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 : *Modèle Intelligent et Décision (M.I.D)*

Thème

Implémentation De La Méthode K-anonymat à La Base D'un Algorithme Génétique

Réalisé par :

- **BOUROUAHA Imene**
- **BOURSALI Rahma**

Présenté le 2 juillet 2022 devant le jury composé de MM.

- *ABDERRAHIM Alaeddine* *(Président)*
- *BELABED Amine* *(Encadrant)*
- *SMAHI Mohamed Ismail* *(Examinateur)*

Année Universitaire : 2021-2022

Résumé :

La méthode k-anonymisation consiste à partager des microdonnées avec des tiers (chercheurs, professionnels) pour améliorer les services en minimisant le risque de ré-identification, tout en préservant l'utilité de ces données. Dans ce travail, nous étudions le modèle k-anonymat en le définissant et en expliquant les techniques et les métriques utilisées pour le mettre en œuvre. Ainsi, nous avons réalisé une implémentation à base d'un algorithme génétique, qui a pour objectif de chercher la méthode K-anonymat la plus optimale.

Mots clés : la vie privée, le k-anonymat, l'utilité, l'algorithme génétique.

Abstract :

The k-anonymization allows data to beshared with third parties (researchers and professionals) to improve services sothat data confidentiality is protected and to minimize the risk of re-identification, while maintaining the utility of this data. In ourwork, wepresent an overview of privacy, itsdefinition, and models for protectingit. Wealsotouched on the anonymity model, by definingit and explaining the techniques and measuresused to implementit, and finally we showed the different steps and tools to implement this approach depending on genetic algorithms.

Keywords : privacy, k-anonymity, utility, genetic algorithm.

ملخص:

تسمح طريقة إخفاء الهوية بمشاركة البيانات مع أطراف ثالثة (باحثون ومهنيون) لتحسين الخدمات بحيث تكون سرية البيانات محمية وللتقليل إلى أدنى حد من مخاطر إعادة تحديد الهوية، مع الحفاظ على فائدة هذه البيانات.
نقدم في عملنا هذا نظرة عامة حول الخصوصية وتعريفها ونماذج لحمايتها كما تطرقنا إلى نموذج إخفاء الهوية وذلك من خلال تعريفها وشرح التقنيات والمقاييس المستخدمة لتنفيذها وأخيرا أظهرنا الخطوات والأدوات المختلفة لتنفيذ هذا النهج اعتمادا على الخوارزميات الجينة.

الكلمات المفتاحية: الخصوصية، إخفاء الهوية، المنفعة، الخوارزمية الجينية.

REMERCIEMENTS

Nous remercions en premier lieu, Dieu le tout puissant, de nous avoir donnés la santé, la volonté et le courage pour suivre le chemin de la science et pour mener à terme, ce mémoire.

En second lieu, nous tenons à remercier notre encadrant **M. BELABED Amine** pour la qualité de son encadrement, pour sa patience et sa rigueur durant la préparation de ce mémoire.

Nous adressons également nos vifs remerciements à :

- **M.ABDERRAHIM Alaeddine** d'avoir accepté de présider le jury.
- **M. SMAHI Mohamed Ismail** pour avoir accepté d'examiner ce travail.
- **Nos enseignants du département d'informatique** qui ont contribué à notre formation au cours de notre parcours universitaire.

Dédicace

En témoignage je t'offre ce modeste travail pour remercier pour tes sacrifices et pour l'affection son tu m'as toujours :

*Ma mère « **DJAMILA** »,*

Tu m'as donnée la vie, la tendresse et le courage pour réussir. Tout ce que je peux t'offrir ne pourra rien exprimer l'amour et la reconnaissance que je te porte. Que Dieu te préserve et te procure santé et longue vie.

*Mon père « **ABDELSALLEM** »,*

L'épaule solide, l'œil attentif compréhensif et la personne la plus digne de mon estime et de mon respect et mon amour. Aucune dédicace ne saurait exprimer mes sentiments, que Dieu te préserve et te procure santé et longue vie.

*À ma fille : « **MERIEM CHAHINNEZ** »,*

Mon petit rayon de soleil qui éclaire ma vie, la chanceux d'avoir une fille comme toi, merci de m'avoir donné la parole « ma mère ». Je demande à Dieu le succès dans votre cursus scientifique.

*A mes chers frères : « **Hicham**, **Abdelfateh**, **Tarek** et son fils **Wassim** ».*
*A ma chère sœur :« **Sara** ».*

*A tous les membres de ma famille et toutes personnes qui porte le nom **BOUROUAHA** et **BENSEDDIK**.*

*A ma chère SOEUR avant d'être mon binôme **Boursali Rahma**.*

Bourouaha Imene

Dédicace

Ce n'est que grâce à Allah que nous avons pu achever ce travail.

Je dédie ce PFE :

*A ma grande mère "**Fadila**",*

Rappelé à Dieu le 15 avril 2017.

Que Dieu l'Accueille dans son vaste paradis.

*A mon père "**Abdeslem**" et ma mère "**Aicha Tidda**", que m'ont toujours encouragée pour évoluer dans mes études. Que Dieu vous préserve et vous garde en bonne santé.*

*A ma chère sœur "**Dounia**" et mes chers frères "**Yacine, Ayoub** et **Djoud**".*

Ainsi qu'à chaque membre de ma famille.

*À ma binôme "**Imene**".*

À tous mes amies.

BOURSALI Rahma

Table Des Matières :

Listes Des tableaux :

Listes Des Figures :

Liste Des Abréviations :

AEMO *:* *Algorithmes Evolutionnaires Multiobjectif*

ANS*:* *Attributs Non Sensibles*

AS *:* *Attributs Sensibles*

CAVG*:* *Normalized Average Equivalence Class Size Metric, Voir*

CM *:* *Classification Metric*

CNIL*:* *Commission Nationale De L'informatique Et Des Libertés*

CPU: *Central Processing Unit*

DM *:* *Discernibility Metric*

IDE*:* *Integrated Development Environment*

IE *:* *Identifiants Explicites*

NAS*:* *Numéro D'assurance Sociale*

PREC *:* *Précision*

QID *:* *Quasi-Identifiants*

RAM: *Random Access Memory*

Introduction Générale

La technologique a évolué de façon spectaculaire dans notre époque actuelle, cela est dû principalement à l'avènement de l'Internet. Cette technologie qui représente un énorme système de communication et de transmission de données, a conduit à augmenter de façon exponentielle la quantité des données produites et transmise de jour en jour.

De nombreuses applications (ex. des réseaux sociaux, les sites Web, les applications de soins, de santé, etc.) recueillent des renseignements auprès des utilisateurs d'Internet, qu'ils y consentent ou non, et exploitent des données personnelles pour certains services qui sont confidentielles et doivent être protégées. Tout cela, a soulevé des préoccupations mondiales au sujet de la vie privée des personnes.

En effet, les données jouent un rôle très important dans le développement de la science et de l'innovation. Cela a conduit les chercheurs à proposer des modèles de protection de la vie privée. Ces modèles visent à rendre l'identité des propriétaires anonyme pour éviter le risque de divulgation d'un coté, et conserver l'utilité des données d'un autre coté. Une des modèles permettant de répondre à ces conditions, est le modèle k-anonymat, l'objet d'étude dans notre mémoire.

