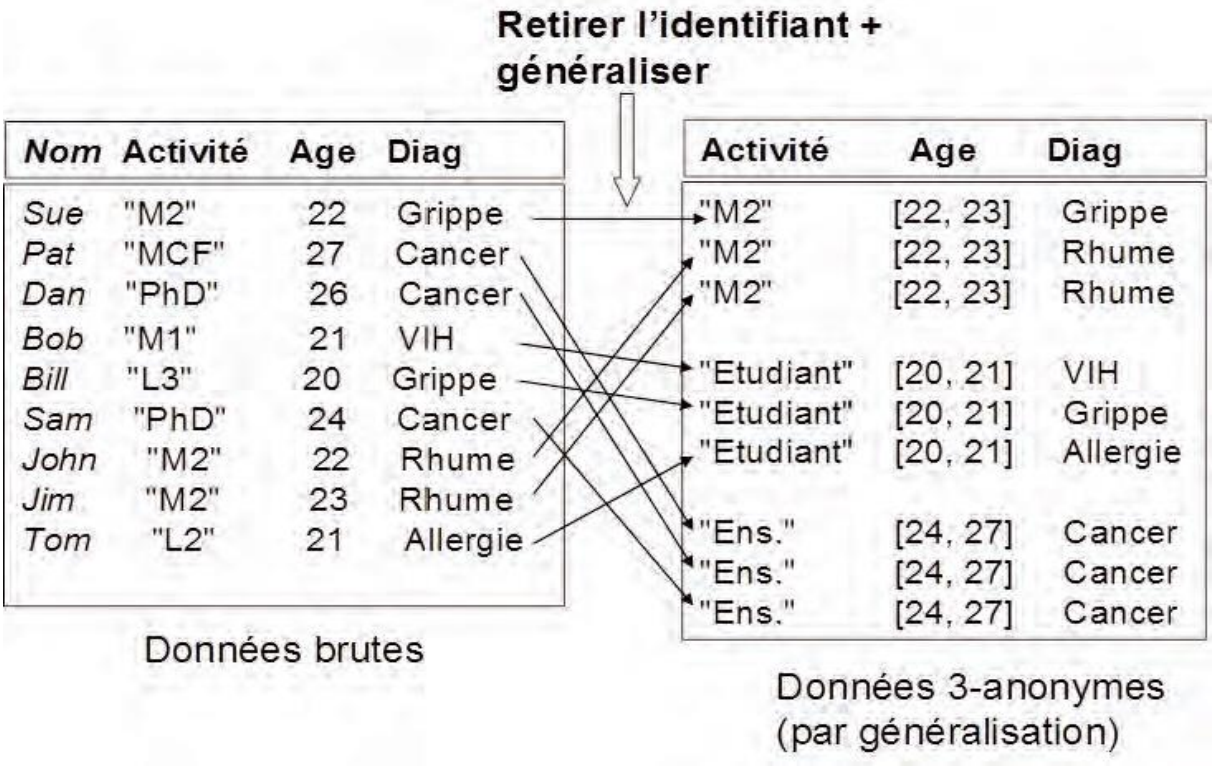

Figure II.4. Anonymisation d'une table sur des données universitaires[44].

il est possible de déduire des informations dans certains cas pathologiques, sans faire le moindre croisement, par exemple si tous les individus d'une classe possèdent la même valeur sensible. Le modèle de la l-diversité répond à ce problème, en rajoutant une contrainte supplémentaire sur les classes d'équivalence : non seulement au moins k_n-uplets doivent apparaître dans une classe d'équivalence, mais en plus le champ sensible associé à la classe d'équivalence doit prendre au moins l valeurs distinctes 5. Dans l'exemple de la Figure II.5., on voit que pour constituer de telles classes on doit parfois regrouper ensemble des étudiants et des enseignants. Leur activité est alors désignée de façon encore plus générale (« université »). Notons qu'on peut également lister les valeurs possibles, par exemple avoir une modalité « Étudiant ou Doctorant » (Etu/PhD).

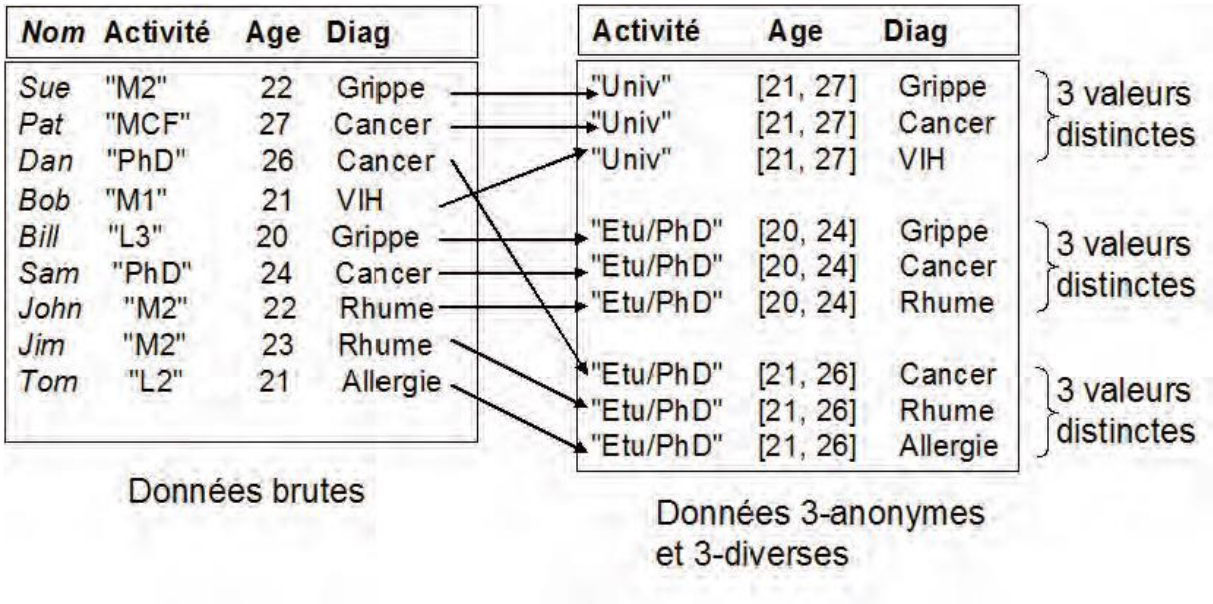

Figure II.5. Données l-diverses[44].

Cependant, en menant une attaque par croisement du même type que celle de Sweeney, il reste possible de déduire des informations. On voit par exemple dans la Figure II.5 qu'on peut déduire qu'un étudiant de 20 ans aura une probabilité 0.33 (soit 1/k) d'avoir la grippe, 0.33 d'avoir le cancer et 0.33 d'avoir un rhume... et surtout aucune chance d'avoir une autre pathologie. Si on sait que Bill est la seule personne de la base dans ce cas de figure, alors on peut déduire des informations sensibles à son sujet.

