République Algérienne Démocratique et Populaire
Université Abou Bakr Belkaid– Tlemcen
Faculté des Sciences
Département d'Informatique

Mémoire de fin d'études

pour l'obtention du diplôme de Master en Informatique

Option: *Système d'Information et de Connaissances (S.I.C)*

Thème

La protection de la vie privée sur Internet: Application sur les données personnelles

Réalisé par :

- Mlle.CHARIF Ismahan

Présenté le 01 Juillet 2013 devant le jury composé de MM :

- *CHAOUCHE Lamya* *(Président)*
- *BELABED Amine.* *(Encadreur)*
- *BENAMAR Abdlkrim* *(Examinateur)*
- *BENMANSOUR Lamya* *(Examinatrice)*

Année universitaire: 2012-2013

Résumé

La Publication des données personnelles sans pour révéler des informations sensibles est un problème important. Au cours de ces dernières années, différentes approches ont été proposées comme solution. Dans le même cadre, le présent travail introduit une approche qui se base sur la génération aléatoire des données, cette approche fournit une meilleure protection du fait que les données générées se diffèrent totalement des données originales.

Mots clés : vie privée, donnée personnelle, protection, génération aléatoire

Abstract

Publishing data about individuals without revealing sensitive information about them is an important problem. In recent years, various approaches have been proposed as a solution. In the same context, this paper introduces an approach based on randomly generated data; this approach provides better protection that the data generated is totally different from the original data.

Keywords: privacy, personal data, protection, random generation.

الملخص

ان نشر البيانات للأفراد دون الكشف عن معلومات حساسة عنها هي مشكلة هامة .في السنوات الأخيرة، اقترحت مناهج مختلفة كحل لهذه المشكلة, في هذا الاطار ,العمل المقدم هو منهج جديد يعتمد على التوليد العشوائي للمعطيات هذا المنهج يزود بحماية جيدة لان المعطيات الجديدة مختلفة تماما عن المعطيات الاصلية.

الكلمات الاساسية :الخصوصية ,البيانات الشخصية, حماية, عشوائي .

Remerciments

Je remercie dieu le tout puissant de m'avoir donné le privilège et la chance d'étudier et de suivre le chemin de la science.

J'adresse mes vifs remerciements à mon encadreur Mr. BELABED A, pour les orientations et les conseils qu'il m' prodigué durant l'évolution de mon projet.

Je tiens également à remercier Mme.CHAOUCHE L. d'avoir accepté de présider le jury de mon projet de fin d'étude.

Aussi je remercie Mr.BENAMAR A. et Mme.BENMANSOUR F. Qui ont bien voulu examiner mon travail. Leur présence va valoriser, de manière certaine, le travail que j'ai effectué.

J'adresse également ma profonde gratitude envers les professeurs de l'université ABOU BEKR BELKAID, ceux du département de l'informatique en particulier.

Enfin je remercie tous ceux qui m'ont aidé de près ou de loin dans la réalisation de ce projet de fin d'étude.

Merci à tous

DEDECACES

Je dédie ce modeste travail

A mes chers parents, que dieu les protège pour

leur amour, leurs encouragements et leurs

sacrifices.

A mes chères sœurs et mon frère ainsi qu'à toute

ma famille.

Tous mes enseignants du primaire à l'université.

A tous ceux que j'aime et qui m'aiment.

A tous mes amis.

Table des matières

Introduction générale

Introduction générale

Avec le développement rapide des base de données, de l'internet et des technologies de l'informatique, une grande quantité de données personnels peuvent être intégrées et analysées numériquement, ce qui conduit à une utilisation accrue des outils d'exploration de données pour dégager des tendances et des modes. Cela a soulevé des préoccupations universelles sur la protection de la vie privée des individus.

L'objectif de ce mémoire est de traiter une problématique universelle et d'actualité qui est la protection de la vie privée des données personnelles, l'idée de base est de créer une nouvelle approche qui se base sur la génération aléatoire des nouvelles données à partir des données originales en utilisant un classificateur automatique. Pour garder un maximum de sens entre les valeurs des attributs générés cette génération est guidée par un ensemble des règles sémantique. 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 des données personnelles, le travail est organisé comme suit :

Dans le chapitre **1** nous présentons une vue générale sur la vie privée, ses principes, les différents niveaux de protection ainsi que quelques attaques et technologies permettant la protection de la vie privée.

Le chapitre **2** est consacré à un état de l'art sur la protection des données personnelles, pour cela nous présentons les risques relatifs à la vie privée et quelques approches de protection.

Dans le chapitre **3** nous présentons notre approche pour protéger les données personnelles. En fin nous terminons par une brève conclusion qui résume notre travail et présente quelques perspectives.

CHAPITRE I : La protection de la vie privée

I. Introduction

L'extension de l'Internet et du commerce électronique ont créé des menaces sans précédent pour le respect de la vie privée du consommateur. Actuellement, les détails intimes de la vie privée de l'internaute sont exposés au vu et au su de tout venant prêt à les localiser. Souvent cependant, les consommateurs ne se rendent pas compte que chaque fois qu'ils utilisent l'Internet, ils laissent derrière eux un chemin jonché de données personnelles.

A titre d'exemple, peu de consommateurs sont au courant que les entreprises exploitent d'importantes bases de données pour dresser le profil de leurs clientèles et bien cibler leur publicité. Certaines de ces entreprises se servent de cette information pour fixer les prix « à la tête du client ». L'information personnelle et financière est systématiquement volée et détournée pour acheter des marchandises ou obtenir du crédit.

Encore plus inquiétant, de nombreux documents confidentiels, tels que les dossiers financiers et médicaux autrefois stockés dans des bases de données séparées, peuvent dorénavant être hébergés en ligne, souvent sans le consentement du consommateur ou une protection suffisante de sa vie privée.

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 protection dans vie privée ainsi que quelques attaques et enfin les technologies permettant la protection de la vie privée.

II. La notion 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 primordiale de bien comprendre que toute information non sécurisée mise en ligne peut être accessible par tout le monde. Cette

prise de conscience de l'universalité d'internet et de sa propension à diffuser rapidement une information important **[2].**

III. Différents niveaux de protection de la vie privée

On peut définir quatre propriétés principales pour la protection de la vie privée :

L'anonymat, pseudonymat, Non-**"**chaînabilité" et Non-observabilité **[4]**.

- **Anonymat :**

Requiert que d'autres utilisateurs ou sujets soient incapables de déterminer le véritable nom de l'utilisateur associé à un sujet, une opération ou un objet.

- **Pseudonymat :**

C'est l'utilisation d'un pseudonyme au lieu du vrai nom.

- **Non-chaînabilité :**

C'est l'impossibilité pour d'autres utilisateurs d'établir un lien entre les différentes opérations faites par un même utilisateur.

- **Non-observabilité :**

Consiste à ce que des utilisateurs ou des sujets ne puissent pas déterminer si une opération est en cours d'exécution.

IV. Les principes fondamentaux de protection de la vie privée

Il existe quelques principes universels liés au respect de la vie privée comme la minimisation des données, souveraineté des données, consentement explicite et la transparence.

IV.1.Minimisation des données

La première mesure de minimisation consiste que la seule information nécessaire pour compléter une application particulière devrait être collectée ou utilisée et pas plus **[3],** par exemple le commerce électronique, impliquant un client, un marchand, un service de livraison, des banques. Le marchand n'a pas besoin en général de l'identité du client, mais doit être sûr de la validité du moyen de paiement. La société de livraison n'a pas besoin de connaître l'identité de l'acheteur, ni ce qui a été acheté (sauf les caractéristiques physiques), mais doit connaître l'identité et l'adresse du

destinataire. La banque du client ne doit pas connaître le marchand ni ce qui est acheté, seulement la référence du compte à créditer, le montant, etc **[4].**

IV.2.Souveraineté des données

Lorsque des données personnelles se trouvent sur un site distant, c'est-à-dire une machine qui n'est pas sous le contrôle direct de la personne concernée (typiquement, un serveur d'une entreprise ou administration), soit pour un court moment (par exemple l'exécution d'une simple transaction), soit pour plus longtemps (par exemple des dossiers médicaux dans un hôpital), l'accès à ces données devrait être strictement limité à l'usage souhaité par leur propriétaire, c'est-à-dire la personne correspondant à ces données. Cela signifie que le propriétaire des données doit pouvoir imposer une politique de protection de la vie privée sur ses données et que le serveur qui conserve et traite ces données doit mettre en œuvre cette politique par des mécanismes de contrôle des accès à ces données. La politique en question peut définir des permissions et des interdictions précisant qui peut ou ne peut pas réaliser quelle opération sur ces données personnelles, mais aussi des obligations précisant, par exemple, que les données expirent (et donc doivent être effacées) après un délai donné suivant la terminaison de la transaction, ou que la divulgation de ces données à un tiers doit être notifiée au propriétaire par courriel, etc. Bien sûr, la politique de vie privée imposée par le propriétaire des données doit être compatible avec la politique de sécurité qui protège les biens de l'entreprise et gouverne l'exécution de l'application, et donc les accès effectifs aux données. La compatibilité entre ces deux politiques doit être vérifiée avant la divulgation par l'utilisateur de ses données personnelles **[18].**

IV.3.Consentement explicite

Ça signifie qu'avant de collecter les données personnelles d'un individu, il faut lui demander son autorisation et lui expliquer comment elles seront utilisées **[3].**

IV.4.Transparence

Ça signifie que le système ne doit pas être considère comme une boite noire dans laquelle l'individu doit avoir une confiance aveugle **[3].**

V. Les attaques sur la vie privée

V.1. Vol d'identité

Le vol d'identité renvoie au processus initial consistant à acquérir les données personnelles d'une personne à des fins criminelles.

Les voleurs d'identité s'approprient des éléments clés de renseignements personnels d'une personne de manière physique ou autrement, sans avertir et ils les utilisent pour usurper l'identité et commettre des crimes en nom d'une personne.

En plus des noms, des adresses et des numéros de téléphone, les voleurs d'identité recherchent: les numéros d'assurance sociale, les numéros de permis de conduire, les renseignements sur les cartes de crédit et les renseignements bancaires, les cartes bancaires, les cartes d'appel, les certificats de naissance et les passeports.