Une table de données publiée est dite K-anonyme si chaque combinaison de valeurs d'enregistrement ne puisse pas être distingué d'au moins K-1 enregistrements dans cette table**[1]**. Cette méthode est réalisée par plusieurs techniques parmi elle la généralisation et la suppression. Une table satisfaite une k-anonymat optimale si la confidentialité des propriétaires de ces données est forte, tout en préservant l'utilité de ces données. La recherche de telle K-anonymat s'est avérée un problème NP-hard **[2].**

Dans cette étude, nous avons essayé d'implémenter l'approche k-anonymat à la base d'un algorithme génétique. Notre objectif est de déterminer la meilleure généralisation qui maximise à la fois, la confidentialité et l'utilité des données.

Afin d'aborder tous les aspects de notre thème " Implémentation de la méthode K-anonymisation à la base d'un algorithme évolutionniste", notre mémoire est structuré en trois chapitres principaux décrits comme suit :

Le Chapitre 1, intitulé " la vie privée " est consacrée à présenter une vue générale sur la vie privée, sa définition, ses principes ainsi que quelques attaques sur la vie privée avec ces différentes techniques utilisées, et enfin, les modèles qui permettent de protéger la vie privée.

Le Chapitre 2 intitulé "Le modèle k-anonymat : notions de base" introduit les connaissances de base de la méthode k-anonymat, en expliquant ses notions préliminaires, sa définition et les techniques utilisées pour mettre en œuvre cette approche, Ensuite, quelques métriques pour évaluer le modèle k-anonymat ainsi que quelques travaux connexes au k-anonymisation sont présentés.

Le Chapitre 3 intitulé "contribution et implémentation " d'écrit les détails de notre implémentation. Pour cela, nous avons donné dans ce chapitre, un aperçu général sue les algorithmes génétiques et comment adapter ce type d'algorithme au problème K-anonymat. Le chapitre présente également les outils de développement ainsi que les données utilisées dans les expérimentations, suivi d'une interprétation et discussion des résultats.

Enfin, une conclusion générale qui conclut notre travail et donne quelques perspectives.

Chapitre I : La Vie Privée sur l'internet

I.1. Introduction

L'apparition de l'internet et le développement des technologies de l'informatique à causer des plusieurs problèmes de la protection de la vie privée et de sécurité. Bien que la protection de la vie privée est hautement considérée pour assurer la protection des utilisateurs et des renseignements personnels (d'individus ou de groupes), le potentiel de nombreuses violations de la vie privée sur Internet est préoccupant.

L'objectif de ce chapitre est de présenter une vue générale sur la vie privée, nous commençons par la définition de la vie privée ainsi que ses principes. Ensuite, nous présentons quelques attaques sur la vie privée avec quelques techniques d'attaques. Enfin nous terminons par les modèles de protection de la vie privée.

I.2. Définition

La vie privée (en anglais "privacy") est un terme flou et difficile à définir car, chaque individu a le droit de décider et déterminer librement les limites de sa vie privée qui doit être respecter **[3]**par les autres personnes.

Le concept de la vie privée peut contenir toutes les informations destinées à être individuel à une personne, telle que la protection d'identité, d'image, le secret de la résidence, du domicile, le secret relatif à la santé, le secret professionnel, la vie familiale et la vie sentimentale…

Dans l'article 12 de la déclaration universelle des droits de l'homme de 1948, il est mentionné que chaque personne a le droit de protéger sa vie privée « Nul ne sera l'objet d'immixtions arbitraires dans sa vie privée, sa famille, son domicile ou sa correspondance, ni d'atteintes à son honneur et à sa réputation. Toute personne a droit à la protection de la loi contre de telles immixtions ou de telles atteintes ».**[4]**

I.3. La vie privée sur l'internet

Sur internet, la vie privée est devenue plus menacée, on ne compte plus les cas de cambriolages survenus à la suite d'une publication sur des réseaux sociaux annonçant

des lieux de vacances, l'usurpation d'identité ou le licenciement pour cause de propos injurieux sur Facebook sont aussi des pratiques récurrentes.

En réalité, deux types de données personnelles sont disponibles sur le net : celles que nous communiquons volontairement lors d'inscriptions sur des sites d'échanges et celles collectées à notre insu.**[5]**

I.4. Pourquoi la protection de la vie privée ?

Imaginer un monde sans intimité sur internet et les réseaux sociaux, où toutes vos informations personnelles qui permettent de t'identifier directement ou indirectement sont partagées au grand public. C'est un grand problème parce que sans intimité l'internet devient un endroit moins sûr pour nous et dangereux.

Actuellement, la protection de la vie privée sur internet devient nécessaire et très importante et certaines informations devraient être privées et protéger pour la sécurité de notre vie.

I.5. Les Principes de la vie Privée

La protection de la vie privée s'évalue avec le développement dans le temps, en particulier avec l'avancement de la technologie. Cette protection exige des principes fondamentaux qui doivent être compatibles avec la protection de la vie privée, parmi les principaux principes on cite :

I.5.1. La Minimisation des données personnelles :

Les données à caractère personnel doivent être convenables, relatives au contenu traitée et limitées à ce qui est nécessaire **[6]**, Comme l'explique la CNIL, « *les informations enregistrées doivent être pertinentes et strictement nécessaires au regard de la finalité du fichier* ».

Par exemple, un responsable de traitement peut avoir besoin des coordonnées de ses prospects pour ses opérations de communication, mais s'il ne procède qu'à des

newsletters par voie électronique, il n'est pas nécessaire a priori de collecter l'adresse postale des personnes concernées. **[7]**

I.5.2. Le Consentement explicite :

Le consentement doit être préalable à la collecte ou traitement des données avec l'autorisation explicite de la part du propriétaire.**[8]**

I.5.3. La Souveraineté des données :

Ce principe donne qu'un individu doit pouvoir pratiquer sa souveraineté sur ses informations et de les garder sous son contrôle direct autant que possible. C'est aussi le droit de les corriger et éventuellement les supprimer.**[6]**

I.5.4. La Transparence :

Une organisation doit faire en sorte que des renseignements précis sur ses politiques et ses pratiques concernant la gestion des renseignements personnels soient facilement accessibles à toute personne. **[9]**

I.5.5. La sécurité :

L'entreprise doit implémenter des mécanismes et utiliser des mesures de sécurité pour protéger les données des utilisateurs contre tout accès illégal, divulgation, reproduction, modification ou toute utilisation inappropriée. **[9]**

I.5.6. L'anonymat :

L'utilisateur n'est pas tenu de publier ses informations personnelles lorsqu'il fait une action ou invoque une opération. Dans un contexte informatique générale, cela signifie de garder le nom et l'identité d'un utilisateur caché à travers diverses applications. **[10]**

I.5.7. La Pseudonymat :

Le système permet à ses utilisateurs d'utiliser un pseudonyme au lieu de leur véritable identité.**[6]**

I.5.8. La Non-chaînabilité :

C'est l'impossibilité (pour d'autres utilisateurs) d'établir un lien entre différentes opérations faites par un même utilisateur.**[6]**

I.5.9. La Non-observabilité :

Ce principe permet aux utilisateurs d'effectuer une action ou d'utiliser une ressource système sans que d'autres puissent observer que la ressource est utilisée.**[6]**

I.6. **Les Menaces sur la vie Privée**

Les menaces les plus courantes pour les données personnelles, sont l'utilisation de mots de passe facile à pirater, le partage de mots de passe et la non-installation de mises à jour de sécurité sur les appareils, mais la liste des risques pour la confidentialité des données ne se limite pas à cela, le paragraphe suivant présente d'autres exemples des menaces à la vie privée :

I.6.1. Divulgation des données personnelles :

Bien que les individus comprennent la valeur de leurs données personnelles, cependant ils partagent largement la facilité de trouver et accéder à leurs données personnelles (le nom, la date de naissance, l'état civil …), que la majorité d'entre nous considèrent triviales et inoffensives notamment, mais qui peuvent être utilisés contre nous par exemple, se venger, punir ou dénoncer des gestes allégués surtout avec l'émergence des réseaux sociaux et divers services.**[6]**

I.6.2. Vol d'identité :

Le vol d'identité (ou usurpation d'identité) est un crime dans lequel des informations personnelles sont volées et usurpées dans le but d'accéder à certaines ressources ou d'obtenir des avantages sous le nom d'une autre personne.