VII.3. La *t*-proximité [45]:

Pour essayer de réduire encore l'information qui peut être observée directement, on introduit le modèle de la t-proximité, toujours à partir d'un regroupement de données en classes d'équivalences selon le processus du k-anonymat. Ce nouveau modèle est basé sur une connaissance globale de la distribution des données sensibles, c'est-à-dire en ce cas les pathologies, pour essayer de faire coller au mieux les valeurs sensibles d'une classe d'équivalence à cette distribution, et ainsi éviter le problème de déduction d'informations soulevé par la l-diversité. Le facteur t que nous ne détaillons pas ici, indique dans quelle mesure on se démarque de la distribution globale.

Age	Sexe	Département	Pathologie	Nombre d'individus
<45	M	75	Grippe	400
<45	M	75	Rhume	800
>45	M	75	Grippe	500
>45	M	75	Rhume	1000
<35	F	75	Grippe	300
<35	F	75	Rhume	600
>35	F	75	Grippe	600
>35	F	75	Rhume	1200
...				

Figure II.6. *t*-proximité[45].

La t-proximité souffre de plusieurs problèmes, le plus important étant sans doute son utilité ! En effet, il parait évident d'exploiter des données k-anonymes ou même l-diverses pour découvrir des corrélations entre des données appartenant au quasi-identifiant et des données sensibles. Toutefois, le but même de la t-proximité est de réduire au maximum ces corrélations, puisque toutes les données sensibles de chaque classe d'équivalence vont se ressembler ! Ainsi, comme on le voit dans la Figure II.6., la t-proximité permet surtout de répondre à la question suivante : comment partitionner mes données de telle sorte que toutes les partitions se ressemblent en termes de distribution ? Par exemple, si on imagine une base de données nationale sur des pathologies, comment regrouper les départements, classes d'âge et sexes, de telle sorte

qu'on ait la même distribution des pathologies dans chaque sous-groupe. On peut s'interroger du jeu de données qui résulte de cette opération lorsqu'on souhaite précisément réaliser une analyse qui fait ressortir les facteurs qui différencient les individus.

VII.4. La confidentialité différentielle (Différential Privacy) [55]:

Nous concluons ce survol des techniques d'anonymisation par la confidentialité différentielle, une méthode très en vogue dans les milieux de la recherche en informatique depuis quelques années, car contrairement aux méthodes précédentes, elle est la seule à donner des garanties formelles, c'est-à-dire des preuves mathématiques, sur la possibilité de borner les informations qu'on peut apprendre sur les individus. Cette méthode introduit un échantillonnage des données vraies (avec une probabilité a), et une génération de données fictives avec une probabilité $\beta >> \alpha$ (mais ces données doivent naturellement rester réalistes...). Les garanties formelles sont cruciales, et permettent de quantifier le risque de ré-identification des n-uplets, d'où l'engouement pour cette méthode. En effet, en observant le jeu de données anonymes, l'information qu'on peut obtenir sur le fait qu'un n-uplet soit vrai ou faux est doublement bornée : on n'est jamais sûr qu' un n-uplet soit vrai avec une probabilité supérieure à α , ni qu'il soit faux avec une probabilité inférieure à β.

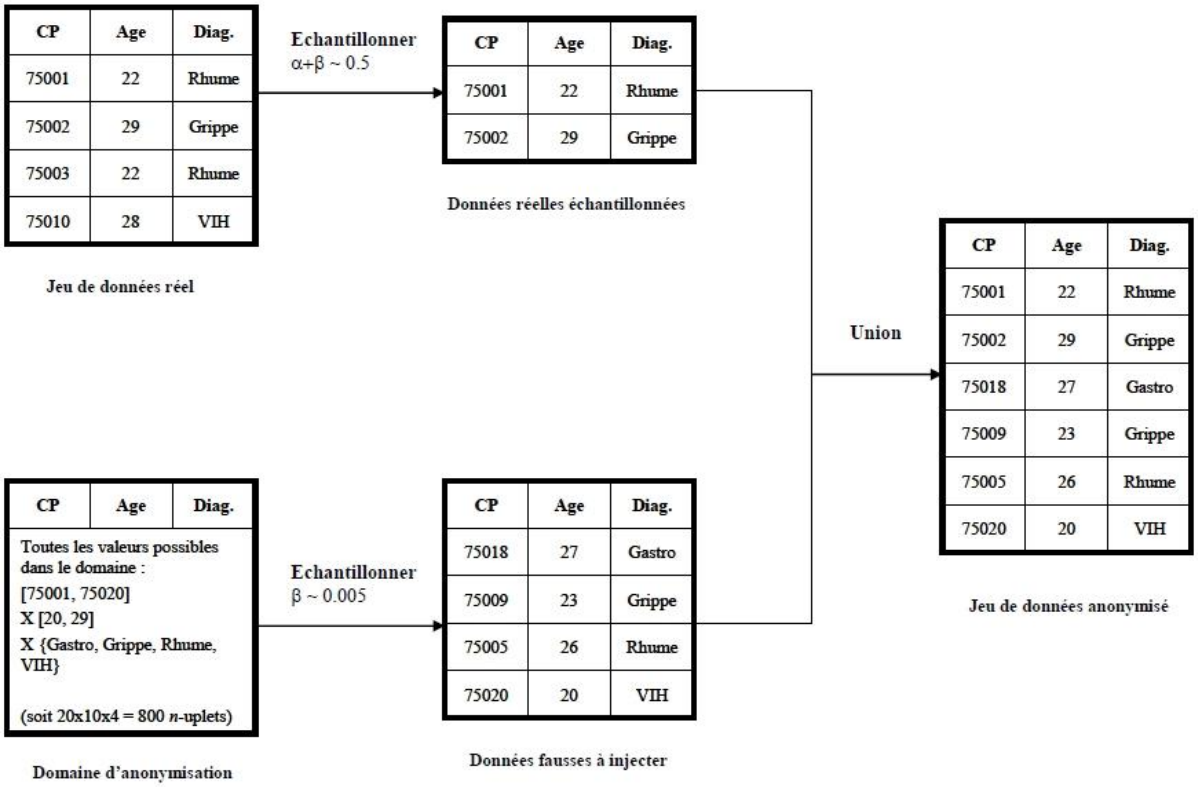

Figure II.7. Confidentialité Différentielle[55].