Les voleurs d'identité peuvent manipuler les renseignements d'une personne. Ils peuvent utiliser les identités volées pour faire des achats extravagants, ouvrir de nouveaux comptes bancaires, détourner le courrier, présenter des demandes d'emprunt, de cartes de crédit et de prestations sociales, louer des appartements et même commettre des crimes beaucoup plus graves. **[9]**

V.2. Le profilage

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 **[16].**

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 **[17].**

VI. Quelques techniques d'attaques

VI.1. Logiciels espions (spywares)

Les logiciels espions sont des logiciels qui recueillent des renseignements sur une personne à son insu. De façon générale, ces logiciels épient les gestes et les habitudes des utilisateurs sur l'Internet et font parvenir ces renseignements à des compagnies de publicité. Ces dernières se servent de ces renseignements pour établir des profils commerciaux qui les aident à mettre leurs produits en marché de façon plus efficace. Ces logiciels espions sont parfois intégrés dans des logiciels gratuits (logiciels à utilisation partagée ou shareware) que les utilisateurs peuvent télécharger sur Internet. Des contrats de licence plutôt longs (que peu de gens lisent) accompagnent souvent ces logiciels.

Les logiciels espions sont une pratique courante dans le monde de l'informatique. Même si cette pratique n'est pas très appréciée, il n'en demeure pas moins qu'elle n'est pas illégale puisque les fabricants de logiciels n'ont pas, en général, d'intentions criminelles **[7].**

VI.2. 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]**.

VI.3.Le phishing

Le phishing, traduit parfois en « hameçonnage », est une technique frauduleuse utilisée par les pirates informatiques pour récupérer des informations (généralement bancaires) auprès d'internautes.

La technique du phishing est une technique d'« ingénierie sociale » c'est-à-dire consistant à exploiter non pas une faille informatique mais la « faille humaine » en dupant les internautes par le biais d'un courrier électronique semblant provenir d'une entreprise de confiance, typiquement une banque ou un site de commerce.

Le mail envoyé par ces pirates usurpe l'identité d'une entreprise (banque, site de commerce électronique, etc.) et les invite à se connecter en ligne par le biais d'un lien hypertexte et de mettre à jour des informations les concernant dans un formulaire d'une page web factice, copie conforme du site original, en prétextant par exemple une mise à jour du service, une intervention du support technique, etc.

Dans la mesure où les adresses électroniques sont collectées au hasard sur Internet, le message a généralement peu de sens puisque l'internaute n'est pas client de la banque de laquelle le courrier semble provenir. Mais sur la quantité des messages envoyés il arrive que le destinataire soit effectivement client de la banque. Ainsi, par le biais du formulaire, les pirates réussissent à obtenir les identifiants et mots de passe des internautes ou bien des données personnelles ou bancaires (numéro de client, numéro de compte en banque, etc.).

Grâce à ces données les pirates sont capables de transférer directement l'argent sur un autre compte ou bien d'obtenir ultérieurement les données nécessaires en utilisant intelligemment les données personnelles ainsi collectées **[6]**.

VI.4.Chevaux de Troie

Un cheval de Troie est un programme informatique néfaste se présentant sous une forme bénigne. Les chevaux de Troie peuvent se présenter sous la forme d'un jeu, ou de n'importe quel programme qui peut être joint à un courriel. Le cheval de Troie est un programme exécutable. Voici quelques-unes des extensions de fichiers exécutables : exe, bat, pif, com, vbs.

Lorsqu'il est exécuté, un cheval de Troie 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 [**7]**.

VII. Les technologies de protection de la vie privée

parmi les technologies de protection de la vie privée on peut citer Privacy Enhancing Technologies (PETs) , Privacy by design.

VII.1. Privacy Enhancing Technologies (PETs)

C'est un ensemble de techniques et d'applications qui permettent à un individu de protéger ses informations personnelles pendant qu'il est en ligne **[3].**

Pour protéger la vie privée des utilisateurs d'Internet, on peut considérer plusieurs catégories de PETs: les systèmes de gestion des identités, les systèmes de communications anonymes et les systèmes d'accès anonymes aux services Internet.

a. Les systèmes de gestion des identités

La gestion d'identité (Identity Management) est un maillon clé dans la chaine de sécurité des organisations**,** Elle permet de renforcer le niveau de sécurité général en garantissant la cohérence et la traçabilité dans l'attribution des droits d'accès aux différentes ressources de système d'information, quelle que soit leur technologie et leur localisation **[13]**.

Exemples: Microsoft passport, Single Sign-On (SSO)

- **Windows Live ID**

Windows Live ID (anciennement appelé Microsoft Passport) est un service qui permet d'utiliser une adresse de messagerie et un mot de passe uniques, appelés authentifiant, pour accéder à la plupart des sites et services de Microsoft ainsi que ceux de ses partenaires choisis.

Il permet d'enregistrer ces authentifiant (adresse de messagerie et mot de passe) à un site ou un service qui utilise Windows Live ID, ou au site Web Windows Live ID. Microsoft utilise cette identité unique pour aider à améliorer l'authentification de Windows Live ID et pour la protection contre les pourriels et l'utilisation malveillante du compte **[14].**

- **Single Sign-On (SSO)**

L'authentification unique (Single Sign-On), est une méthode permettant à un utilisateur de ne procéder qu'à une seule authentification pour accéder à plusieurs ressources (machines, systèmes, réseaux)

L'objectif du SSO est ainsi de propager l'information d'authentification aux différents services du réseau, voire aux autres réseaux et d'éviter ainsi à l'utilisateur de multiples identifications par mot de passe **[15].**

b. les systèmes de communications anonymes

L'écoute passive (c'est-à-dire l'observation sans modification) de communications est une menace importante contre la vie privée puisque, même lorsque le contenu d'une communication est chiffré, la simple observation des adresses source et destination peut révéler des informations sensibles. Ainsi, sur Internet, une adresse IP permet d'identifier un utilisateur, mais aussi les services auxquels il se connecte. Par exemple, l'adresse IP 67.192.121.169 correspond au serveur Web des Alcooliques Anonymes. Simplement en observant juste les adresses source et destination d'une connexion sur Internet, on peut déduire : l'identité de l'utilisateur, sa localisation) au moment de la connexion, ses domaines d'intérêt (par exemple les problèmes d'alcoolisme) **[18]**.

Les PETs permettant de communiquer de manière anonyme dans un réseau.

Exemples : Mix, Tor, Crowds, etc.

- **Les Mix**

Le Mix est un routeur qui cache le lien entre les messages entrants et sortants par un mécanisme de chiffrement et de permutation des messages **[11].**

Le fonctionnement d'un Mix simple (voir la figure I.1)

Le mix reçoit en entrée plusieurs paires du type (message ; adresse du destinataire) qui ont été préalablement chiffrées puis déchiffre les messages et envoie en sortie les messages à leurs destinataires correspondants.

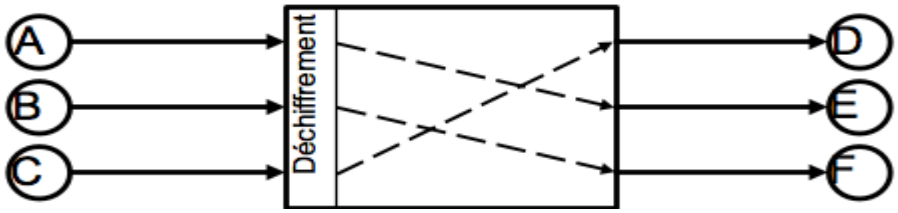

Figure I.1 : un Mix simple **[11]**

- **Crowds**

Protocole de communication anonyme qui protège l'anonymat de l'envoyeur d'un message en le routant de manière aléatoire vers des groupes d'utilisateurs similaires. L'idée principale est de cacher l'origine d'un message en le dispersant. **[11]**

- **Tor**

Tor (The Onion Router) est un réseau mondial décentralisé, anonyme organisée en couche autour de routeurs qui jouent le rôle de nœud, Permet d'anonymiser tout type de communication faite sur Internet **[11]**.

Exemples : sites web visités, messagerie instantanée, courriel.

Le logiciel Tor est basé sur un réseau, composé de tous les ordinateurs/serveurs utilisant Tor. Le principe du réseau est le suivant : le routage-oignon des informations. En bref, un paquet de données, passe directement au serveur (habituellement). Cependant, en utilisant Tor, ce logiciel va faire passer les données chiffrées d'ordinateur en ordinateur (tous utilisant Tor), jusqu'à atteindre le serveur. Chaque relais est appelé un "nœud". Le dernier nœud est chargé de déchiffrer tous les paquets et de les envoyer au serveur **[12]**.

c. les systèmes d'accès anonymes aux services Internet

Fournir des communications anonymes ne suffit pas pour obtenir un accès anonyme à un service : les messages envoyés au fournisseur de service peuvent contenir des informations identifiants, qu'il faut effacer ou transformer par un mandataire (proxy) avant qu'elles ne soient transmises au fournisseur. Cette transformation dépend de la sémantique du message (C'est-à-dire de la signification de son contenu), et la tâche peut donc être très ardue. Si le mandataire est dédié à un service spécifique, il est relativement aisé d'analyser la syntaxe des en-têtes, par exemple, pour éliminer une partie des informations sensibles. Cependant, la structure des messages requis par la

plupart des services peut être très variable, et donc très difficile à anonymiser. Bien sûr, utiliser un seul mandataire pour accéder à un service suppose que l'on ait confiance dans ses administrateurs, puisqu'ils peuvent enregistrer des informations sensibles sur l'application comme sur les communications **[18]**.

VIII. Privacy by design

C'est l'intégration de la problématique du respect de la vie privée de la conception d'un système. Considère la question de la vie privée a priori, plutôt que de réagir a posteriori une fois que le système a été déployé et qu'on constate un bris de vie privée **[3]**.

- **Les principes fondamentaux**

Les principes fondamentaux de le Privacy by Design décrivent les mesures proactives nécessaires pour faire de la protection de la vie privée le mode implicite de fonctionnement de toutes les organisations.