Le piratage peut aller au-delà de la confidentialité personnelle et identitaire avancée, comme le numéro d'assurance sociale (NAS), les cartes de téléphone, les mots passe de boit mail, les certificats de naissance et les passeports. **[11]**

I.6.3. Le profilage :

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. **[12]**

I.7. Les techniques d'attaques

Parmi les techniques les plus utilisées dans les attaques sur la vie privée, on cite :

I.7.1. Logiciels espions (spywares) :

Un logiciel espion est un programme informatique qui accumule, dans le but de les transmettre à des tiers, des informations sur l'utilisateur d'un ordinateur, et ce, sans avoir obtenu au préalable son véritable consentement. Il s'agit d'un type particulier dans la famille des logiciels malveillants **[13]**. Cette pratique porte atteinte aux données personnelles et à la vie privée des internautes dans la mesure où « un logiciel espion est généralement employé pour effectuer du cybermarketing « agressif » et – dans de plus

rares cas – prendre possession de renseignements personnels pour réaliser des activités illicites ». **[14]**

À certains égards, l'utilisation de logiciels espions a pour principal objectif l'aspiration d'adresses électroniques afin d'adresser aux internautes des courriels non-sollicités, c'est-à-dire des « spams ». Dans un arrêt du 14 mars 2006, la Chambre criminelle de la Cour de cassation française a considéré que l'utilisation de « logiciels permettant d'aspirer sur des espaces accessibles au public d'internet (forums de discussion, sites…) des adresses électroniques de personnes physiques afin de leur adresser des courriels publicitaires non sollicités (ou spams) » **[16] [15]** était illégale.

I.7.2. Les cookies :

Un cookie est un petit fichier stocké par un serveur dans le terminal (ordinateur, téléphone, etc.) d'un utilisateur et associé à un domaine web (c'est à dire dans la majorité des cas à l'ensemble des pages d'un même site web). Ce fichier est automatiquement renvoyé lors de contacts ultérieurs avec le même domaine. (CNIL)

Les cookies furent inventés au milieu des années 1990 par les Américains John Giannandrea et Lou Montullise présentent sous la forme de fichiers textes, qui sont automatiquement enregistrés par le navigateur sur le disque dur lorsqu'un visiteur se rend sur un site web. Ils ont été développés pour améliorer l'expérience de l'utilisateur, pour permettre aux sites web de se souvenir du passage de telle personne en plus proposer du contenu adapté à votre culture locale **[17]**mais permet aussi de les utiliser pour le Traking à travers l'enregistrement et l'analyse du comportement des utilisateurs.

I.7.3. Le 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**. [18]**

I.7.4. Attaques par inférence :

Une attaque par inférence est une technique d'exploration de données utilisée pour accéder illégalement à des informations sur un sujet ou une base de données en analysant des données. Il s'agit d'un exemple de violation de la sécurité de l'information. Une telle attaque se produit lorsqu'un utilisateur est capable de déduire des informations clés ou critiques d'une base de données à partir d'informations triviales sans y accéder directement. **[19]**

I.8. Quelques modèles de protection de la vie privée

Dans les années récentes, de nombreuses approches ont été proposées dans la littérature de la préservation de la confidentialité des données publiées. L'objectif est de publier des données anonymisées avec la préservation de l'identité de l'utilisateur contre la divulgation et sans violations de la vie privée. Il existe plusieurs modèles telle que : k-anonymat, l-diversité, t-proximité, δ-Présence et la différentiel Privacy.

I.8.1. K-anonymat :

Le modèle de k-anonymat est le premier modèle proposé dans la littérature (Samarati et Sweeney 1998), est une technique d'anonymisation qui utilise les opérations de généralisation et de suppression. **[20]**

Ce modèle de protection de la vie privée bien connu vise à protéger les ensembles de données contre la ré-identification dans le modèle du procureur. Un jeu de données est K-anonyme si chaque enregistrement ne peut pas être distingué d'au moins k-1 autres enregistrements concernant les quasi-identificateurs. Chaque groupe d'enregistrements indiscernables forme une classe dite d'équivalence. Le k-anonymat et les techniques d'anonymisation seront présentées en détail dans le chapitre 2. **[21]**

I.8.2. l-diversité :

Le k -anonymat protège une table contre la divulgation de l'identité de ses individus, mais ne protège pas suffisamment les attributs sensibles qui ne font pas partie du quasi- identifiant. Mettent en évidence cette faiblesse dans le modèle du k -anonymat, avec deux types d'attaques : l'attaque par homogénéité et l'attaque avec connaissance a priori Pour remédier à ces limitations du k-anonymat, Machaanavajjhala et al. ont introduit la l-diversité comme un formalisme de protection plus fort. **[22]**

Example :

Le Tableau I-1 représente une version anonymisée satisfaisant distincte et entropique 3- diversité, il compose de deux quasi-identifiants (Code ZIP et Age) et deux attributs sensibles (Salaire et Maladie) :

Tableau I-1: Table Original **[23]**

	ZIP Code	**Age**	**Salary**	**Disease**
1	476**	2*	3k	Heart Disease
2	476**	2*	4k	Heart Disease
3	476**	2*	5k	Heart Disease
4	4790*	≥ 40	6k	Flu
5	4790*	≥ 40	11k	Heart Disease
6	4790*	≥ 40	8k	Cancer
7	476**	3*	7k	Heart Disease
8	476**	3*	9k	Cancer
9	476**	3*	10k	Cancer

Tableau I-2: Une version 3-diversitiée **[23]**

	ZIP Code	**Age**	**Salary**	**Disease**
1	476**	2*	3k	Heart Disease
2	476**	2*	4k	Heart Disease
3	476**	2*	5k	Heart Disease
4	4790*	≥ 40	6k	Flu
5	4790*	≥ 40	11k	Heart Disease
6	4790*	≥ 40	8k	Cancer
7	476**	3*	7k	Heart Disease
8	476**	3*	9k	Cancer
9	476**	3*	10k	Cancer

I.8.3. T-proximité :

Bien que le formalisme de protection de la l-diversité représente une avancée au-delà du k-anonymat en protégeant les tables contre les attaques de divulgation d'attribut, il a cependant quelques faiblesses.

Le modèle de t-proximité a été proposé par Li Ninghui et al. **[23]**.Ce modèle utilise la propriété selon laquelle la distance entre la distribution de l'attribut sensible dans un groupe anonymisé ne devrait pas être différente de la distribution dans la table entière de plus d'un seuil "t".

I.8.4. δ-Présence :

C'est la probabilité de déduire la présence de l'enregistrement de toute victime potentielle dans un intervalle spécifié δ = (δmin, δmax).

Formellement, étant donné un tableau E publique et une table T privée, où T ⊆ E, une table généralisées T* satisfait (δmin, δmax) -présence si :

$$\delta min \leq P\ (t \in T \mid T^*) \leq \delta max \text{ pour tout } t \in E$$

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 à accès à la même table E comme l'attaquant fait. Cela peut ne pas être une hypothèse pratique. **[24]**

Exemple :

Le tableau 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 **(**Tableau I-2**)** 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 **(**Tableau I-3**)**, mais précise quels tuples sont dans l'ensemble de données privées **(**Tableau I-2**)**.