La confidentialité différentielle oblige à calculer un estimateur d'un agrégat que l'on souhaite connaître. Prenons l'exemple du calcul du nombre moyen de malades de la grippe par département, et supposons pour simplifier que les données fictives sont générées de manière équiprobable. On peut estimer le nombre total de malades de la grippe par la fonction suivante, dont l'objectif est de soustraire le bruit (connu) introduit:

$$\mathrm{Nb}_{\mathrm{Rhume}_{\mathrm{estimé}}} = \frac{(\mathrm{Nb}_{\mathrm{Rhume}_{\mathrm{anonyme}}} - \beta \times \mathrm{Nb}_{\mathrm{Rhume}_{\mathrm{domaine}}})}{\alpha} = (2 - 200 \times 0.005)/0.5 = 2$$

Le taux d'erreur peut également être estimé. Cependant, seules certaines fonctions d'agrégation peuvent être calculées avec une erreur bornée : moyenne, nombre total, etc. En revanche, on voit bien que calculer la valeur maximale d'une donnée numérique ne fait pas sens. Outre cette restriction, le problème principal de la mise en œuvre de la confidentialité différentielle réside dans la vraisemblance des données fictives. Ainsi, cette technique s'applique surtout lorsqu'on cherche à protéger des données de géo-localisation, où il est facile de générer des données fausses « plausibles », et où les

fonctions qu'on peut calculer avec cette technique d'anonymisation restent utiles (en particulier la densité et la distance). En revanche, comme on le voit sur l'exemple, il parait plus difficile d'exploiter cette méthode d'anonymat sur des données médicales.

VIII. Conclusion :

Nous avons présenté dans ce chapitre un état de l'art sur la protection des Micro-Donnée. Nous avons commencé par quelques définitions relatives à la protection des micro-données, après, une classification des techniques de protection, ainsi que les approches les plus populaires de protection ont été survolées. Dans le chapitre suivant nous présenterons la partie pratique de notre travail.

CHAPITRE III: Conception et implémentation

I. Introduction :

Ce chapitre décrit notre approche pour protéger des micro-données . Nous commençons par décrire le principe de cette approche et ses détails et nous terminons par un test et quelques résultats.

II. L'approche proposée :

L'idée principale de l'approche proposée est de publier des données fictives au lieu de vrais données qui exposent les individus à des risques d'attaques sur leur confidentialité. Les données fictives sont générées en utilisant des modèles issues des données originales, ce qui permet aux nouvelles données de garder certaines propriétés des données originales. L'approche se compose de deux étapes principales (figure III.1) :*la génération des données*, et *l'évaluation des données générées*. L'étape d'évaluation permet de tester la qualité des données générées. Si la qualité n'est pas satisfaisante on refait l'étape de génération.

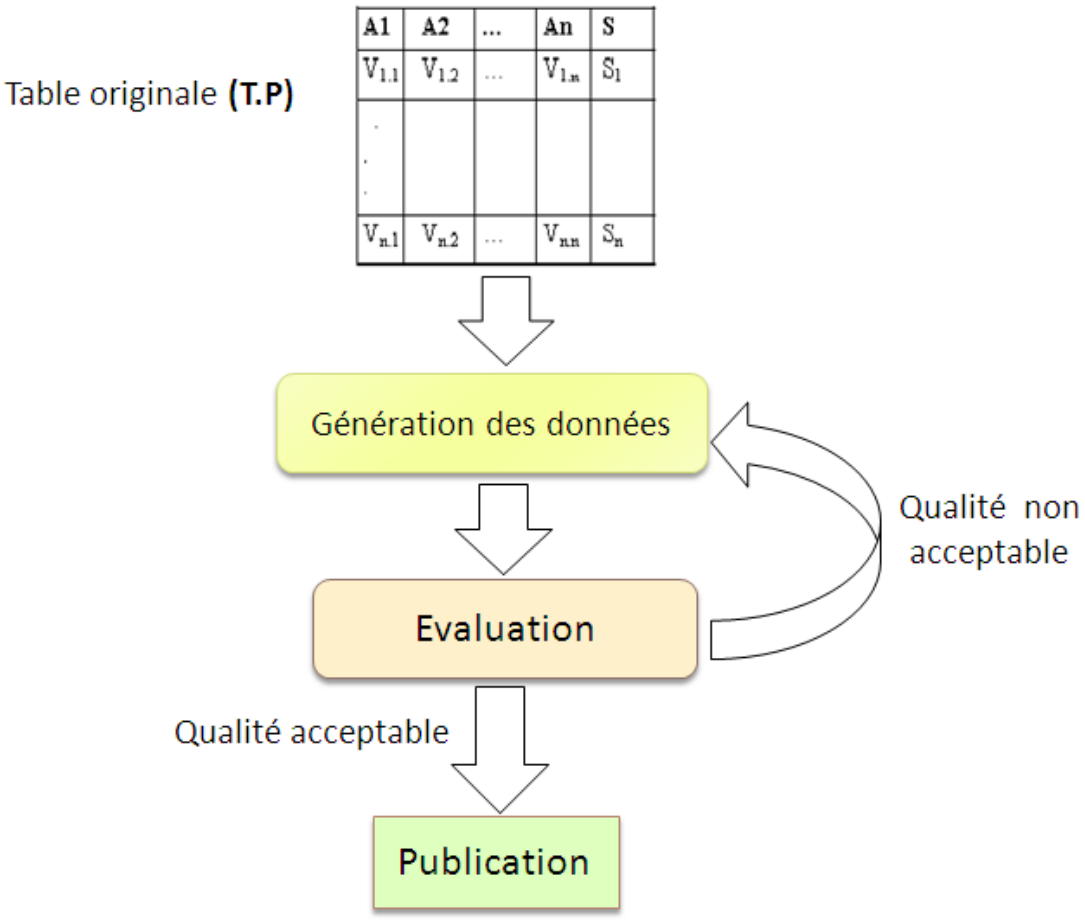

Figure III.1. Le principe de l'approche

La suite de cette section introduit une brève formulation du problème, avant d'expliquer les étapes de génération des données ainsi que le mécanisme d'évaluation.

II.1.La formulation du problème :

Soit une table " TP " dite table privée qui illustre les données originales d'un ensemble de personnes. La table " TP " contient " N " attributs, ces attributs sont divisés en deux catégories : les attributs non sensibles : $A = A_1, A_2, \ldots A_m$ (ex : âge, sexe, niveau d'études,..) et les attributs sensibles : $S = S_1, S_2, \ldots S_k$ (ex : maladie, salaire,...), avec $m + k = N$. Dans notre modèle d'attaque, un attaquant essaie d'acquérir les valeurs des attributs sensibles d'un individu en se basant sur ses connaissances partielles sur les attributs non sensibles. L'objectif de notre approche est de générer à partir de la table " TP " une nouvelle table " TG " dite générée, cette dernière sera publiée à la place de la première. La table " TG " doit contenir exactement le même nombre d'attributs.

Dans notre travail nous avons traité le cas d'un seul attribut sensible, la généralisation sur plusieurs attributs sensibles n'affecte pas les principes de notre approche.

II.2.La génération des données :

La génération de la table " TG " passe par trois étapes (figure III.2) :

1. La construction d'un modèle de classification ;
2. La génération des attributs non sensibles ;
3. L'application des règles sémantique
4. La prédiction de l'attribut sensible.