- **Proactive et non réactif**

Le Privacy by Design est une approche qui se caractérise par des mesures proactives plutôt que réactives. Il prévoit et empêche des événements de la vie privée avant qu'ils se produisent. En bref, Privacy by Design vient avant le fait, non pas après, Il s'agit de l'application du proverbe : « mieux vaut prévenir que guérir ».

- **La vie privée comme un réglage par défaut**

Nous pouvons tous être certains d'une chose (les règles par défaut). Privacy by Design vise à offrir le maximum de la vie privée en faisant en sorte que les données personnelles sont automatiquement protégées dans un système d'information et de gestion. Si une personne ne fait rien, leur vie privée demeure intact. Aucune action n'est exigée de la part de l'individu pour protéger leur vie privée, elle est établie dans le système par défaut.

- **La vie privée est intégrée dans la conception**

Privacy by Design est intégré dans le design, l'architecture des systèmes et les pratiques commerciales.

Le résultat est que la vie privée devient une composante essentielle du fonctionnement. La vie privée fait partie intégrante du système, sans diminuer la fonctionnalité.

- **Fonctionnalité complète (à somme positive)**

Privacy by Design, ne doit pas empêcher la mise en œuvre d'autres fonctionnalités, mais doit être un avantage concurrentiel. Par exemple : la prise en compte de la vie privée ne doit pas empêcher un haut niveau de sécurité. Il est possible de réaliser plusieurs objectifs à la fois sans les compromettre.

- **Protection du cycle de vie complet**

Privacy by Design, ayant été intégrés dans le système avant l'assemblage du premier élément alors des mesures de sécurité solides sont essentiels à la vie privée, du début à la fin. Cela garantit que toutes les données sont bien conservées puis détruits à la fin du processus en toute sécurité.

- **Visibilité et transparence**

Privacy by Design vise à assurer à tous les intervenants que toutes les pratiques sont exploitables selon les promesses et les objectifs énoncés. Ses composants et les opérations restent visibles et transparentes, pour les utilisateurs et les fournisseurs.

- **Respect de La vie privée de l'utilisateur**

La conception exige à des architectes de conserver les intérêts de l'individu le plus élevé en offrant des mesures telles que la vie privée forte par défaut **[10]**.

➢ **Les domaines d'application du « privacy by design »**

La protection intégrée de la vie privée s'applique aux nouvelles technologies, et notamment aux systèmes informatiques et aux infrastructures des réseaux. Ses principes peuvent s'appliquer à tous les types de renseignements personnels, mais ils devraient l'être avec une rigueur particulière aux données délicates telles que les renseignements médicaux et financiers. Plus les données sont délicates, plus les mesures de protection de la vie privée tendent à être strictes **[8].**

Exemple : aux réseaux sociaux, la santé en ligne, les systèmes de transport intelligents,…

IX. Conclusion

Nous avons présenté dans ce chapitre des notions théoriques sur la protection de la vie privée, En commençant par une définition avec les principes et les différents niveaux de protection dans la vie privée ainsi que quelques attaques et enfin nous avons terminé par les technologies permettant la protection de la vie privée.

Le chapitre suivant sera consacré à un état de l'art sur la protection des données personnelles.

CHAPITRE II : La protection des données personnelles

Etat de l'art

I. Introduction

La protection des données personnelles est devenue une question essentielle dans notre société, soucieuse de mettre l'informatique au service du citoyen, Avant d'être un problème juridique, l'utilisation des données personnelles est un problème éthique, philosophique et politique. Il s'agit de concilier progrès et libertés. Les données sont très diverses, les raisons de constituer des fichiers de données sont de plus en plus étendues, Les possibilités sont infinies, les risques le sont aussi.

L'enjeu de la protection des données personnelles est dès lors compréhensible. Ces données revêtent un caractère fondamental, et relèvent du champ des libertés individuelles, il est donc indispensable de les protéger. Cependant, les données personnelles sont omniprésentes, on les retrouve sous différentes formes, dans divers secteurs, ce qui pose des problèmes complexes rendant la protection difficile Il en résulte que, si la mise en œuvre de la protection aux problèmes classiques est assez efficace, la protection des données personnelles face à Internet est loin d'être optimale.

Ce chapitre est consacré à un état de l'art sur la protection des données personnelles, pour cela nous commençons par une définition des données à caractère personnelle ainsi que les risques relatifs à la vie privée, les attaques sur ces données et enfin nous présentons quelques approches de protection.

II. Définition des données personnelles

On appelle **données personnelles** les informations qui permettent, notamment sur Internet, d'identifier directement ou indirectement une personne physique [19].

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 **[19].**

III. Les risques relatifs à la vie privée

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 **[20].**

- **L'exemple suivant illustre une telle possibilité [21]** :

On considère la table II.1, qui illustre les données médicales. Dans ce tableau, que nous appelons la table privée (TP), les données ont été rendus anonymes par la suppression des noms et les numéros de la Sécurité Sociale (NSS) de sorte à ne pas divulguer explicitement l'identité des patients.

Cependant, les valeurs d'autres attributs, notamment de race, la date de naissance, le Sexe, le code ZIP et l'état matrimonial peuvent apparaître dans certains tables externes (la liste électorale publique par exemple illustré dans la Table II.2) conjointement avec l'identité individuelle, qui permet leur suivi. Ainsi, le code ZIP, la date de naissance, le sexe et l'état matrimonial peuvent être liés à la liste électorale dans la Table II.2 et on peut révéler ainsi le Nom, l'adresse et la Ville d'un individu. Dans notre exemple et dans la Table privé, il n'y a qu'une seule femme divorcée (F) né le 64/04/12 et vivant dans la zone 94142. La combinaison de ces informations avec celles de la table II.2, si unique, identifie d'une manière unique que « Sue J. Doe » habitant à « 900 Market Street, San Francisco souffre de l'hypertension artérielle **[21].**

SSN	Nom	Race	DN	Sexe	Zip	Etat civil	Maladie
		asiatique	64/04/12	F	94142	Divorcé	Hypertension
		asiatique	64/09/13	F	94141	Divorcé	obésité
		asiatique	64/04/15	F	94139	marié	Douleur à la poitrine
		asiatique	63/03/13	H	94139	marié	obésité
		asiatique	63/03/13	H	94139	marié	souffle court
		noir	63/03/18	F	94138	unique	souffle court
		noir	64/09/27	F	94139	unique	obésité
		blanc	64/09/27	F	94139	unique	Douleur à la poitrine
		blanc	64/09/27	F	94141	veuve	souffle court

Table II.1. Tableau privée **[21]**

Nom	Adresse	ville	Zip	DN	Sexe	Statut
……..	…….	……	…….	……	……	……..
Sue J.Doe	900 Market St	San Francisco	94142	64/04/12	F	divorcé
……..	……	…….	……	……	……	……

Table II.2. Tableau public (liste électorale) [**21]**

IV. Les attaques des données personnelles

Il y a risque de violation de vie privée quand l'identité d'un individu est liée à un enregistrement ou quand elle est liée à une valeur d'un attribut sensible. Ces brèches d'anonymat sont appelées : record linkage, attribute linkage, table linkage et l'attaque probabilist.

IV.1. L'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-anonymity (voir la section IV.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” **[30].**

IV.2. L’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-diversity[voir la section IV.2],t-closeness(voir la section IV.3), Personalized Privacy(voir la section IV.6) **[30].**

IV.3.L’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-Presence (voir la section IV.4) a été proposé pour combattre les attaques par ”Table linkage” **[26].**

IV.4.l’attaque probabiliste

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

Le modèle differential Privacy a été proposé pour combattre les attaques ”probabilistes” **[26].**

V. Les opérations de protection

En règle générale, la table d'origine ne satisfait pas l'exigence de confidentialité spécifiée et le tableau doit être modifié avant d'être publié. La modification se fait par l'application d'une séquence d'opérations d'anonymisation à la table. Une opération d'anonymisation peut être : une généralisation, suppression, anatomization et permutation, ou perturbation.

V.1. Généralisation et Suppression

La Généralisation consiste à substituer les valeurs d'un attribut donné avec des valeurs plus générales. A cet effet, la notion de domaine (l'ensemble des valeurs que l'attribut peut prendre) est définie pour pouvoir appliquer la généralisation (figure II.1) **[24].**

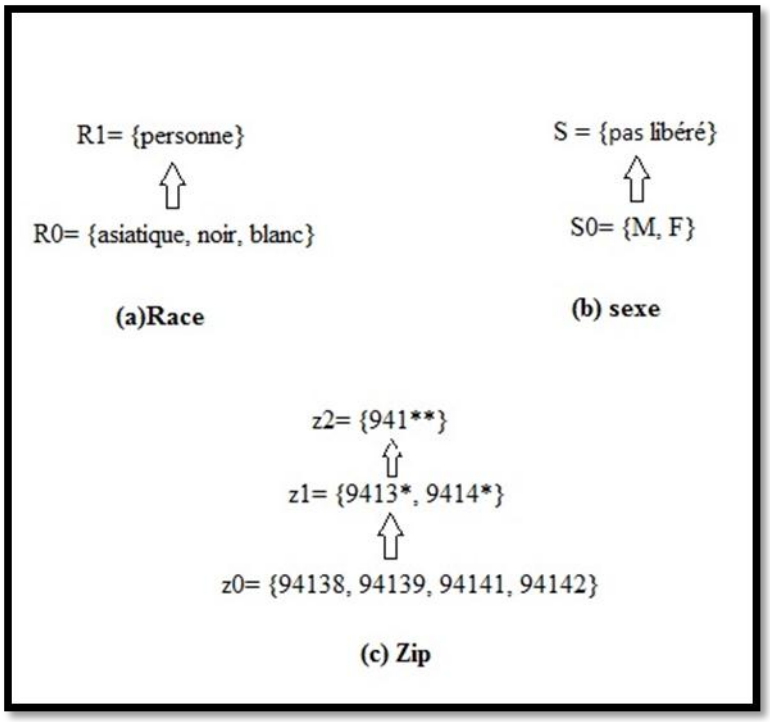

Figure II.1. Des exemples de hiérarchies de généralisation **[24].**

Le principe de suppression peut réduire le montant de la généralisation nécessaire.
La Suppression sert donc à «modérer» le processus de généralisation quand un nombre limité de valeurs aberrantes obligerait une grande quantité de généralisation. Par exemple, considérons les tables généralisées dans la figure.II.2 Les tuples en italique sont celles qui devraient être supprimés dans chaque tableau généralisée **[24].**