Tableau I-3: Tableau publique(E) **[25]**

	Nom	**Zip**	**Age**	**Nationalité**	**Sen**
A	Alice	47906	35	USA	0
B	Bob	47903	59	Canada	1
C	Chirstine	47906	42	USA	1
D	Drink	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

Tableau I-4: Tableau privée(T) **[25]**

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

E* est la généralisation de E **(**Tableau I-5**)** et T* est (0.5, 0. 66) - présence généralisation de T **(**Tableau I-6**)** Les deux généralisations ont la même cartographie de généralisation.

Tableau I-5: Tableau publique (E*) **[25]**

	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*	*	Blugaria	1
I	48*	*	France	1

Tableau I-6: Tableau privée (T*) **[25]**

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

I.8.5. Différentiel-privacy :

Dans ce modèle, la protection de la vie privée n'est pas considérée comme une propriété d'un jeu de données, mais comme une propriété d'une méthode de traitement des données. De manière informelle, il garantit que la probabilité d'une sortie possible du processus d'anonymisation ne change pas « de beaucoup » si les données d'une personne sont ajoutées ou supprimées des données d'entrée. Par conséquent, il devient très difficile pour les attaquants d'obtenir des informations sur des individus spécifiques et les ensembles de données sont protégés contre la divulgation de l'appartenance, de l'identité et des attributs. La confidentialité différentielle ne fait pas non plus d'hypothèses solides sur les connaissances de base des attaquants, par exemple sur les attributs qui pourraient être utilisés pour établir des liens. Au lieu de cela, tous les attributs devraient être définis comme étant quasi-identificateurs.**[6]**

I.9. Conclusion

Dans ce chapitre nous avons présenté une vue générale de la vie privée sur internet, nous avons introduit par une définition avec les différents principes de la vie privée, ensuite nous avons parlé de quelques attaques et des techniques d'attaque et nous avons terminé par les modèles de la protection de la vie privée.

Chapitre II : Le Modèle k-anonymat

II.1. Introduction

Comme nous avons mentionné dans le chapitre précédant, le k-anonymat est parmi les modèles les plus utilisés dans le domaine de la protection de la vie privée.

Ce modèle traite la problématique qui consiste à partager des micro-données avec des tiers (ex. chercheurs, professionnels, etc.) pour fournir les différents services et applications, de sorte que la confidentialité des propriétaires de ces données soit préservée. La protection de vie privée dans ce modèle consiste à rendre difficile d'établir d'un lien entre l'identité d'un individu particulier et l'une de ces données (ou attributs) privées et sensibles.

Ce chapitre introduit les connaissances de base nécessaires à la compréhension de l'anonymisation en générale et le modèle K-anonymat en particulier.

II.2. Notions préliminaires

Généralement, les données sont collectées et publiées sur de nombreux formats différents. On peut distinguer deux grands formats telle que : les microdonnées (microdata) et les macrodonnées (macrodata), les macrodonnées sont des données qui ont été agrégées **[26]**,tandis que le terme " microdonnées" signifie un enregistrement contenant des informations concernent un individu spécifique (un citoyen ou une entreprise). **[27]**

Dans notre chapitre nous focalisons plus sur le format " microdonnées " qui peuvent être représentées sous forme de tableau (matrice), ou chaque enregistrement (ligne) comporte l'information sur un individu (physique ou morale) spécifique et chaque colonne est un attribut partagé par tous les individus du tableau.**[26] [27]**le TableauII-1 représente une microdonnées avec 6 enregistrements et 5 attributs ("Patient Name, "Gender", "Age", "Zip code" et "Health Problem").

Tableau II-1: Un exemple de microdata

(IE)	(QID)			(AS)
Patient Name	**Gender**	**Age**	**Zip code**	**Health Problem**
Amit	Male	35	400071	Viral lnfection
Pankaj	Male	37	400182	Viral lnfection
Vishal	Male	39	400095	Heart problem
Sheetal	Female	54	440672	Flu
Pallavi	Female	58	440123	Heart problem
Nilesh	Male	54	440893	Viral lnfection

Les attributs dans microdata peuvent être divisés en quatre catégories : **[28] [29]**

- **Les identifiants explicites (IE)** : sont les attributs qui permettent d'identifier un individu. Par exemple : nom, numéro de sécurité sociale, etc.
- **Les quasi-identifiants (QID)** : Ce sont les attributs qui ne peuvent pas identifier un individu directement mais s'ils sont liés à des données publiques, ils peuvent identifier un individu facilement. Par exemple, le code postal, l'âge, le sexe, etc.
- **Les attributs sensibles (AS)** : Ce sont les attributs qu'un individu veut cacher aux autres. Par exemple la maladie.
- **Les attributs non sensibles (ANS)** : attribut n'appartenant à aucune des trois catégories précédentes.

A titre d'exemple le Tableau II-1 représente une table qui se compose de l'attribut "Patient Name" est un identifiant explicite (IE), les attributs "Gender", "Age" et "Zip code" sont des quasi-identifiants (QID) et l'attribut "Health Problem" est un attribut sensible (AS).

D'autre part, les attributs d'une microdonnées peuvent être de différents types : **[30]**

- **Continu :** (ou bien numérique) si l'attribut sur lequel des opérations numériques et arithmétiques peuvent être effectués, exemple : les attributs "Age " et "Zip code" dans le TableauII-1.

- **Catégoriel :**si l'attribut prend des valeurs sur un ensemble fini et sur lequel les opérations arithmétiques ne peuvent pas être effectués, exemple : les attributs "Patient Name","Gender" et "Health Problem" dans la Tableau II-1.

Dans le cas où une relation d'ordre peut être appliquée sur ses valeurs possibles, on dit qu'il est un attribut Ordinal sinon il est un attribut Nominal.

Microdata est le format le plus flexible parmi tous les types de données, cependant c'est le plus problématique pour la vie privée des individus. **[27]**

II.3. K-anonymisation

K-anonymat *(en anglais k-anonymity)* est un concept clé qui a été introduit pour faire face au risque de ré-identification des données anonymisées par la liaison avec d'autres ensembles de données (l'attaque par liaison d'enregistrements). Ce modèle a été proposé la première fois par Pierangela Samarati et Latanya Sweeney en **[31][1][32]**. L'idée derrière le k-anonymat est d'empêcher la divulgation des attributs sensibles d'un individu particulier.

Définition 1 : (k-anonymity) [1]

Une table T est dite K-anonyme si chaque enregistrement $t \in T$, il existe $k - 1$ autres enregistrements qui partagent les mêmes valeurs pour tous les attributs du quasi-identifiant.

Définition 2 : (Classe d'équivalence) [28]

Dans une table T, une classe d'équivalence est un ensemble d'enregistrements qui ont la même valeur pour tous les attributs de quasi-identifiants.

Tableau II-2: microdata original

QID			AS
Zip	**Age**	**Nationalité**	**Etat**
13053	28	Russie	Heart Disease
13068	29	Américaine	Heart Disease
13068	21	Japonais	Infection virale
13053	23	Américaine	Infection virale
14853	50	Indian	Cancer
14853	55	Russie	Heart Disease
14850	47	Américaine	Infection virale
14850	49	Américaine	Infection virale
13053	31	Américaine	Cancer
13053	37	Indian	Cancer
13068	36	Japonais	Cancer
13068	35	Américaine	Cancer

Tableau II-3: microdata de 4-anonymity

QID			AS
Zip	**Age**	**Nationalité**	**Etat**
130**	<30	*	Heart Disease
130**	<30	*	Heart Disease
130**	<30	*	Infection virale
130**	<30	*	Infection virale
1485*	40	*	Cancer
1485*	40	*	Heart Disease
1485*	40	*	Infection virale
1485*	40	*	Infection virale
130**	3*	*	Cancer
130**	3*	*	Cancer
130**	3*	*	Cancer
130**	3*	*	Cancer

Dans le contexte des problèmes de k-anonymat, une base (ou une table) de données est sous format de micro-données. La base se compose de « n » enregistrement et « m » attributs. Généralement, pour rendre une base anonyme, la première étape consiste à supprimer ou chiffrer les IE afin d'éviter toute identification directe. Le Tableau II-2 représente une table originale et le Tableau II-3 est son anonymisation pour k=4. Nous pouvons voir clairement dans le Tableau II-3 que chaque enregistrement est identique à moins 3 autres pour les attributs QID. La probabilité qu'un individu est ré-identifié dans une table anonymisée est égale à 1/k.