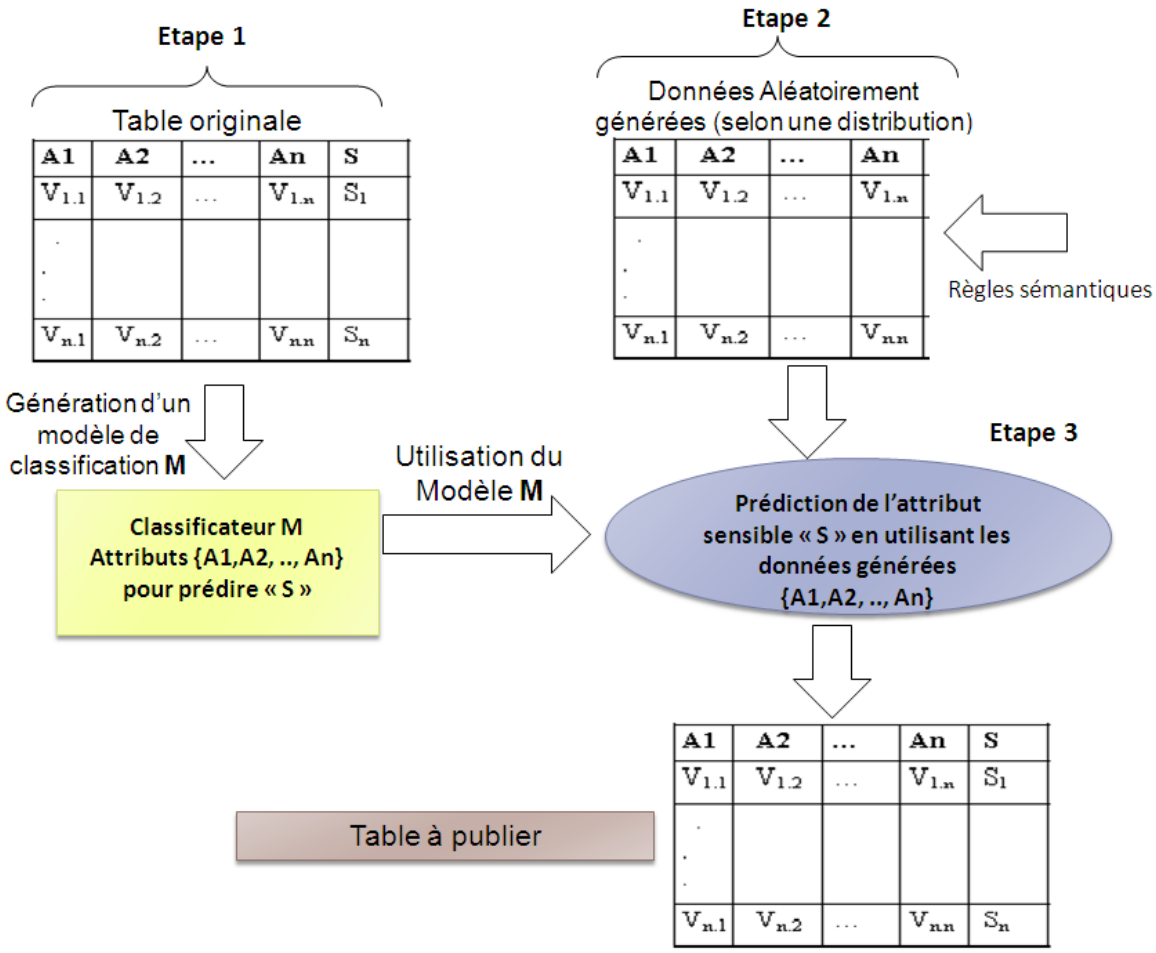

Figure III.2. Les étapes de génération des données

- **Etape 1 :** la construction du modèle de classification Dans cette étape un modèle de classification est construit à partir de la table " TP ". Ce modèle est issu d'un mécanisme d'apprentissage sur les données de la table " TP ". A la fin de cette étape on obtient un modèle capable de prédire la valeur de l'attribut sensible " S " en prenant comme entrés les attributs non sensibles (A1...An).
- **Etape 2 :** la génération des attributs non sensibles Cette étape consiste à générer aléatoirement un ensemble de valeurs pour chaque attribut non sensible en utilisant les intervalles des valeurs issues des données originales.

- **Etape 3 :** Cette étape est guidée par un ensemble de règles sémantiques appliquées sur les données générées pour éviter les cas réellement impossibles. Un exemple de règle sémantique est : " Une personne qui a un bas niveau d'éducation c'est impossible d'avoir un bon travail ou occupation", par exemple une personne qui a un niveau d'éducation de 7ème ou 8ème moyenne est impossible d'occuper un poste de travail comme un professeur ou un ingénieur dans un domaine particulier. D'autres exemples de règles sémantique utilisées dans notre travail est comme suit :

- Un enfant est impossible d'avoir un travail.
- Un enfant ne peut jamais avoir un niveau universitaire ;
- Une personne d'état civil divorcé ne peut pas avoir un mari.
- Une femme ou un homme d'état civil marié doit avoir un époux ou une épouse.

La génération des attributs est orientée par ces ensembles de règles sémantiques.

Le processus d'application de règles sémantiques dans cette étape est présenté par la figure suivante :

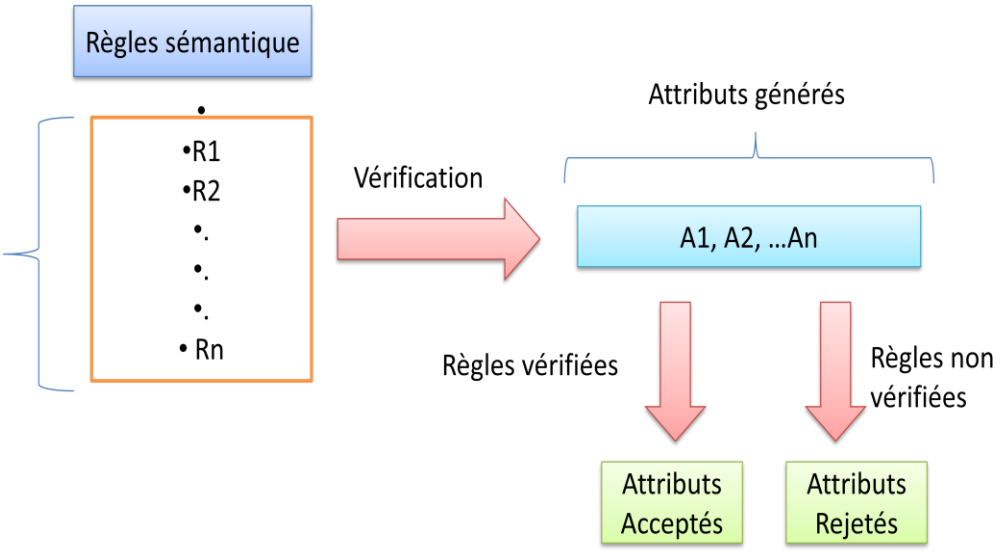

Figure III.3. Le processus d'application de règles sémantiques

- **Etape 4 :** la prédiction de la classe sensible En utilisant le modèle de classification construit à l'étape 1 et les données générées à l'étape 2, on peut prédire les différentes valeurs de l'attribut sensible " S ". La fin de cette étape

produit la table " TG ". Si cette table est approuvée par le mécanisme d'évaluation, elle sera publiée au lieu de la table " TP ".