Race:R0	Le code ZIP:Z0
Asiatique	*94142*
Asiatique	*94141*
Asiatique	94139
Asiatique	94139
Asiatique	94139
Noir	*94138*
Noir	*94139*
Blanc	*94139*
Blanc	*94141*

(a)

Race:R1	Le code ZIP:Z0
personne	*94142*
personne	94141
personne	94139
personne	94139
personne	94139
personne	*94138*
personne	94139
personne	94139
personne	94141

(b)

Race:R0	Le code ZIP:Z1
Asiatique	9414*
Asiatique	9414*
Asiatique	9413*
Asiatique	9413*
Asiatique	9413*
Noir	9413*
Noir	9413*
Blanc	*9413**
blanc	*9414**

(c)

Figure II.2. Un exemple d'un tableau privé (a) et ses généralisations **[21]**

V.2. Anatomization et Permutation

Contrairement à la généralisation et à la suppression, l'anatomization ne modifie pas les quasi-identifiants « QID » comme l'âge, le code zip, ou les attributs sensibles comme la maladie, le salaire, mais dissocier la relation entre les deux. Plus précisément, la méthode divise les données sur les « QID » et les données sur attributs sensibles en deux tables distinctes : une table de quasi-identificateurs « QIT » qui contient les attributs « QID », et une table sensible « ST » qui contient les attributs sensibles, et les deux tables « QIT » et « ST » possèdent un attribut commun, GroupID. Tous les enregistrements du même groupe auront la même valeur du GroupID dans les deux tables, et sont donc liées à des valeurs sensibles dans le groupe de la même manière. Si un groupe a « l » valeurs sensibles distincts et chaque valeur distincte se produit qu'une seule fois dans le groupe, alors la probabilité de lier un enregistrement à une valeur sensible par GroupID est de « 1 /l ». Une attaque de type liaison peut être réduite en augmentant la valeur de « l ».L'avantage majeur de l'anatomie est que les données de « QIT » et « ST » ne sont pas modifiées. **[26].**

Les tableaux suivant représentent un exemple d'une telle opération, le tableau II.3.a représenté le tableau d'origine, la table II.3.b est une table intermédiaire nécessaire pour déterminer les GroupID, les deux tableaux II.3.c et II.3.d représentent le résultat final de cette opération.

Age	Sexe	Maladie (Sensible)
30	Homme	Hépatite
30	Homme	Hépatite
30	Homme	VIH
32	Homme	Hépatite
32	Homme	VIH
32	Homme	VIH
36	Femme	Grippe
38	Femme	Grippe
38	Femme	Cœur
38	Femme	Cœur

Table II.3.a. Tableau d'origine **[26]**

Age	Sexe	Maladie (Sensible)
[30-35]	Homme	Hépatite
[30-35]	Homme	Hépatite
[30-35]	Homme	VIH
[30-35]	Homme	Hépatite
[30-35]	Homme	VIH
[30-35]	Homme	VIH
[35-40]	Femme	Grippe
[35-40]	Femme	Grippe
[35-40]	Femme	Cœur
[35-40]	Femme	Cœur

Table II.3.b. Tableau intermédiaire **[26]**

Age	Sexe	GroupID
30	Homme	1
30	Homme	1
30	Homme	1
32	Homme	1
32	Homme	1
32	Homme	1
36	Femme	2
38	Femme	2
38	Femme	2
38	Femme	2

Table II.3.c. Tableau quasi-identifiant (QIT) **[26]**

GroupID	Maladie (Sensible)	GroupID
1	Hépatite	3
1	VIH	3
2	Grippe	2
2	Cœur	2

Table II.3.d. Tableau sensible (ST) **[26]**

La Permutation à presque le même principe que l'anatomization, L'idée est de dissocier la relation entre une quasi-identifiant et un attribut sensible par la division d'un ensemble des données en groupes et traînant leurs valeurs sensibles à l'intérieur de chaque groupe **[26].**

V.3. Perturbation

L'idée générale est de remplacer les valeurs de données d'origine avec des valeurs de données synthétiques, donc l'information statistique calculée à partir des données perturbées ne diffère pas significativement de l'information statistique calculée à partir des données d'origine, donc l'attaquant ne peut pas récupérer des informations sensibles à partir des données publiées **[26].**

V.4. échantillonnage

Le tableau de micros données protégé sous forme d'un un échantillon de la table d'origine. En d'autres termes, la table de micro donnée protégé ne comprend que les données (tuples) d'un échantillon de l'ensemble de la population. Comme il ya une incertitude à savoir si un Individu particulier est dans l'échantillon ou non, le risque de ré-identification dans le micro donnée publié diminue Cette technique fonctionne sur les attributs catégoriques seulement **[32].**

VI. Les approches de la protection des données personnelles

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-Closeness, δ-Présence et differential Privacy ainsi que la Personalized Privacy .

VI.1. le modèle 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 **[27].**

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

Le quasi-identificateur d'une table T, dénoté « QIT », est un ensemble d'attributs de T qui, si utilisés conjointement, peuvent mener à l'identification d'un individu avec une probabilité égale à 1. L'objectif principal de la méthode k-anonymat est de transformer

une table de manière à ce que personne ne puisse établir de lien entre la table T et un individu avec une probabilité inférieure à 1/k.

On dit qu'une table « T » est k-anonyme en rapport avec un quasi-identificateur « QIT » si et seulement si, pour tout enregistrement « r » dans « T », il existe au moins (k − 1) autres enregistrements dans T qui ne peuvent être distingués de « r » par rapport à « QIT » **[22].**

Exemple :

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

Table II.4. Tableau initial des données **[23].**

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

Table II.5. Tableau de 4-anonyme **[23].**

La **Table II.4** présente les enregistrements médicaux d'un hôpital. On remarque que la table ne contient pas d'attributs identifiant comme le nom, le numéro de sécurité sociale, etc..

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

La **Table II.5** montre un tableau 4 anonyme dérivé du la Table II.4 ("*" représente une valeur supprimée, par exemple, "code zip = 1485 *" signifie que le code est dans l'intervalle [14850-14859] et «âge = 3 *" signifie que l'âge est compris entre [30 − 39]) **[30].**

VI.2. le modèle l-diversité

La l-diversité est une technique d' anonymisation plus avancée que le k-anonymat dans la mesure où elle garantit en plus que dans un groupe de k individus il y aura au moins « l »valeurs sensibles distinctes (voir la **table II.6.a**, la **table II.6.b**). En effet, si on n'effectue pas cette vérification, il se pourrait que les valeurs sensibles ne soient que 1-diverses et donc qu'on puisse retrouver la valeur sensible associée à un individu, bien que cet individu ait été anonymisé par une technique de k-anonymat **[25]**. Dans la **table II.5** si un attaquant est sûr qu'un individu d'âge<30 est dans la table alors il peut inférer avec une probabilité de 1 qu'il souffre d'un cancer.

Exemple :

130**	**<30**	*****	**Heart Disease**
130**	**<30**	*****	**Cancer**
130**	**<30**	*****	**Infection virale**

Table II.6.a. Les données sont 3-anonymes et 3-diverses

130**	**3***	*****	**Cancer**
130**	**3***	*****	**Cancer**
130**	**3***	*****	**Cancer**

Table II.6.b. Les données s sont 3 anonymes et 1-diverses

Un autre exemple est celui de la **table II.7.a**, un hôpital qui possède un ensemble de données (**Table II.7.a**) et qu'il veuille le rendre disponible à la recherche. Il décide d'apporter les modifications à l'ensemble de données de manière à ce qu'il soit impossible de distinguer chacun des items de l'ensemble d'au moins un autre item. L'ensemble est considéré 2-diverse dans ce cas et le résultat est présenté à la Table **II.7.b**. Par exemple, si un attaquant savait que l'information de Tom, un jeune homme de 21 ans était dans l'ensemble, il ne pourrait prédire le diagnostic de la maladie dont il souffre avec une probabilité supérieure à 1/l où l = 2. Dans ce cas, tout ce qu'il pourrait apprendre est que Tom souffre d'asthme ou de la grippe. Plus tard, l'ensemble de données est mis à jour et de nouveaux items sont ajoutés avec comme résultat données présentée à la **Table II.7.c**. Les nouvelles données sont présentées en italique dans la table. L'hôpital publie donc une nouvelle version de sa table 2-diverse et cette version est présentée à la **Table II.7.d**. La vie privée de Tom est toujours protégée dans cette nouvelle version, mais ce n'est pas le cas pour les autres **[22].**

Age	Sexe	Diagnostic
21	Homme	Asthme
23	Homme	Grippe
52	Homme	Alzhiemer
57	Femme	Diabéte

Table II.7.a. Ensemble de données initial **[22].**

Age	Sexe	Diagnostic
[21-25]	Homme	Asthme
[21-25]	Homme	Grippe
[50-60]	Personne	Alzhiemer
[50-60]	Personne	Diabéte

Table II.7.b. Ensemble de données initial anonymisé en divers **[22].**

Age	Sexe	Diagnostic
21	Homme	Asthme
23	Homme	Grippe
52	Homme	Alzhiemer
57	Femme	Diabéte
27	*Femme*	*Cancer*
53	*Homme*	*Cardiaque*
59	*Femme*	*Grippe*

Table II.7.c. Nouvel ensemble de données mis à jour **[22]**

Age	Sexe	Diagnostic
[21-30]	Personne	Asthme
[21-30]	Personne	Grippe
[21-30]	Personne	Cancer
[51-55]	Homme	Alzhiemer
[51-55]	Homme	Cardiaque
[56-60]	Femme	Grippe
[56-60]	Femme	Diabéte

Table II.7.d. Nouvel ensemble anonymisé en 2-diverse**[22].**

VI.3.le modèle t-closeness

Une classe d'équivalence est dite avoir t-closeness si la distance entre la distribution d'un attribut sensible dans cette classe et la distribution de l'attribut dans la table entière n'est pas supérieure à un seuil t. Un tableau est dit avoir t- closeness si toutes les classes d'équivalence ont t- closeness **[28].**

VI.4. le modèle δ-Presence

C'est la probabilité de déduire la présence de l'enregistrement de toute victime potentielle dans un intervalle spécifié $\delta = (\delta min, \delta max)$. Formellement, étant donné un tableau E publique et une table T privée, où $T \subseteq E$, une table généralisées T* satisfait $(\delta min, \delta max)$-présence si $\delta min \leq P(t \in T \mid T^*) \leq \delta max$ pour tout $t \in E$. δ-présence peut prévenir indirectement l'enregistrement et les liens entre les attributs parce que si l'attaquant a au plus δ% de confiance que l'enregistrement de la victime cible est présent dans le tableau publié, alors la probabilité d'un lien succès de son enregistrement et de l'attribut est sensible au plus δ%. Bien que δ-présence soit un modèle de vie privée relativement «sûre», il suppose que l'éditeur de données a accès à la même table E comme l'attaquant fait. Cela peut ne pas être une hypothèse pratique **[26].**

Exemple :

Les tableaux suivant représente un exemple sur le risque de la vie privée dans δ-présence où l'adversaire sait E et veut identifier les tuples dans les données privé T.