En générale, ce modèle est suffisant pour contrer les attaques par liaison d'enregistrements parce qu'il focalise plus sur les QID, mais il est vulnérable à de nombreuses attaques telle que les attaques par liaison d'attributs et plus particulièrement les attaques par homogénéité et les attaques fondées sur des connaissances de base. **[33]**

Pour assurer le k-anonymat, on utilise deux techniques principales : la généralisation et la suppression sur les QID. Nous discuterons ces techniques dans la section suivante.

II.4. Les techniques d'anonymisation

Les bases conformes au k-anonymat peuvent être construites par généralisation et suppression. Nous allons d'écrire dans cette section ces techniques en utilisant un exemple. Donc, nous exploitons le Tableau II-4 qui représente une base de données originale médicale.

Tableau II-4: microdata original

Patient Name	Gender	Age	Zip code	Health Problem
Amit	Male	35	400071	Viral lnfection
Pankaj	Male	37	400182	Viral lnfection
Vishal	Male	39	400095	Heart problem
Sheetal	Female	54	440672	Flu
Pallavi	Female	58	440123	Heart problem
Nilesh	Male	54	440893	Viral lnfection
Sagar	Male	41	400022	Flu
Mahesh	Male	46	400135	Flu
Sujata	Female	44	400182	Flu

Les identifiants explicite (IE) telle que "Patient Name" doivent être supprimés de la Tableau II-4.

Tableau II-5: microdata après la suppression de IE

Gender	Age	Zip code	Health Problem
Male	35	400071	Viral lnfection
Male	37	400182	Viral lnfection
Male	39	400095	Heart problem
Female	54	440672	Flu
Female	58	440123	Heart problem
Male	54	440893	Viral lnfection
Male	41	400022	Flu
Male	46	400135	Flu
Female	44	400182	Flu

II.4.1. Généralisation :

La généralisation est une procède qui consiste à remplacer les quasi-identifiants par des valeurs moins spécifiques, mais cohérentes. **[31]**

Une hiérarchie est souvent utilisée pour généraliser les valeurs **[29].** Comme les Figures suivantes montrent, les valeurs du niveau le plus bas sont les valeurs les plus spécifiques (valeurs d'origine). Ainsi, la racine de la hiérarchie est la valeur la plus générale, représente le plus haut niveau. Exemple : dans la Figure II.1 le nœud "Male" est au niveau 0 de la hiérarchie et Le nœud "4*****" dans la Figure II.2 est au niveau 3 de la hiérarchie.

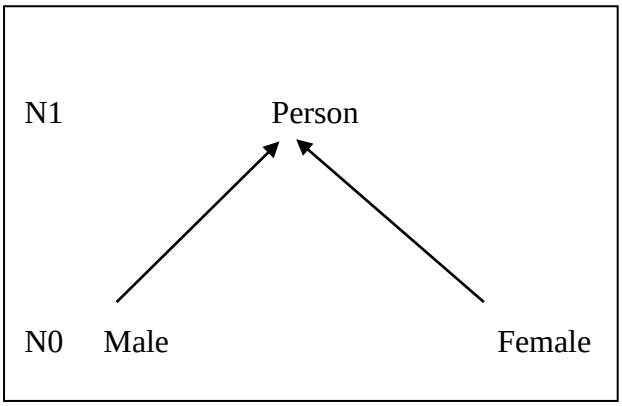

Figure II.1: Hiérarchie de généralisation de l'attribut "Gender".

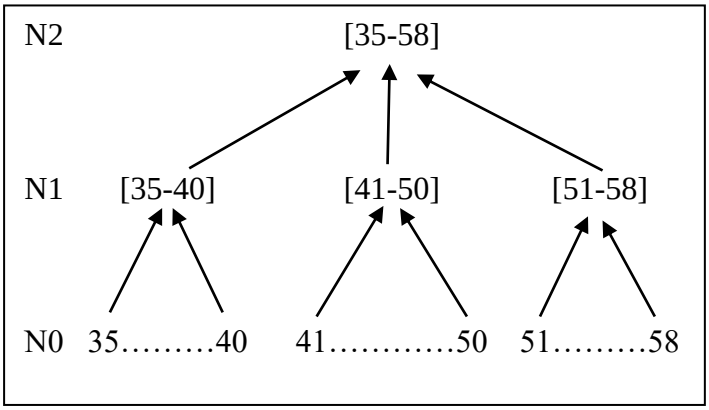

Figure II.2 : Hiérarchie de généralisation de l'attribut "Age".

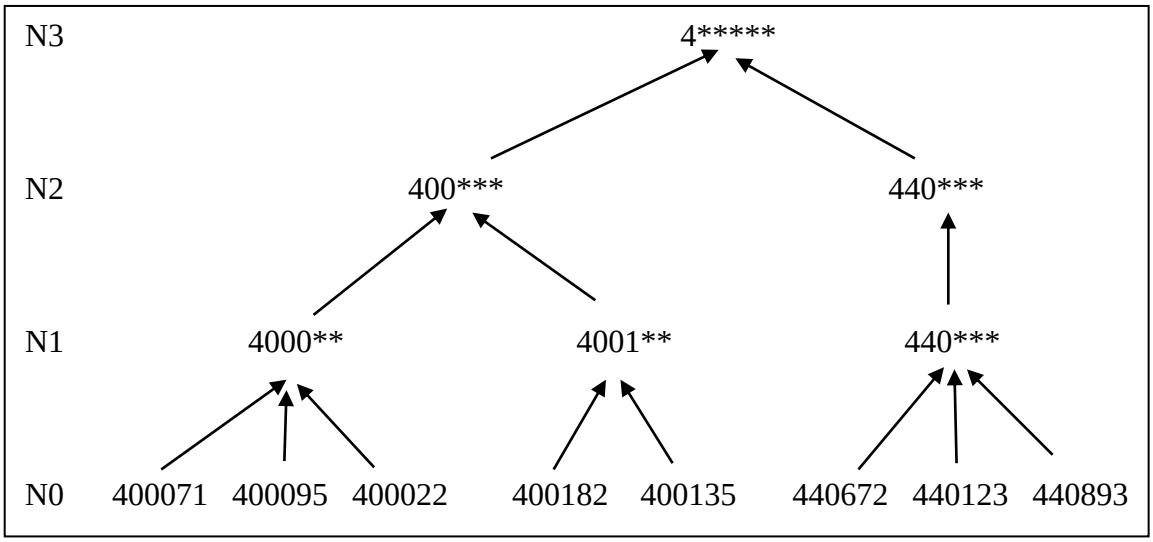

Figure II-1: Hiérarchie de généralisation de l'attribut "Zip code".