II.3.Le mécanisme d'évaluation :

Le mécanisme d'évaluation consiste à comparer les performations d'un modèle de classification issu des données originales (table TP), avec un modèle issu des données générées (table TG), si la différence ne dépasse pas un certain seuil on peut valider les données générées. Les étapes d'évaluation se déroulent comme suit (figure III.4) :

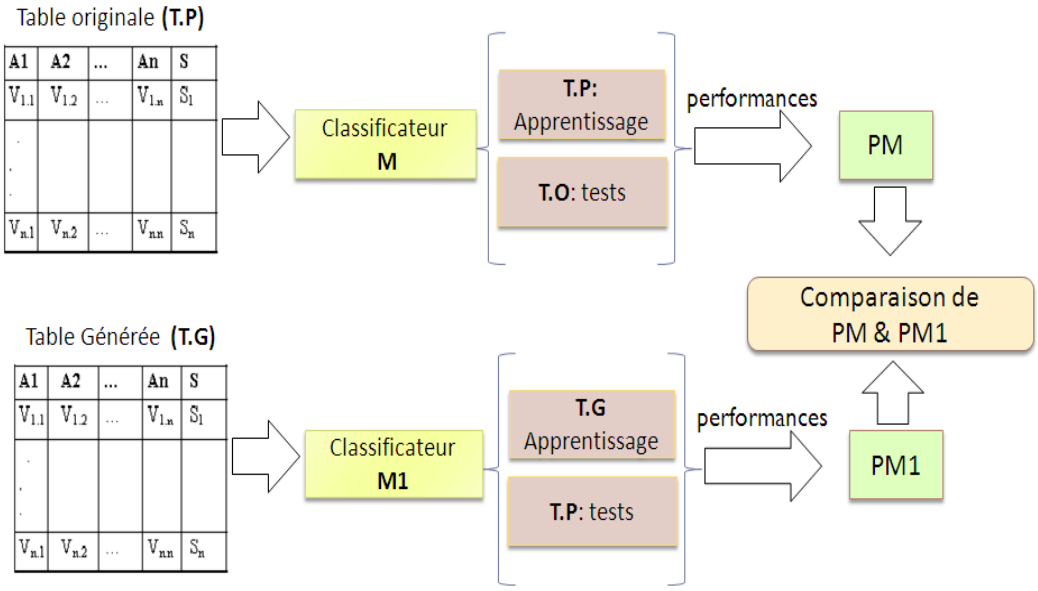

Figure III.4. Le mécanisme d'évaluation

- **Etape 1 :** Dans cette étape on construit un modèle de classification " M " à partir de la table " TP ". Cette construction se base sur l'utilisation d'une partie de données de la table " TP " comme base d'apprentissage. Les performances du modèle construit notées " PM " sont tirées en utilisant une partie de la même table comme base de tests.
- **Etape 2 :** Dans cette étape on construit aussi un modèle de classification " M1 " à partir de la table générée " TG". Cette construction se base sur l'utilisation d'une partie de données de " TG " comme base d'apprentissage. Les performances du modèle construit, notées " PM1 " sont tirées en utilisant cette fois la table " TP " comme base de tests.

- **Etape 3 :** cette étape compare " PM " avec " PM1 ". Si les différences ne dépassent pas certains seuils (ex : la différence de la précision de " M " et " M1 " ne dépasse pas un seul S), on peut juger les données de la table " TG " comme valides.

III. Expérimentation et résultats :

Nous avons testé notre approche sur la base " Adulte Data Set " (http ://archive.ics.uci.edu/ml/datasets/Adult), cette base est contient 14 attributs dont un est considéré comme sensible (le revenu). Cette base est considérée comme un benchmark dans ce domaine. Dans la base de test l'attribut sensible est un attribut binaire (revenu >50K ou <=50K), pour cette raison on a choisi pour la génération des données un classifieur de type (J48) .

Pour la phase d'évaluation nous avons utilisé trois algorithmes : Naive bayes [50], foret Aléatoire (Random Forest) [51] et SMO(Sequential Minimal Optimization) . L'utilisation de plusieurs algorithmes donne plus de crédibilité et d'assurance que les données générées sont valides et ne dépendent pas d'un seul algorithme. Notre choix des algorithmes est fait d'une manière à valider les données générées sur plusieurs familles d'algorithme : la famille des classifieurs bayésiens est représenté par l'algorithme Naive bayes, la famille des classifieurs fonctionnels est représenté par l'algorithme SMO(Sequential Minimal Optimization) et la famille des classifieurs de type arbres de décisions est représenté par l'algorithme RandomForest .

Les performances des modèles sont évaluées en termes de " précision " et de " rappel ". Dans l'étape de générations des données nous avons généré une base de 40000 individus, la totalité de cette base est utilisée comme base d'apprentissage pour élaborer le modèle de classification pour l'évaluation de la qualité des données générées. La base de test pour ce modèle est composée de 16000 individus choisis aléatoirement à partir des données originales.

La table III.1 représente les résultats d'évaluation des modèles issues des données générées en faisant varier le nombre des règles sémantiques utilisées dans la génération. Notons que les règles sémantiques utilisées dans chaque étape ont été choisies aléatoirement.

Algorithme	Random Forest		Naïve Bayes		SMO	
	précision	Rappel	précision	rappel	précision	rappel
Données Originales	84.0	84.8	82.1	83.1	84.3	85.0
Données Générées(12 R.S)	82.0	76.4	76.0	78.6	82.3	80.6
8 Règles sémantiques	80.8	76.8	72.9	65.6	81.7	79.8
6 Règles sémantiques	74.1	76.4	75.2	78.0	77.0	77.7
4 Règles sémantiques	58.4	76.4	71.9	74.5	76.3	77.1
2 Règles sémantiques	58.4	76.4	68.1	62.4	70.6	76.4
Sans les Règles sémantiques	75.2	76.5	78.9	78.8	58.4	76.4

Table III.1. Comparaison des performances des modèles issus des données originales et les données générées selon le nombre de règles sémantiques.

Les résultats montrent que globalement la qualité des données générées s'améliore avec l'augmentation de nombre de règles sémantiques utilisées. Cette amélioration est très apparente lorsque le nombre de règles est grand (8 et 12 règles). Des exceptions sont faites dans les cas de nombre de règles 2 et 4, où on remarque une dégradation dans la qualité des données générées par rapport à une base qui n'utilise aucune règle sémantique et cela pour tout les algorithmes de classification utilisés.