Les attributs dans le tableau privé (**Table II.8.b**) est un sous-ensemble de celui de l'ensemble de données dans le tableau publique. L'attribut "sen" ne fait pas partie du la table publique (**Table II.8.a**), mais précise qui tuples sont dans l'ensemble de données privées (**Table II.8.b**).

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

Table II.8.a. Tableau publique (E) **[29].**

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

Table II.8.b. Tableau privée (T) **[29].**

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

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

Table II.9.a. Tableau publique (E*) **[29].**

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

Table II.9.b. Tableau privée (T*) **[29].**

VI.5.Le concept « differential Privacy »

Ce concept a été introduit récemment comme modèle de protection de données personnelles. Ce modèle est largement adopté par la communauté scientifique du fait qu'il assure une protection stable et indépendante des connaissances a priori et a posteriori (avant et après l'accès à des données publiées) de l'attaquant sur un individu particulier. Autrement dit La « differential Privacy » exige que le risque qu'un individu soit victime d'une attaque sur son confidentialité (privacy) ne devrait pas augmenter sensiblement à la suite de la participation à une base de données statistiques.La majorité des méthodes proposées pour réaliser une « differential Privacy » utilisent une table de contingence (contingency table) en

ajoutant un bruit sur la fréquence de chaque groupe. La table de contingence est issue aussi d'une généralisation de la table des données originale. La problématique consiste à trouver un algorithme *A* qui vérifier « ε- differential Privacy » c-à-dire :

Pour toute base *D* et pour tout uplet $t \in D$, et pour tous ensemble $S \in Rang(A)$ l'inégalité suivante soit vérifiée :

$$Pr[A(D)=S] \leq e^{\varepsilon} Pr[A(D_{\pm t}].$$

Ou : $e^{\varepsilon} \cong 1+\varepsilon$ et $D_{\pm t}$ est une base qui ne diffère de D que sur un seul uplet t **[31].**

VI.6.Personalized Privacy

La Confidentialité personnalisé propose la notion de vie privée personnalisée pour permettre à chaque propriétaire d'enregistrement de spécifier son propre niveau de confidentialité. Ce modèle suppose que chaque attribut sensible à un arbre taxonomique et que chaque propriétaire d'enregistrement spécifie un nœud à garder dans cet arbre. La vie privée d'un enregistrement propriétaire est violée si un attaquant est capable de déduire une valeur de domaine sensible dans le sous arbre de son nœud garde avec une probabilité, appelée probabilité de manquement, supérieur à un certain seuil **[26].**

Exemple :

Supposons que le VIH et le SRAS sont des nœuds enfants des maladies infectieuses dans l'arbre de taxonomie. Un patient Alice VIH peut définir le nœud garde aux maladies infectieuses, ce qui signifie qu'elle permet aux gens d'en déduire qu'elle a certaines maladies infectieuses, mais pas tel ou tel type de maladie infectieuse. Un autre patient atteint du VIH, Bob, ne le dérange pas de divulguer son dossier médical, alors il ne fixe pas de nœud garde pour cet attribut sensible **[26].**

VII. Conclusion

Nous avons présenté dans ce chapitre un état de l'art sur la protection des données personnelles, En commençant par une définition des données à caractère personnelle, ainsi que les risques relatifs à la vie privée , les attaques sur vie privée et enfin nous avons terminé par quelque approche de protection.

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 les données personnelles.

Nous commençons par décrire le principe de cette approche et ses détails et nous terminons par un test et quelques résultats.

II. Le principe de l'approche

II.1.La formulation de problème

Avant de présenter les détails de notre approche, nous définissons quelques concepts de base :

Soit un tableau privée « TP » qui illustre les données originales d'un ensemble de personnes, le tableau contient des attributs ($A=a_1, a_2, ... a_n$) et divise ces dernier en deux groupes : les attributs sensibles ($S=s_1, s_2, \ldots s_m$) et les attributs non sensibles ($A=a_1, a_2, ...a_{m1}$) telle que $m_1+m=n$.

Notre objectif est de générer des nouvelles données dans la table « TG » à partir des données originales de la table « TP » de tel sort à protéger la vie privée des personnes.

II.2.les étapes de génération

Dans cette approche nous avons utilisé un classificateur automatique pour générer des nouvelles données (table « TG ») à partir des données originales de la table « TP ». Les nouvelles données générées diffèrent totalement des données originales ce qui implique une grande protection. Cette procédure se déroule en trois étapes (figure III.1) :

- **Etape 1: la génération des attributs non sensibles**

  Cette étape consiste à générer aléatoirement un ensemble de valeurs pour les attributs non sensibles. Cette étape est guidée par un ensemble de règles sémantiques appliquées sur les données générées pour éviter les cas réellement impossible, par exemple « un enfant ne peut être marié».

- **Etape 2 : la génération du classificateur**

  L'objectif est de générer un modèle de classificateur ce modèle est issu d'un mécanisme d'apprentissage sur les données de la table « TP », il prend comme entrée un ensemble d'attributs jugés non sensibles (les attributs $A_1 \ldots A_n$ dans la

figure III.1) et essaye de prédire l'attribut sensible « S » (ex : la maladie d'une personne ou son revenu).

- **Etape 3 : la prédiction de la classe sensible**

En utilisant le classificateur généré à l'étape 2 et les données de l'étape 1, on peut prédire les différentes valeurs de l'attribut sensible « S ». La fin de cette phase produit un ensemble de données générées « TG » qui sera publié au lieu de l'ensemble des données « TP ».

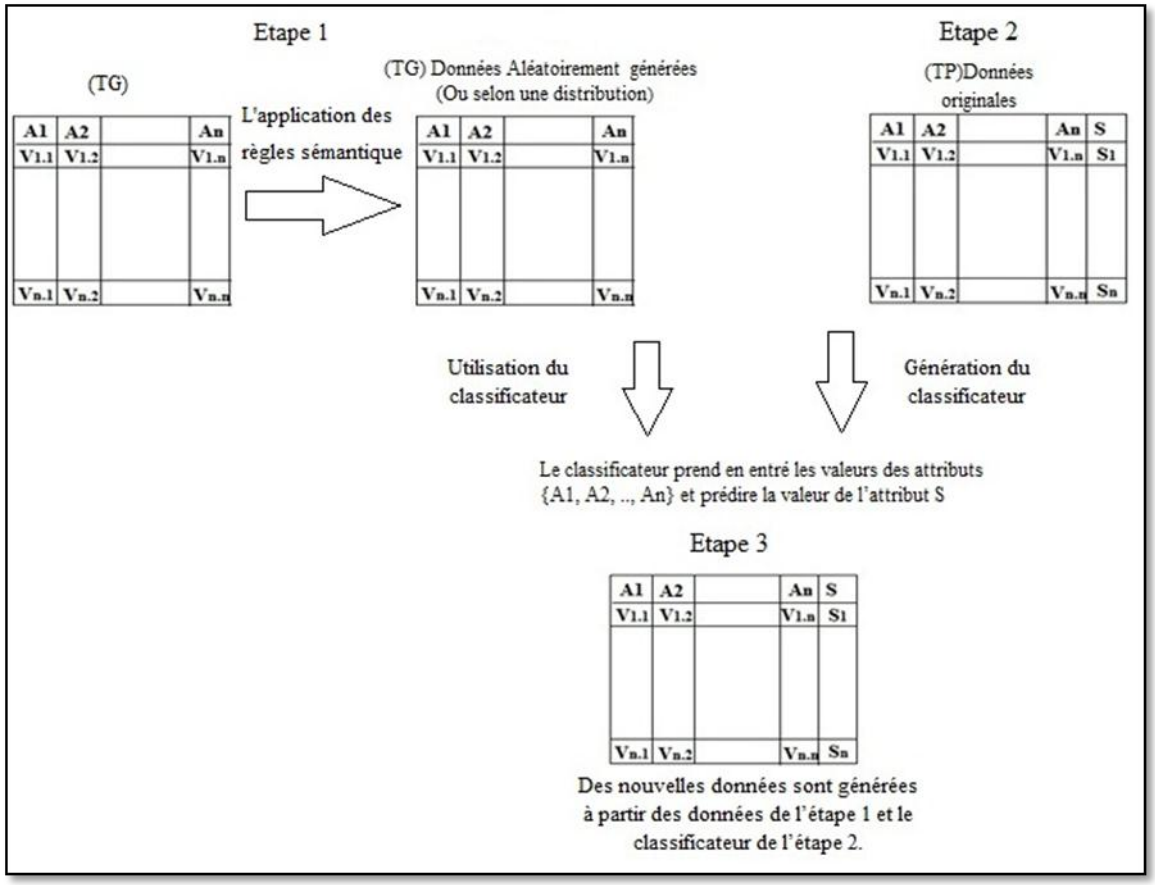

Figure III.1. Les étapes de génération des données

II.3.Le mécanisme d'évaluation

En comparant les performances entre les données générées (Table « TG ») et les données originales (Table « TP), le mécanisme d'évaluation se déroule en deux phases (voir la figure III.2) :

- **Phase 1 :**

L'objectif de cette phase est de trier les performances du classificateur issu de l'utilisation de la table « TP » (comme base d'apprentissage et base de test).