Dans le cas où la hiérarchie des QID est définie, on remplace dans la table anonyme, une valeur d'un attribut du QID de la table originale par un de ses ancêtres dans la hiérarchie de généralisation. Le niveau de généralisation appliqué n'est pas nécessairement le même pour chacun des attributs du QID. A titre d'exemple, si l'on considère que "Gender", "Age" et "Zip code" sont des attributs du type QID, nous obtenons, par le procédé de généralisation la Tableau II-6**.**

Tableau II-6: Application de la technique de généralisation aux attributs "Gender", "Age" et "Zip code"

Gender	**Age**	**Zip code**	**Health Problem**
Person	[35-40]	40****	Viral lnfection
Person	[35-40]	40****	Viral lnfection
Person	[35-40]	40****	Heart problem
Person	[41-50]	44****	Flu
Person	[51-58]	44****	Heart problem
Person	[51-58]	44****	Viral lnfection
Person	[41-50]	40****	Flu
Person	[41-50]	40****	Flu
Person	[41-50]	40****	Flu

II.4.2. Suppression :

Lawrence a introduit cette technique en 1980 dans **[11][34]**, qui consiste à retirer une colonne totalement (est nommé suppression globale) ou bien remplacer quelques données de colonne par la valeur nulle (suppression locale) pour satisfaire la contrainte k-anonymat.

A titre d'exemple : nous appliquons une suppression globale sur l'attribut "Zip code" et suppression locale sur l'attribut "Age" dans la Tableau II-7 pour éviter le risque de ré-identification.

Tableau II-7: Application de la technique de suppression (locale et globale)

Gender	Age	Health Problem
Male	35	Viral lnfection
Male	37	Viral lnfection
****	39	Heart problem
Female	54	Flu
Female	58	Heart problem
****	54	Viral lnfection
Male	41	Flu
****	46	Flu
****	44	Flu

II.5. Les métriques d'anonymisations

L'objectif principale d'anonymisation **[35] [36] [37]**consiste à répondre à la fois à la confidentialité des données (la vie privée) et l'utilité des données. La confidentialité des données est la capacité de résister à une attaque de ré-identification. Tandis que l'utilité qui est une exigence principale **[31]** de k-anonymisation est définie comme étant la capacité d'extraire des informations utiles pour améliorer les services. Ainsi, L'évaluation de l'utilité est difficile sans préciser le besoin de l'utilisation de ces ensembles de données. **[38]**

Un point intéressant concerne la relation entre la confidentialité et l'utilité des données : plus les données résistent à la ré-identification (plus le k est grand), plus l'utilité de ces ensembles de données est faible. Donc, pour évaluer et quantifier l'anonymisation, il existe deux métriques principales : Les métriques de la vie privée et les métriques de l'utilité des données. Dans cette section nous présentons certaines mesures d'utilité telle que : La précision PREC **[32]**, la Classification Metric (CM) **[39]** et La métrique de CAVG.**[40]**

II.5.1. La métrique de précision PREC :

Sweeney a proposé cette métrique en 2002 dans **[32]** afin de mesurer la perte d'information à travers la hauteur de la hiérarchie de généralisation. Cette métrique est basée sur le principe de plus que les données sont généralisées, plus la précision est perdue. Cette métrique est représentée par la formule suivante :

$$\mathrm{PREC}(\mathrm{T}') = 1 - \frac{\sum_{j=1}^{|QID|}\sum_{i=1}^{|T|}\frac{h}{|VGHAj|}}{|\mathrm{T}| \ * \ |\mathrm{QID}|}$$

- |T| : est le nombre total d'enregistrements de la table T
- |QID | : est le nombre des attributs du QID
- h représente la hauteur de la hiérarchie de généralisation de l'attribut Aj dans l'enregistrement i après la généralisation
- |VGHAj| est la hauteur totale de la hiérarchie de généralisation de l'attribut Aj.

Exemple : nous considérons le Tableau II-6. Pour cette table |T| = 9 et |QID| = 3.Selon les arbres de généralisations de QID dans Figure II.1, Figure II.2 et Figure II.3, on a : $|VGHA_1|$= 1, $|VGHA_2|$=2 et $|VGHA_3|$= 3, représentent respectivement la hauteur totale de la hiérarchie de généralisation de "Gender", "Age" et "Zip code". Ainsi, h_1=1, h_2=1, h_3=2 représentent respectivement les niveaux de généralisation de "Gender", "Age" et "Zip code" dans le Tableau II-6, donc :

$$\mathrm{PREC}\ (\mathrm{Tableau\ II}-6) = 1 - \frac{\left(9 * \frac{1}{1}\right) + \left(9 * \frac{1}{2}\right) + \left(9 * \frac{2}{3}\right)}{9 * 3} = 0.28$$

II.5.2. La métrique de CAVG :

L'objectif de cette métrique est d'essayer à s'approcher au cas idéal, où le nombre des enregistrements dans chaque classe d'équivalence d'une table anonymisée est égale au seuil de k-anonymat (k), de sorte que minimiser l'effet négatif sur la qualité des données. Dans le cas idéal, cette métrique prendrait la valeur 1. On peut obtenir la

métrique CAVG (normalized average equivalence class size metric) à l'aide de la formule suivante :

$$CAVG(T') = \frac{|\mathrm{T}|}{|\mathrm{E}| * \mathrm{K}}$$

- |T| : la taille de table originale T
- |E| : le nombre total de classe d'équivalence dans la table T '
- K : le seuil de k-anonymat

II.5.3. La métrique de classification :

Cette mesure fait référence au principe d'évaluer l'exactitude d'un processus spécifique sur des données anonymisées. Parmi ces processus est la construction d'un modèle de classification. Pour cela Iyengar a introduit la métrique de classification en 2002 dans **[39]**, qui est définie comme la somme des enregistrements dont les étiquettes de classe sont différentes de celle de la majorité dans sa classe d'équivalence, normé par la taille de table. Elle peut être obtenue à l'aide de la formule suivante :

$$CM = \frac{\sum \text{touslesenregistrementspénalité (enregistremente)}}{\mathrm{N}}$$

$$\text{Pénalité } (enregsitremente) = \begin{cases} 1 \; sieestsupprimé \\ 1 \; siclass(e) \neq majorité(G(e)) \\ 0 \; si \text{ non} \end{cases}$$

II.6. Travaux similaires

On dit qu'une table satisfaite le k-anonymat optimale si elle est k-anonyme et préserve au maximum la qualité des données. La complexité d'une k-anonymisation optimale s'est avérée NP-hard**[2]**, donc plusieurs algorithmes et méthodes ont été proposés pour réaliser le k-anonymat. Certains de ces algorithmes sont basés sur des méthodes complètes qui utilisent des approches de recherche gourmande, d'autres sont basés sur des méthodes heuristiques.

Pierangela Samarati **[31]** a proposé un algorithme complet pour identifier toutes les k généralisations minimales et choisit la k anonymisation optimale parmi elles selon certains critères. L'algorithme est basé sur l'axiome que si un nœud au niveau h, dans la hiérarchie de généralisation de domaine satisfait k anonymat, alors tous les niveaux de hauteur supérieure à h satisfont également k anonymat. Pour exploiter cela, l'algorithme utilise une recherche binaire sur les niveaux de la hiérarchie de généralisation de domaine, c'est-à-dire que pour une hiérarchie généralisée de domaine de hauteur h, la recherche commence à un niveau de h/2 et si ce niveau satisfait le k-anonymat alors il recherche au niveau h/4. Sinon, il cherche au niveau 3h/4 et il continue de la même manière jusqu'à ce qu'il trouve la solution au niveau le plus bas possible. Cette approche suppose que le niveau optimal est celui avec le moins de généralisation et, à l'intérieur de ce niveau, elle prend le nœud qui a la perte d'informations minimale comme solution.