L'explication de ce phénomène nécessite une étude plus approfondie. Une première hypothèse c'est que l'utilisation d'un nombre restreint de règles peut provoquer la suppression d'un sous ensemble de valeurs de certains attributs, cela peut affecter une corrélation (si elle existe) entre les attributs et par conséquent le modèle de classification issu de ces données.

La figure III.5 montre la différence entre les qualités des modèles issue par une génération qui n'utilise pas de règles sémantiques et une génération qui utilise 12 règles sémantiques.

On remarque qu'une amélioration peut atteindre jusqu'à 23% de précision et plus de 4% de rappel (pour l'algorithme SMO). Pour l'algorithme Naive Bayes il n'y a pas d'amélioration dans la qualité mais plutôt une légère dégradation (de 2% de précision et 0.2% de rappel).

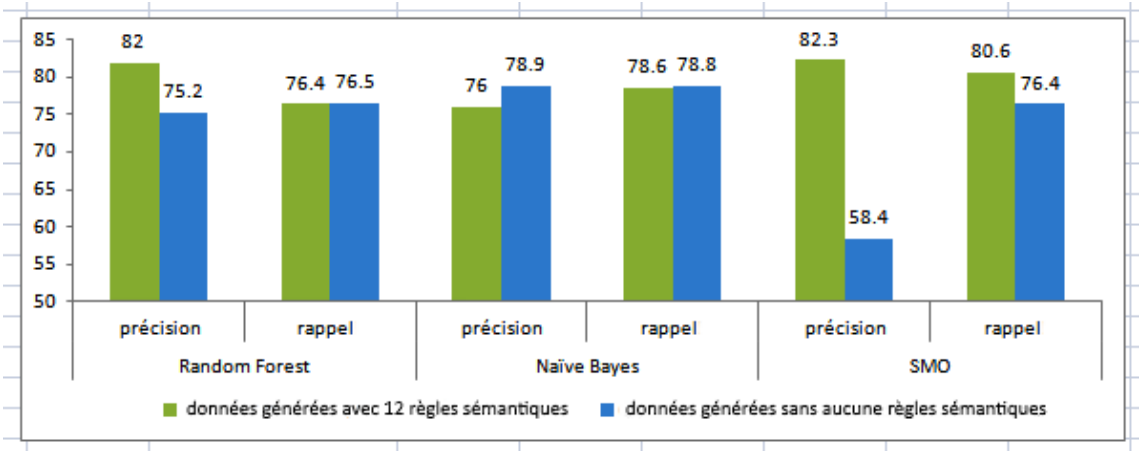

Figure III.5.Comparaison des performances des modèles issus des données générées avec 12 règles sémantiques et les données générées sans aucune règles sémantiques.

La table III.2 et la Figure III.6 représentent la comparaison des performances entre le modèle de classification issu des données originales et le modèle de classification issu des données générées avec 12 règles.

Algorithme	Random Forest		Naïve Bayes		SMO	
	précision	Rappel	précision	Rappel	précision	rappel
Données Originales	84.0	84.8	82.1	83.1	84.3	85.0
12 Règles sémantiques	82.0	76.4	76.0	78.6	82.3	80.6
Dégradation	-2%	-8.4%	-6.1%	-4.5%	-2%	-4.4%

Table III.2.Comparaison des performances des modèles issus des données originales et les données générées avec 12 règles.

Les résultats montrent que le modèle de classification issu de données générées a subi une dégradation de précision entre 2% (Rndom Forest et SMO) et 6% (Naive Bayes) et de rappel entre 4% (SMO et Naive Bayes) et 8% (Random Forest). Nous jugeons que le niveau de dégradation est acceptable et qu'il reflète bien le compromis à faire entre l'utilité des données publiées et la protection de la vie privée des individus.

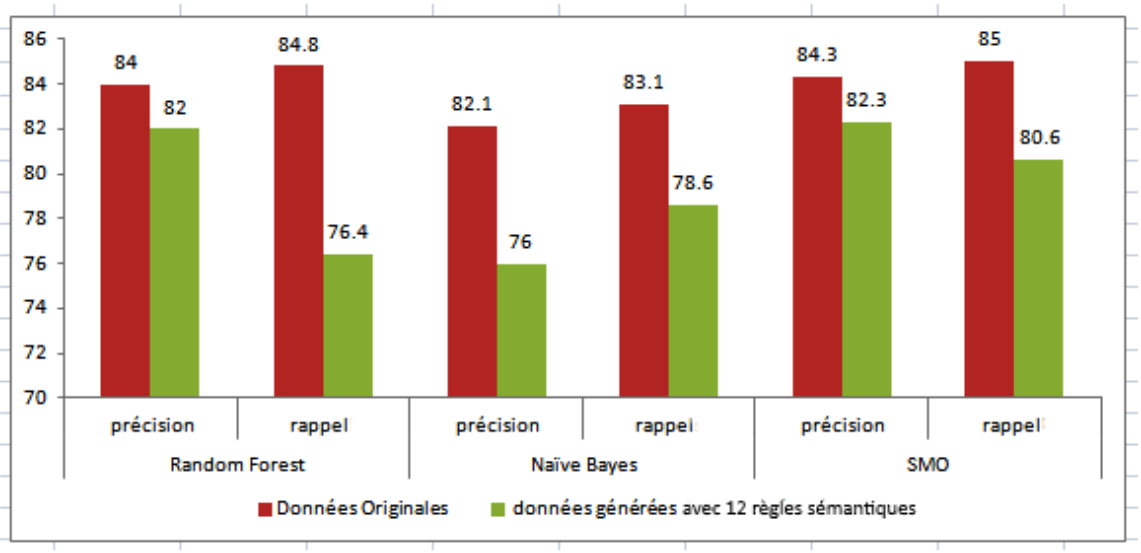

Figure III .6.Comparaison des performances des modèles issus des données originales et les données générées avec 12 règles sémantiques.

IV. Conclusion :

Dans ce travail nous avons proposé une approche qui génère aléatoirement des nouvelles données à partir des données originales en utilisant un classifieur automatique pour garder le maximum de corrélation entre les attributs sensibles et les attributs non sensibles ainsi qu'un ensemble de règles pour garder une sémantique acceptable entre les différentes valeurs des attributs générées . Les nouvelles données générées se diffèrent totalement des données originales ce qui implique une grande protection.

Conclusion générale

Le travail présenté dans ce mémoire entre dans le cadre de la protection de la vie privée sur Internet et plus précisément la protection des données personnelles (les micro-données).. Nous avons donné une vue générale sur ce domaine en introduisant des connaissances sur la vie privée sur l'internet, les types des attaques relatives, les différents risques associés à la perte des données personnelles ainsi qu'un état de l'art sur les techniques et les approches les plus populaires pour la protection des micro-données. Nous avons proposé une approche de protection dont l'idée est de publier des données fictives au lieu de vrais données.