- **Phase 2 :**

L'objectif de cette phase est de trier les performances du classificateur issu de l'utilisation de la table « TG » » comme base d'apprentissage et les données originale de la table « TP » comme base de test.

- **Phase 3 :**

Dans cette phase on va comparer les performances des deux classificateurs.

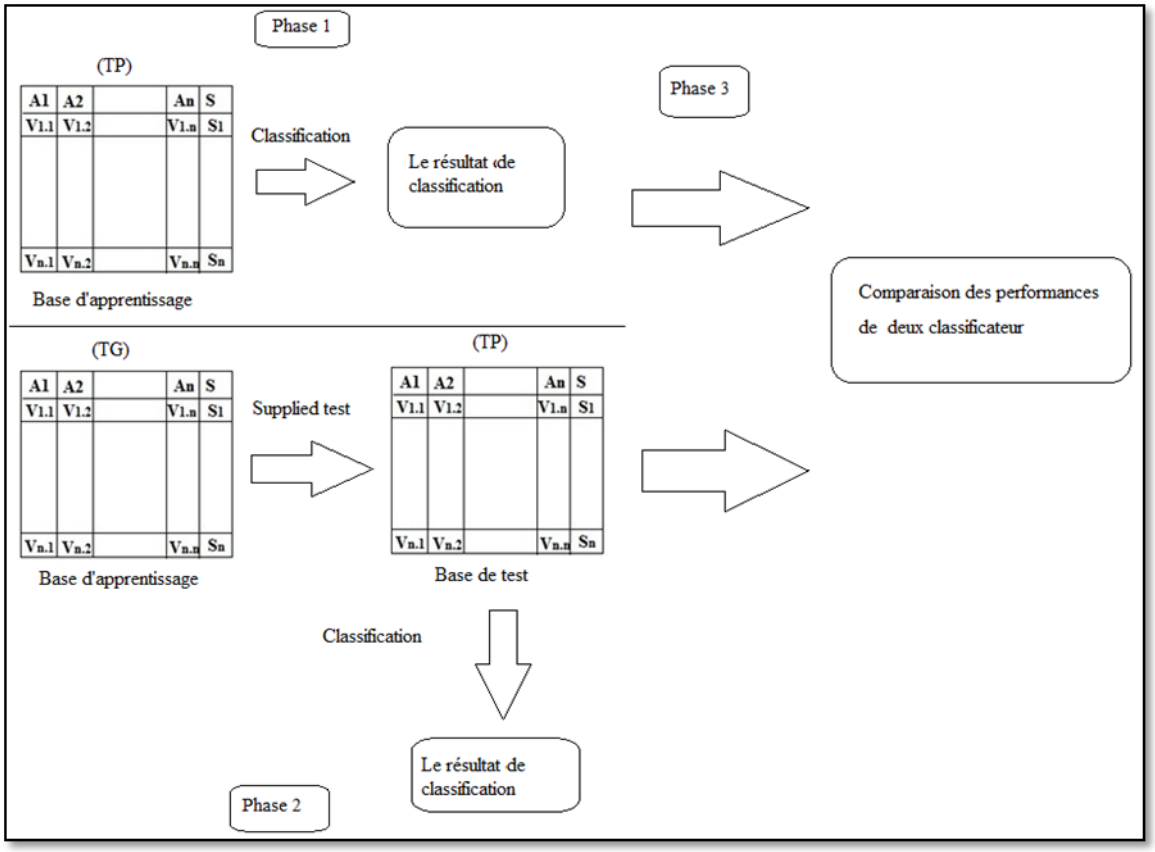

Figure III.2. Le déroulement des tests

III. Implémentation et Résultats

IV.1.la description de la base

Dans cette partie on va tester notre approche sur la base «Recensement de revenu», **[37],** les attributs de cette base sont divisés en deux groupes : les attributs sensibles (salaire) et les attributs non sensibles (age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country).

Le tableau suivant décrit quelques caractéristiques de cette base **[37]** :

Nombre d'instances :	48842
Nombre d'attributs :	15
Valeurs manquantes ?	Oui
Caractéristiques des attributs :	Multi variée

Nous avons fait un prétraitement sur cette base pour supprimer les valeurs manquantes des individus, la base obtenue contient « **30162 »** instances.

IV.2.Les outils de développement

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

- Langage de programmation : Java
- L'environnement de développement (IDE) : Netbeans 7.0.1.
- L'API Weka 3.6

a. Java comme langage de programmation :

Le langage **Java** est un langage de programmation informatique orienté obje, La particularité et l'objectif central de Java est que les logiciels écrits dans ce langage doivent être très facilement portables sur plusieurs systèmes d'exploitation tels que UNIX, Windows, Mac OS ou GNU/Linux, avec peu ou pas de modifications. Pour cela, divers plate-formes et

frameworks associés visent à guider, sinon garantir, cette portabilité des applications développées en Java **[33].**

b. Neatbeans comme environnement de développement :

NetBeans est un environnement de développement intégré (EDI), NetBeans constitue par ailleurs une plateforme qui permet le développement d'applications spécifiques (bibliothèque Swing (Java)) **[34].**

c. l'API weka :

WEKA est un Banc d'outil complet pour l'apprentissage automatique et l'extraction de données. Ses principaux atouts résident dans la zone de classification **[35].**

IV.3.Application et test

Notre objectif est de générer des nouvelles données dans un tableau « TG » à partir des données originale de la table « TP ».

a. les étapes de génération des données

On a utilisé un classificateur automatique pour générer des nouvelles données (table « TG ») à partir des données originales de la table « TP ». Cette procédure se déroule en trois étapes :

- **Etape 1 : la génération des attributs non sensibles**

  Dans Cette étape on a générer aléatoirement un ensemble de valeurs pour les attributs non sensibles on a générer 500 individus. Cette étape est guidée par un ensemble de règles sémantiques appliquées sur les données générées pour éviter les cas réellement impossible, par exemple « un enfant ne peut être marié». «On ne peut pas dire que une femme mariée ou un homme marié sont jamais marié » « on ne peut pas dire qu'un prof ne travaillez pas ».

- **Etape 2 : la génération du classificateur**

  Pour générer un modèle de classification nous avons utilisé, Dans cette étape le classificateur J48 **[36].**

- **Etape 3 : la prédiction de la classe sensible**

Dans cette étape on a utilisé le classificateur généré à l'étape 2 et les données de l'étape 1, on peut prédire les différentes valeurs de l'attribut sensible « S ». La fin

de cette phase produit un ensemble de données générées « TG » qui sera publié au lieu de l'ensemble des données « TP ».

- L'interface suivant permet de charger la base originale, créer des nouvelles données et visualiser les données générée.