Sweeney**[32]** a utilisé un algorithme complet MinGen (2002) qui examine de manière exhaustive toutes les généralisations potentielles pour identifier la généralisation optimale mesurée dans la DM. Sweeney a reconnu que cette recherche exhaustive n'est pas pratique même pour les ensembles de données de taille modeste, motivant une deuxième famille d'algorithmes de k-anonymisation.

Bayardo et Agrawel **[41]** ont proposé un algorithme optimal. Au lieu d'utiliser l'ensemble de données d'origine et en les généralisant systématiquement en un ensemble anonyme, ils ont commencé à partir de l'ensemble de données le plus général et ont

recherché la spécialisation optimale en utilisant des mesures de coût bien définies. Cet algorithme spécialise systématiquement le jeu de données en un ensemble d'inconnues minimales, et il utilise une stratégie de recherche arborescente qui exploite à la fois l'élagage basé sur les coûts et la réorganisation dynamique des recherches.

Plusieurs algorithmes basés sur des heuristiques ont été proposés. Iyengar **[39]** a utilisé un algorithme génétique qui permet des généralisations plus flexibles des données qui élargissent l'espace pour les transformations potentielles des données susceptibles de conduire à de meilleures solutions. Lunacek et al. **[42]** ont amélioré l'approche d'Iyengar en introduisant un nouvel opérateur de croisement pour la généralisation d'attributs contraints et en éliminant celui qui est inutile. Dewri et al.**[43]** ont formulé l'anonymisation k comme un problème d'optimisation multi-objectifs pour parvenir à un compromis entre le niveau de confidentialité et la perte de données. Ils ont utilisé l'algorithme NSGA-II (AEMO), qui a été démontré comme l'un des algorithmes les plus efficaces et les plus célèbres pour l'optimisation multi objective.

II.7. Conclusion

Nous avons consacré ce chapitre à la présentation de modèle k-anonymat, nous avons commencé par introduire des notions préliminaires et la définition du k-anonymat, puis nous avons présenté les différentes techniques de l'anonymisation telle que la généralisation et la suppression et par la suite les métriques d'anonymisation, en particulier : la métrique de vie privée et les métriques de l'utilité. Enfin, nous avons cité quelques travaux réalisés dans ce domaine.

Chapitre III : contribution et Implémentation

III.1. Introduction

Après avoir décrit l'aspect théorique du modèle k-anonymisation, nous passons à notre objectif qui est l'implémentation de la méthode k-anonymat à la base d'un algorithme évolutionniste.

Dans cette partie, nous allons présenter les détails de notre implémentation. Pour cela, nous allons introduire une vue générale sur les algorithmes génétiques et comment nous avons adapté de tels algorithmes pour le problème de K-anonymat. Nous présenterons aussi les outils de développement (langage, plateforme, IDE, etc.) ainsi que les données utilisées dans les expérimentations. Enfin, nous introduirons les résultats des expérimentations et la discussion des résultats.

III.2. Le principe de notre approche

Soit T (A_1, A_2, …. A_n) est une table privée qui comporte des données originales sur un ensemble de personnes, cette table contient N attributs (A_1, A_2, …. A_n). Il existe deux catégories de ces attributs : les attributs Quasi-identificateurs (QID) et les attributs sensibles (SA).

L'objectif de cette approche est d'utiliser la méthode k-anonymat de sorte à générer une nouvelle table anonyme T 'à partir cette table T, La table T'doit contenir exactement le même nombre d'attribut que la table T. Nous utilisons un algorithme évolutionniste afin de sélectionner la meilleure généralisation qui maximise la confidentialité des propriétaires de ces données et au même temps maximise l'utilité des données. La Figure III-1 les étapes en suivant pour réaliser cette approche.

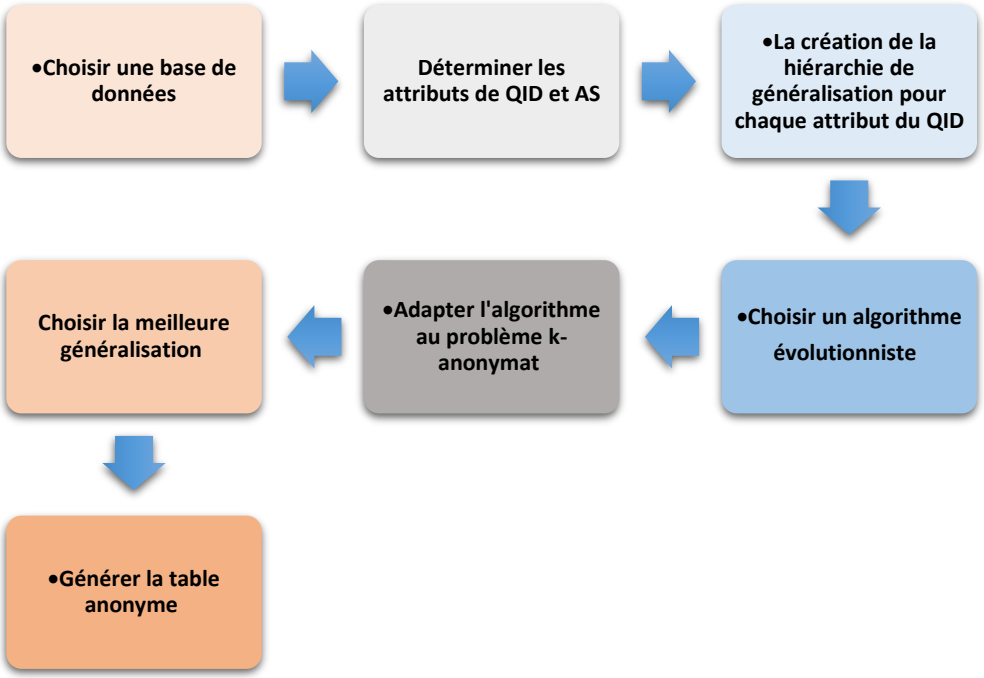

Figure III-1: les étapes suivantes pour réaliser notre approche.

III.3. Outils de développement

Pour pouvoir réaliser notre projet, nous avons utilisé un ensemble d'outils informatiques adéquats et performants pour obtenir les meilleurs résultats.

Nous présentons ci-dessous les outils que nous avons choisi durant le développement :

III.3.1. Software :

III.3.1.1. Langage utilisé :

Python :

Python est un langage de programmation interprété inventé par "Guido van Rossum". La première version de python est sortie en 1991 et est l'un des langages de programmation les plus intéressants du moment, facile à apprendre, et son code est plus lisible et plus facile à maintenir. Python est simple et puissant. **[44]**

De plus, ce langage est utilisé dans le domaine de la science des données, l'intelligence artificielle, l'apprentissage automatique, l'apprentissage en profondeur, etc. Avec des bibliothèques et packages riches tels que : Pandas, Agate, Bokeh, NumPy, Scikit-learn, PyBrain, TensorFlow, Matplotlib...**[45]**

III.3.1.2. Bibliothèque utilisée :

Pandas :

Pandas **[45]** est l'une des bibliothèques de science des données les plus populaires. Elle a été développée par des « Data Scientists » habitués à R et Python, Elle est aujourd'hui utilisée par un grand nombre de scientifiques et d'analystes.

Elle fournit de nombreuses fonctionnalités natives très utiles. En particulier, il est possible de lire des données provenant de nombreuses sources, de créer de grandes bases de données à partir de ces sources et d'effectuer des analyses agrégées en fonction des questions pour lesquelles des réponses sont attendues. Elle peut également être utilisée pour la manipulation de tableaux numériques et de séries chronologiques.