Les données fictives sont générées en utilisant des modèles issues des données originales en se basant sur les Techniques de Machine Learning (classification automatique). Le classificateur automatique dans notre approche a pour rôle d'évaluer et de garder un maximum de corrélation entre les attributs sensibles et les attributs non sensibles. Notons aussi que la génération des données fictives est guidée par un ensemble de règles sémantiques pour garder une sémantique acceptable entre les différentes valeurs des attributs générées ce qui permet aux nouvelles données de garder certaines propriétés des données originales. Les nouvelles données générées diffèrent totalement des données originales ce qui implique une grande protection.

Comme perspectives nous envisageant de :

- Tester notre approche sur d'autres bases, et avec d'autres règles sémantiques, pour ces règles on va essai de faire des règle qui ne fait pas réduise plusieurs instance de la base original ;
- Introduire des nouveaux mécanismes dans la phase de génération qui permettent de capturer les corrélations entre attributs et de garder quelques propriétés statistiques des attributs (moyenne, écart-type,..).
- La comparaison des performances de notre approche avec d'autres travaux.
- Tester d'autres familles de classifieurs pour voir quel type est mieux adapté pour notre approche.

Références Bibliographiques

[1] Le droit de la vie privée, disponible sur :
http://www.chairelrwilson.ca/cours/drt3805/vieprivee.html, consulté le 15 mars 2017.

[2]La vie privée sur l'internet, disponible sur :
http://en.wikipedia.org/wiki/Internet_privacy, consulté le 17 juin 2017.

[3] Données personnelles définition et explications, disponible sur :
http://www.technoscience.net/?onglet=glossaire&definition=11079 , consulté le 14 mars 2017,

[4] L'anonymat et l'internet, disponible sur :
http://www.cga-canada.org/fr-ca/AboutCGACanada/CGAMagazine/2003/May-Jun/Pages/ca_2003_05-06_dp_doubleclick.aspx, consulté le 19 mars 2017

[5] Les cookies (biscuits empoisonnées) disponible sur :
http://fr.wikipedia.org/wiki/Cookie_%28informatique%29, consulté le 15 mars 2017.

[5] Jean-Philippe Walter, Le profilage des individus à l'heure du cyberespace : un défi pour le respect du droit à la protection des données, Stéphanie Lacour (Ed.), La sécurité de l'individu numérisé. Réflexions prospectives et internationales, 2008

[6] TCP Session Hijacking disponible sur :
https://www.aldeid.com/wiki/Attaques/Reseau/Vol-session, consulté le 12 mars 2017.

[7]I'injection de commandes SQL disponible sur :
http://php.net/manual/fr/security.database.sql-injection.php, consulté le 15 mars 2017.

[8] Eric Freyssinet. Lutte contre les botnets : analyse et stratégie. Cryptographie et sécurité [cs.CR]. Université Pierre et Marie Curie - Paris VI, 2015. Français. <NNT : 2015PA066390>.

[9] Chevaux de Troie disponible sur :
https://www.avast.com/fr-fr/c-trojan#academy, consulté le 26 mars 2017.

[10]. Federal Committee on Statistical Methodology (1994). Statistical policy working paper 22. USA. Report on Statistical Disclosure Limitation Methodology.

[11]. Duncan GT, Lambert D (1989). The risk of disclosure for microdata. Journal of Business and Economic Statistics, 7:207{217.

[12]. Domingo-Ferrer J, Mateo-Sanz JM, Torra V (2001). Comparing SDC methods for micro data on the basis of information loss and disclosure risk. In Preproceedings of ETK-NTTS'2001, vol. 2, pp. 807{826, Luxemburg. Eurostat.

[13]. Domingo-Ferrer J, Torra V (2001). A quantitative comparison of disclosure control methods for microdata. In Doyle P, Lane JI, Theeuwes J, and Zayatz L'editors, Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. North-Holland, Amsterdam.

[14]. Domingo-Ferrer J, Torra V (2002). Distance-based and probabilistic record linkage for re-identification of records with categorical variables. Butlleti de l'Associacio Catalana d'Intelligencia Artificial, 27.

[15]. Fellegi IP, Sunter AB (1969). A theory for record linkage. Journal of the American Statistical Association, 64(328):1183{1210.

[16]. Computer Science and Telecommunications Board National Research Council, editors (1997). For the record protecting electronic health information. National Accademy Press, Washington, D.C., USA.

[17]. Winkler WE (1999). Re-identification methods for evaluating the confidentiality of analytically valid micro data. In Domingo-Ferrer J, editor, Statistical Data Protection. Office for Official Publications of the European Communities, Luxemburg.

Références Bibliographiques

[18]. Burridge J, Franconi L, Polettini S, Stander J (2002). A methodological framework for statistical disclosure limitation of business microdata. Technical Re-port 1.1-D4, CASC Project.

[19]. Winkler WE (2004). Masking and re-identification methods for public-use microdata: Overview and research problems. In Domingo-Ferrer J, editor, Privacy in Statistical Databases 2004. Springer, New York.

[20]. Domingo-Ferrer J, Torra V (2001). Disclosure protection methods and information loss for microdata. In Doyle P, Lane JI, Theeuwes J, Zayatz L, editors, Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. North-Holland, Amsterdam.

[21]. Denning DE (1982). Inference controls. In Cryptography and Data Security, pp. 331{392. Addison-Wesley Publishing Company, Reading, Massachusetts; Menlo Park, California; London; Amsterdam; Don Mills, Ontario; Sydney.

[22]. Mateo-Sanz JM, Domingo-Ferrer J, Sebe F (2004). Probabilistic information loss measures for continuous microdata. Technical report, University of Tarragona, Department of Computer Engineering and Mathematics, Research Triangle Park, NC 27709-4006 USA.

[23]. Kooiman PL, Willenborg L, Gouweleeuw J (1998). PRAM: A method for disclosure limitation of microdata. Technical report, Statistics Netherlands, Voorburg, NL.

[24]. Willenborg L, DeWaal T (2001). Elements of Statistical Disclosure Control. Springer-Verlag, New York, USA.

[25. Duncan GT, Keller-McNulty SA, Stokes SL (2001). Disclosure risk vs. data utility: The R-U confidentiality map. Technical report, Los Alamos National Laboratory. LA-UR-01-6428.

[26].5. Cox LH (1980). Suppression methodology and statistical disclosure analysis. Journal of the American Statistical Association, 75(370):377{385.

[27]. Samarati P (2001). Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13(6):1010{1027.

[28]. Domingo-Ferrer J, Torra V (2001). A quantitative comparison of disclosure control methods for microdata. In Doyle P, Lane JI, Theeuwes J, and Zayatz L, editors, Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. North-Holland, Amsterdam.