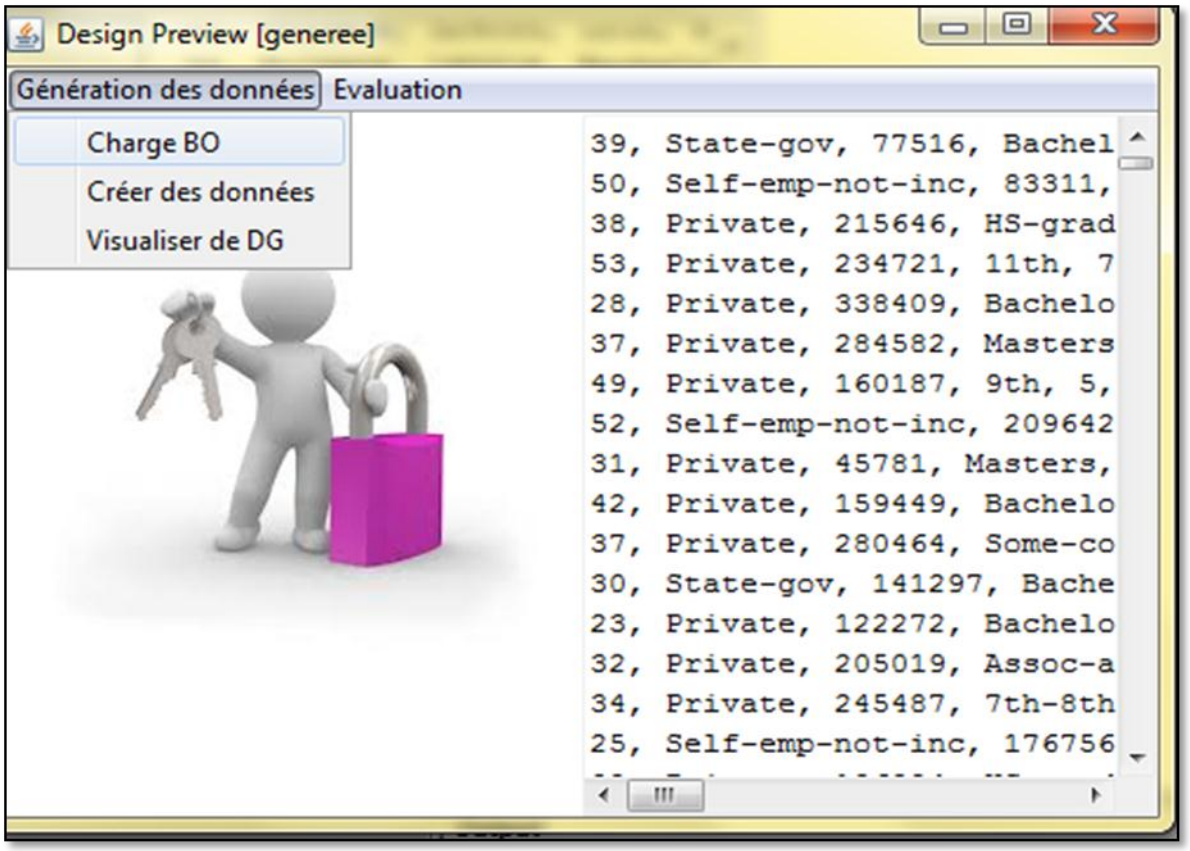

Figure III.3. Interface graphique pour générer les données

- L'interface suivant permet d'afficher les données générée

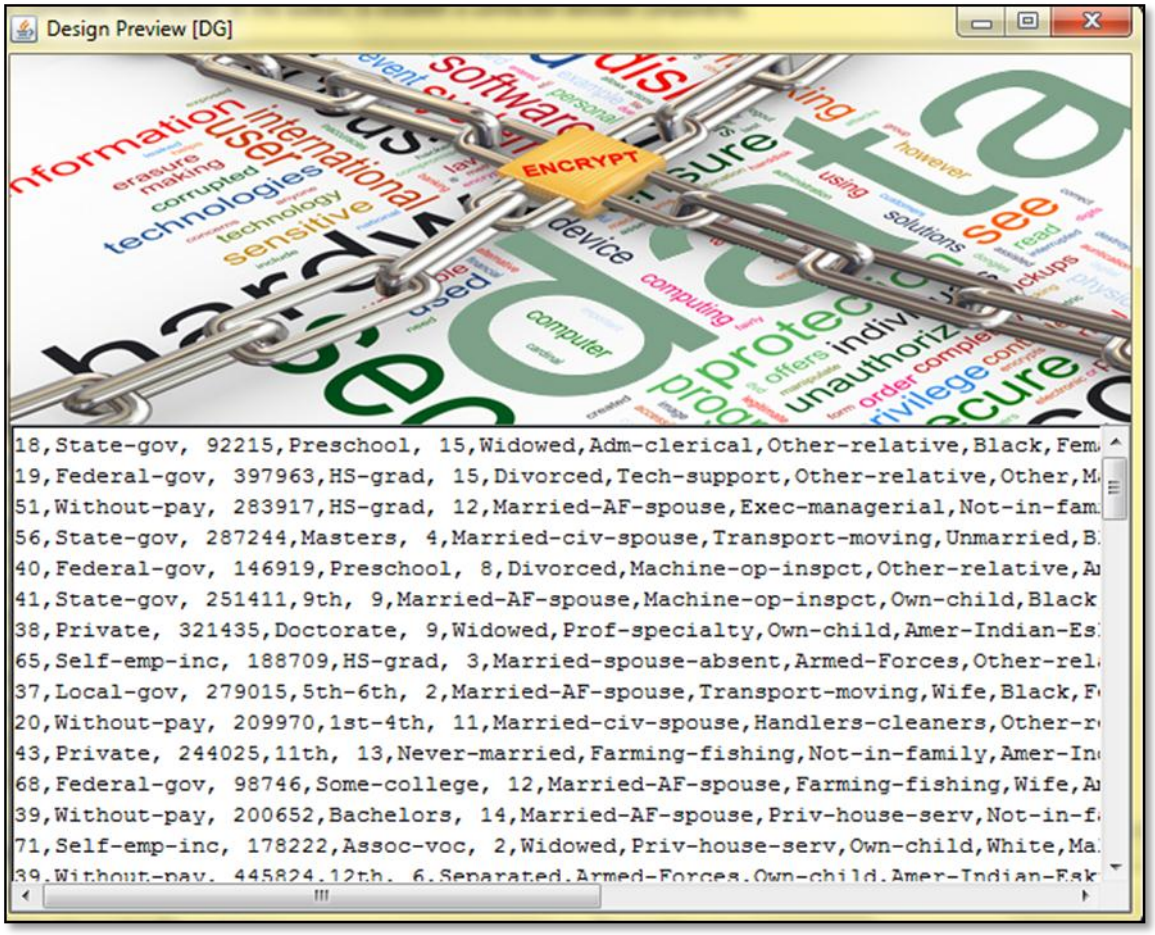

Figure III.4.la visualisation des données générée

IV.4.l'évaluation des résultats

Pour évaluer les résultats nous avons utilisé deux tests dans l'API WEKA.

- La figure suivante représente l'interface pour les tests et visualisation des résultats.

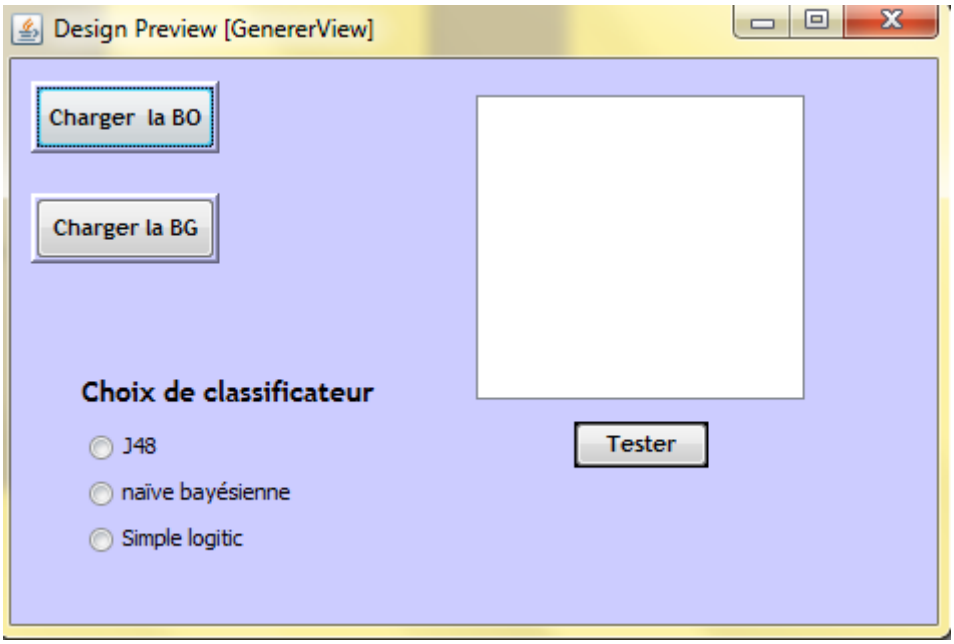

Figure III.4. Interface graphique pour l'affichage des résultats

Nous avons utilisé trois classificateurs « J48, naïve bayésienne et Simple logistic» pour tester les performances de notre approche.

Les tableaux et les diagrammes suivants présents les résultats de cette approche relatifs à chaque classificateur.

	Modèle de TP	**Modèle de TG**
Précision	85.7271 %	80.2898 %
Erreur	14.2729 %	19.7102 %

Table III.1.le résultat de la classification J48

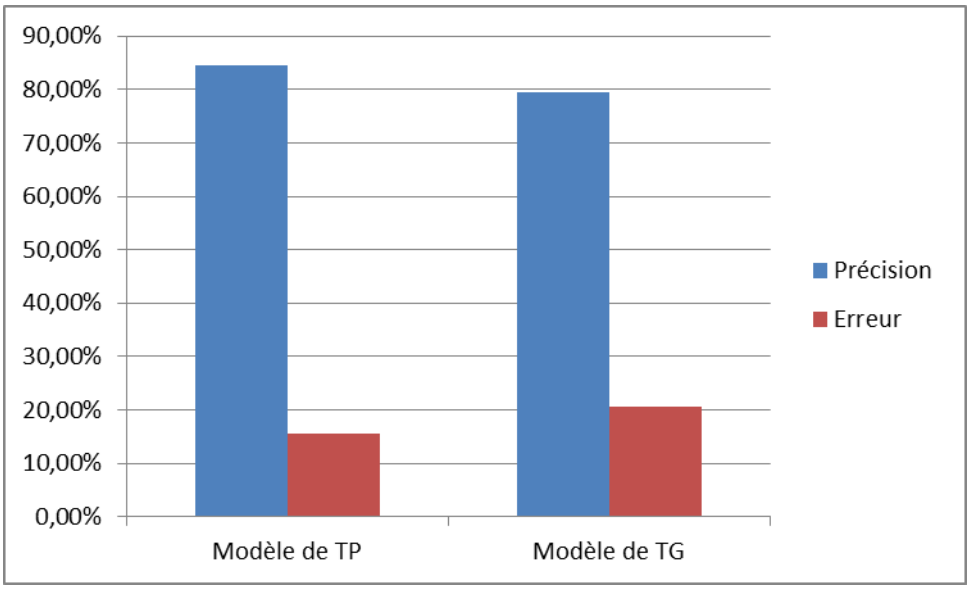

Figure III.5. Le diagramme des résultats de la classification J48

	Modèle de TP	**Modèle de TG**
Précision	82.8758 %	76.2052 %
Erreur	17.1242 %	23.7948 %

Table III.2.le résultat de la classification naïve bayésienne

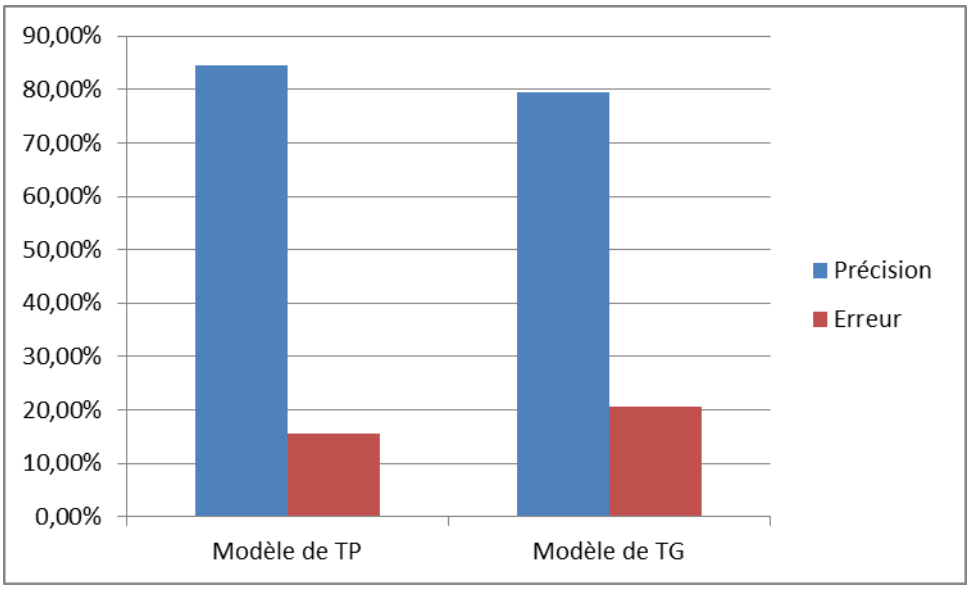

Figure III.6. Le diagramme des résultats de la classification naïve bayésienne

	Modèle de TP	Modèle de TG
Précision	84.5344 %	79.3979 %
Erreur	15.4656 %	20.6021 %

Table III.3.le résultat de la classification Simple logistic

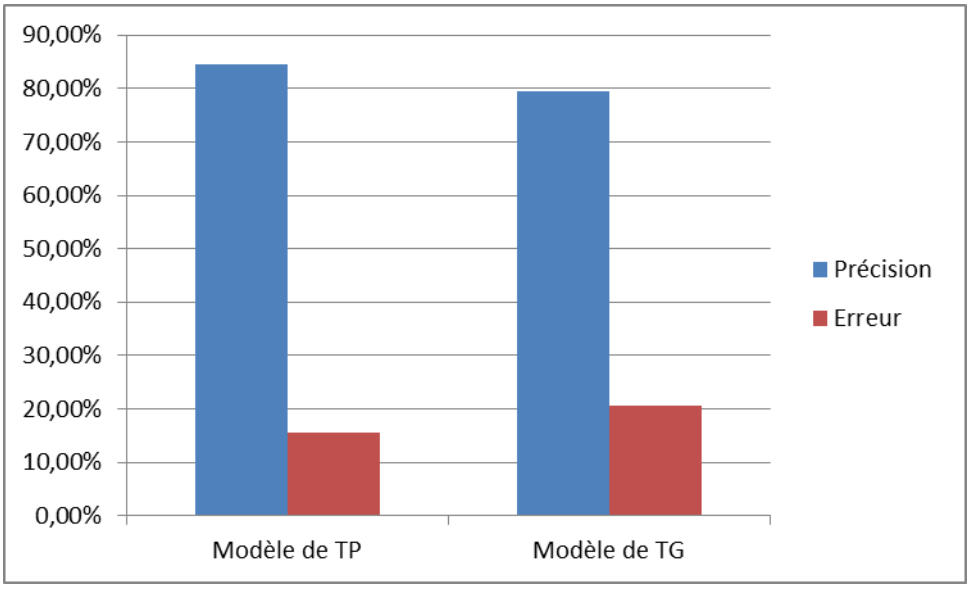

Figure III.7. Le diagramme des résultats de la classification Simple logistic

- **Discussion :**

Les résultats montrent que l'efficacité de notre approche est presque la même pour les trois classificateurs, la différence en terme de précision entre le modèle généré à partir de la table « TP » et le modèle généré à partir de la table « TG » est autour de 6% (La précision dans chaque classificateur est respectivement : «85.7271 % pour j48, 82.8758 % pour naïve bayésienne, 84.5344 %pour Simple logistic » pour la table privée, « 80.2898% pour j48, 76.2052 % pour naïve bayésienne et 79.3979 % pour Simple logistic» pour la table générée), la même chose pour le taux d'erreur avec une différence 6% pour la classificateur (L'erreur dans chaque classificateur est respectivement : «14.2729% pour j48, 17.1242% pour naïve bayésienne,15.4656 %pour Simple logistic » pour la table privée, « 19.7102 % pour j48, 23.7948 % pour naïve bayésienne et 20.6021 %pour Simple logistic» pour la table générée) , d'après ces

résultats on remarque une légère dégradation des performances des classificateurs issu des données générées, a notre avis cette dégradation reste acceptable vue la forte protection de la vie privée des propriétaires des données du fait que les données publies se diffère totalement des données originale .

IV. Conclusion

Dans ce chapitre nous avons présenté la partie pratique de notre travail de ce fait nous avons présenté les détails de notre approche pour la protection des données personnelles avec quelques tests et quelques résultats.

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

Conclusion générale

Conclusion générale

Le travail présenté dans ce mémoire s'inscrit dans le contexte de la protection de la vie privée. Nous avons donné une vue générale sur ce domaine en introduisant la notion de la vie privée, les attaques et les technologies ainsi qu'un état de l'art sur la protection des données personnelles.de ce fait nous avons développé une approche qui générer aléatoirement des nouvelles données à partir des données originales en utilisant un classificateur automatique, pour garder le maximum de corrélation entre les attributs sensibles et les attributs non sensibles ainsi qu'un ensemble de règle pour garder une sémantique acceptable entre les différentes valeurs des attributs générées , Les nouvelles données générées diffèrent totalement des données originales ce qui implique une grande protection.

Comme perspectives nous envisageant de :

- tester notre approche sur des autres bases.
- Améliorer la qualité des données générer en gardant par exemple la même distribution que les données originales.
- La comparaison des performances de notre approche avec d'autres travaux.

REFERENCES BIBLIOGRAPHIQUES

[1] le droit de la vie privée, disponible sur :

http://www.chairelrwilson.ca/cours/drt3805/vieprivee.html, consulté le 15 mars 2013.

[2]La vie privée sur l'internet, disponible sur :
http://en.wikipedia.org/wiki/Internet_privacy, consulté le 17 juin 2013.

[3] Sébastien Gambs, Introduction à la protection de la vie privée, cours, 2012.

[4] Yves Deswarte, Intelligence ambiante et protection de la vie privée,cours, 2004.

[5] les cookies (biscuitsempoisonnées) disponible sur :
http://fr.wikipedia.org/wiki/Cookie_%28informatique%29, consulté le 15 mars 2013.

[6] Le phishing (hameçonnage), disponible sur :

http://www.commentcamarche.net/contents/attaques/phishing.php3, consulté le 19 mars 2013.

[7] La sécurité sur Internet, disponible sur :

http://www.rcmp-grc.gc.ca/qc/pub/cybercrime/cybercrime-fra.htm, consulté le 14 avril 2013.

[8] Respect de la vie privée dès la conception (fr), disponible sur :

http://fr.jurispedia.org/index.php/Respect_de_la_vie_priv%C3%A9e_d%C3%A8s_la_conception_%28fr%29, consulté le 15 avril 2013.

[9] Vol d'identité introduction, disponible sur :

http://www.sse.gov.on.ca/mcs/fr/Pages/Identity_Theft.aspx, consulté le 15 mars 2013,

[10] Privacy by design - The IT Law Wiki, disponible sur:
http://itlaw.wikia.com/wiki/Privacy_by_design, consulté le 09 mars 2013.

[11] Sébastien Gambs, Réseaux de communication anonyme, cours, (2011).

[12] Comprendre et utiliser Tor pour préserver son anonymat, disponible sur :

http://www.scout123.net/comprendre-et-utiliser-tor-pour-preserver-son-anonymat.html, consulté le 17 mars 2013.

[13] IAM,SOO,Fédération d'identités, disponible sur :

http://www.normation.com/fr/services/identity-and-access-management-iam, consulté le 24 mars 2013.

[14] Windows Live ID / Microsoft Passport Network Privacy Supplement , consulté le 24 mars 2013, disponible sur :

http://privacy.microsoft.com/FR-FR/windowsliveid.mspx

[15] Single Sign-On (SSO), disponible sur : http://www.commentcamarche.net/contents/92-single-sign-on-sso, consulté le 24 mars 2013.

[16] 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 2013

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

[18] Yeves Deswarte, Sébastien Gambs. Protection de la vie privée : Principes et technologies, CNRS ; France Université de Toulouse.

[19] Données personnelles définition et explications, disponible sur : http://www.technoscience.net/?onglet=glossaire&definition=11079 , consulté le 14 mars 2013,

[20] A. Abou El Kalam, Y. Deswarte, G. Trouessin, E. Cordonnier, "Une démarche méthodologique

pour l'anonymisation de données personnelles sensibles", Actes du 2ème Symposium sur la Sécurité des Technologies de l'Information et des Communications (SSTIC 2004), Rennes (France), 2-4juin 2004..

[21] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati,K-anonymity Università degli Studi di Milano, 26013 Crema, Italia,2007.

[22] LEMAY, Alain, INFRASTRUCTURE LOGICIELLE VISANT À PROTÉGER LA CONFIDENTIALITÉ DU PATIENT DANS LES IMAGES MÉDICALES UTILISÉES EN RECHERCHE, MÉMOIRE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE, 2009.

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

[24] P. Samarati, L. Sweeney, Protecting Privacy when Disclosing Information : k-Anonymity and Its Enforcement through Generalization and Suppression, Technical Report SRI-CSL-98-04,

Computer Science Laboratory, SRI International, 1998.

[25] l-diversité, disponible sur : http://consultation.demotis.org/glossaire/l-diversit%C3%A9 , consulté le 25 mai 2013.

[26] 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

[27] k-anonymat, disponible sur :

http://consultation.demotis.org/glossaire/k-anonymat, consulté le 25 mai 2013,

[28]Ninghui Li, Tiancheng Li, t-Closeness: Privacy Beyond k-Anonymity and _-Diversity, Department of Computer Science, Purdue University

[29] Mehmet Ercan Nergiz, Suleyman Cetintas, Ahmet Erhan Nergiz, Ferit Akova, Generalizations with Probability Distributions for Data Anonymization, Purdue University Purdue e-Pubs,2010

[30] 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.

[31] Cynthia Dwork, Frank McSherry, Kobbi Nissim_ and Adam Smith,Differential Privacy: A Primer for the Perplexed, CONFERENCE OF EUROPEAN STATISTICIANS,2011.

[32] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, Microdata Protection, © Springer US, Advances in Information Security (2007) http://www.springerlink.com/content/q72ph05204n65128/fulltext.pdf

[33] java, disponible sur :http://fr.wikipedia.org/wiki/Java_%28langage%29, consulté le 17 juin 2013.

[34] netbeans, disponible sur : http://netbeans.developpez.com/faq/?page=Introduction, consulté le 17 juin 2013.

[35] l'API weka , disponible sur : http://weka.wikispaces.com/Primer, consulté le 17 juin 2013.

[36] Comparaison de méthodes de classifications disponible sur : https://www.lri.fr/~antoine/Courses/Master-ISI/ISI-10/Projets_2012/Projet_DM.pdf, consulté le 17 juin 2013.

[37] UCI Machine Learning Repository Adult Data Set,http://archive.ics.uci.edu/ml, consulté le 17 juin 2013.

Liste des tables

Liste des figures