[29] Takemura A (2001). On recent developments in statistical disclosure control techniques. In Proc. of the IAOS Satellite Meeting on Statistics for the Information Society, Tokyo, Japan.

[30]. Domingo-Ferrer J, Torra V (2002). Distance-based and probabilistic record linkage for re-identification of records with categorical variables. Butlleti de l'Associacio Catalana d'Intelligencia Artificial, 27.

[31]. Samarati P (2001). Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13(6):1010{1027.

[32]. Dalenius T, Reiss SP (1978). Data-swapping: a technique for disclosure control (extended abstract). In Proc. of the ASA Section on Survey Research Methods, pp. 191{194, Washington DC.

[33]. Fienberg SE (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Technical Report 611, Carnegie Mellon University Department of Statistics.

[34]. Dandekar R, Domingo-Ferrer J, Sebe F (2002). LHS-based hybrid micro data vs rank swapping and micro aggregation for numeric micro data protection. In Domingo-Ferrer J, editor, Inference Control in Statistical Databases, vol. 2316

[35]. Cox LH (1995). Network models for complementary cell suppression. Journal of the American Statistical Association, 90(432):1453{1462.

Références Bibliographiques

[36]. Gouweleeuw JM, Kooiman P, Willenborg RCLJ, DeWolf PP (1997). Post randomization for statistical disclosure control: Theory and implementation. Technical Report 9731, Voorburg: Statistics Netherlands, Netherlands.

[37]. Karr AF, Sanil AP (2004). Data quality and data confidentiality for microdata: Implications and strategies. Technical Report 149, National Institute of Statistical Sciences, Research Triangle Park, NC 27709-4006 USA.

[38]. Singh AC, Yu F, Dunteman GH (2004). MASSC: A new data mask for limiting statistical information loss and disclosure. In Linden H, Riecan J, Belsby L, editors, Work Session on Statistical Data Confidentiality 2003, pp. 373{394. Eurostat, Luxemburg. Monographs in OfficialStatistics.

[39]. Reiss S (1982). Non-reversible privacy transform. In Proc. of the ACM Symposium on Principles of Database Systems, Los Angeles, CA, USA.

[40]. Mateo-Sanz JM, Martunez-Balleste A, Domingo-Ferrer J (2004). Fast generation of accurate synthetic microdata. In Domingo-Ferrer J, Torra V, editors, Privacy in Statistical Databases, vol. 3050 of LNCS, pp. 298{306. Springer, Berlin Heidelberg

[41]. L. Sweeney : " k-anonymity: a model for protecting privacy" International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5), 2002.

[42]. Pierangela Samarati et Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, 1998. URL http://epic.org/privacy/reidentification/ Samarati_Sweeney_paper.pdf. [Accédé le 2016-02-07].

[43]Allard, N. Anciaux, L. Bouganim, Y. Guo, L. Le Folgoc, B. Nguyen, P. Pucheral, Ij. Ray, Ik. Ray, S. Yin : Secure Personal Data Servers: a Vision Paper, dans *Very Large Data Base*s, 3(1): 25-35, 2010.

Références Bibliographiques

[44]. A. Machanavajjhala , D. Kifer , J. Gehrke , et M. Venkatasubramaniam : "L-diversity : Privacy beyond k-anonymity" ACM Transactions on Knowledge,Discovery from Data, 1(1):2007.

[45]. N. Li, T. Li, S. Venkatasubramanian : " t-closeness: Privacy beyond k-anonymity and l-diversity", International Conference on Data Engineering, 2007.

[46] . V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati "Microdata Protection "Universitμa degli Studi di Milano, 26013 Crema, Italia fciriani, decapita, foresti, samaratig@dti.unimi.it

[47]. Définition sur la vie privée, disponible sur :https://fr.wikipedia.org/wiki/Vie_priv%C3%A9e#Le_concept_de_vie_priv.C3.A9e , consulté le 02 juin 2017.

[48] 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.

[49] Fung, B. C. M., Wang, K., Chen, R., and Yu, P. S. 2010. Privacy-Preserving data publishing: A survey of recent developments. ACM Comput. Surv. 42, 4, Article 14 (June 2010), DOI = 10.1145/1749603.1749605 http://doi.acm.org/10.1145/1749603.1749605

[50]. Harry Zhang "The Optimality of Naive Bayes". FLAIRS2004 conference, 2004.

[51]. Breiman, Leo . "Random Forests". Machine Learning 45 (1) : 5-32, 2001.

[52] L'attaque par Phishing , disponible sur : https://www.arobase.org/phishing/phishing.htm,consulté le 04 juin 2017.

Références Bibliographiques

[53] le processus d'attaque par phishing, disponible sur : http://blog.conixsecurity.fr/le-spear-phishing-une-variante-plus-efficace-du-phishing/, consulté le 04 juin 2017.

[54] botnet technique d'attaque disponible sur : https://steemit.com/steem/@kushbuddy/botnet, consulté le 02 juin 2017.

[55]. Yves Deswarte. "Protection de la vie privée : Principes et technologies".LAAS CNRS Toulouse, France.

[55]..C. Dwork : "Differential Privacy", International Colloquium on Automata, Languages and Programming, 2006.

Liste des tables

Liste des figures

Résumé

L'utilisation et le partage de données personnelles (micro-données) est devenu une nécessité dans plusieurs domaines, la publication de ces données sans révéler les informations sensibles est un problème important dans la vie privée. Le présent travail propose une approche de protection des micro-données qui se base sur la génération aléatoire des données fictives, la génération est guidée par un ensemble de règles sémantiques qui gardant une sémantique acceptable entre les différentes valeurs des attributs générées. Cette approche fournit une meilleure protection du fait que les données générées se diffèrent totalement des données originales.

Abstract

The use and sharing of personal data (micro-data) has become a necessity in several areas, the publication of this data without revealing sensitive information is an important problem in privacy. The present work proposes a micro-data protection approach based on the random generation of fictitious data. The generation process is guided by a set of semantic rules that maintain an acceptable semantics between the different values of the generated attributes. This approach provides better protection because the generated data is totally different from the original data.

الملخص

ان استخدام وتبادل البيانات الشخصية أصبح شيء ضروري في العديد من المجالات و نشر هذه البيانات دون الكشف عن المعلومات الحساسة هي مشكلة هامة في الحياة الخاصة و في نفس الإطار ان حماية الحياة الخاصة أضحت مهمة أكثر فأكثر في السنوات الأخيرة لهذا السبب عدة مناهج تم اقتراحها كحلول, العمل المقدم يعتمد على منهج يرتكز على الإنتاج أو الإنشاء العشوائي للبيانات. هذه البيانات تتحكم في إنتاجها مجموعة من القواعد الدلالية والتي تعطي معنى مقبول بين مختلف قيم البيانات المنتجة. هذا المنهج يزودنا بحماية جيدة لأن المعطيات الجديدة مختلفة تماما عن المعطيات الاصلية