III.3.1.3. Plateforme et l'environnement de développement :

Anaconda :

Anaconda est une distribution gratuite et open-source des langages de programmation Python et R pour le développement d'applications dédiées à la science des données et à l'apprentissage automatique (traitement de données à grande échelle, analyse prédictive, calcul scientifique), conçue pour simplifier la gestion et le déploiement des packages. Les versions de package sont gérées par le système de gestion de packages conda.**[46]**

Jupyter :

Jupyter est une application Web open source pour la programmation dans plus de 40 langages de programmation, dont Python, Julia, Ruby, R ou Scala.

Jupyter est une évolution du projet IPython 2014 de "Fernando Pérez" et "Brian Granger", conçu pour prendre en charge la science des données interactive et le calcul scientifique dans tous les langages de programmation. [**47] [48]**

Google Colab :

Google Colab ou Colaboratory est un service cloud fourni par Google (gratuitement) et basé sur Jupyter Notebook. Il s'agit d'un environnement particulièrement adapté à l'apprentissage automatique et à l'analyse de données. Cette plateforme permet d'exécuter du code python directement dans le cloud. Par conséquent, il n'est pas nécessaire d'installer quoi que ce soit sur notre ordinateur autre qu'un navigateur. **[49]**

III.3.2. Hardware :

Nous avons utilisé une machine virtuelle Colab pour réaliser notre projet. Le Tableau III-1 décrit les différentes caractéristiques de cette machine.

Tableau III-1: caractéristique de machine virtuelle Colab.

Modèle du processeur	Intel Xeon Processor withtwocores@
Fréquence CPU	2.30GHz
Disque dur	108 GB
RAM	13GB

III.4. Dataset utilisé

Avant de passer à la réalisation de notre projet, nous aurons besoin d'un ensemble de données pour tester notre approche.

Pour tester et valider notre implémentation, il est nécessaire d'utiliser un ensemble de données publiques. Nous avons choisi la base "Adult Data Set" (Figure III-2). Aussi, appelé ensemble de données sur le revenu du recensement ("Census Income" dataset) qui représente les caractéristiques démographiques d'un échantillon de la population américaine. Notons que la description de la base est faite avant la partie expérimentation, Cela est dû au fait que nous utiliserons les éléments de cette base comme des exemples pour expliquer les étapes et les concepts de notre travail, Dans Le Tableau III-2 suivante, nous présentons les différentes caractéristiques de la base "Adult ".**[50]** (Lien de téléchargement du dataset "Adult")

Tableau III-2: caractéristique du dataset "Adult".

Nombre d'instances	32561
Nombre d'attributs	14
Caractéristique des attributs	Multivariées
Type des attributs	Catégorial, Continu
Valeurs manquantes ?	Oui

	age	workclass	fnlwgt	education	educationNum	maritalStatus	occupation	relationship	race	sex	capitalGain	capitalLoss	hoursPerWeek	native
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	Unite
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	Unite
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	Unite
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	Unite
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
32556	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	Unite
32557	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	Unite
32558	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	Unite
32559	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	Unite
32560	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	Unite

32561 rows × 15 columns

Figure III-2: Dataset utilisé.

III.4.1. Les attributs :

Comme nous avons indiqué dans le TableauIII-2 la base contient 14 attributs. On distingue deux catégories (5 sont numériques et 8 sont catégoriels) et un attribut représente l'attribut de classe qui signifie le revenu annuel d'un individu (>50K ou <=50K).**[6][50]**

- **Les attributs numériques :**

Tableau III-3: les attributs numériques du dataset "Adult".

Les attributs	***Signification***	***Les valeurs possibles***
Age	Age	[17-90]
fnlwgt	C'est un attribut de poids l'enquête qui représente un score démographique attribué à un individu	[12285-1484705]
educationNum	Le code du niveau d'étude	[1-16]
capitalGain	Le gain en capital	[0-99999]
capitalLoss	La perte en capital	[0-4350]

- **Les attributs catégoriels :**

Tableau III-4: les attributs catégoriels du dataset "Adult".

Les attributs	***Signification***	***Les valeurs possibles***
Workclass	La classe d'emploi	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
Education	Le niveau d'étude	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
Maritalstatus	L'état civil	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
Occupation	La profession	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
Relationship	La relation familiale	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
Race	*Race*	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
Sex	*Sexe*	Female, Male.
Nativecountry	Le pays d'origine	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

Une fois on a choisi le dataset, on passe à l'étape suivante, le prétraitement de données qui consiste de supprimer les données manquantes, après la réalisation de cette étape, nous allons obtenir une base de taille 30183 enregistrements. Voir la Figure III-3.

	age	workclass	fnlwgt	education	educationNum	maritalStatus	occupation	relationship	race	sex	capitalGain	capitalLoss	hoursPerWeek	nativ
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	Unit
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	Unit
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	Unit
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	Unit
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
30178	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	Unit
30179	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	Unit
30180	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	Unit
30181	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	Unit
30182	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	Unit

30183 rows × 15 columns

Figure III-3: le dataset après le prétraitement.

III.5. La généralisation

Nous avons considéré les 8 attributs (*Age, workclass, education,* maritalStatus, occupation, Race, Sex, nativeCountry) comme des attributs QID et l'attribut de classe est AS.

Pour appliquer le modèle k-anonymat, nous avons utilisé la technique de généralisation qui consiste à remplacer les valeurs du QID par des valeurs plus générales. Cette technique nécessite la définition d'une hiérarchie de généralisation pour chaque attribut du QID. Chaque hiérarchie contient au moins deux niveaux.**[1]** La

Figure III-4, Figure III-5 et Figure III-6 représentent respectivement la hiérarchie de généralisation de l'attribut "workclass", du "race" et du "age".

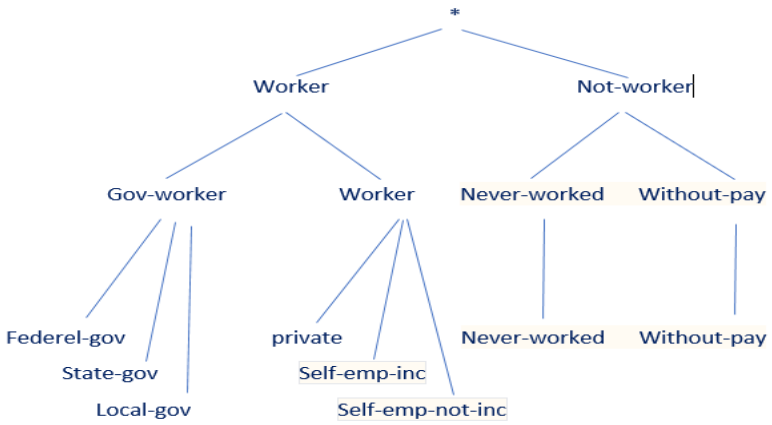

Figure III-4: La hiérarchie de généralisation de l'attribut "workclass".

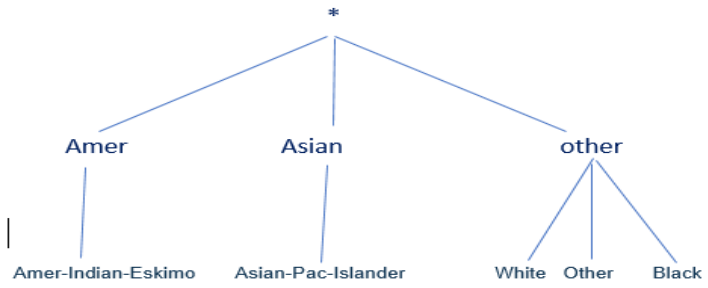

Figure III-5: La hiérarchie de généralisation de l'attribut "race".

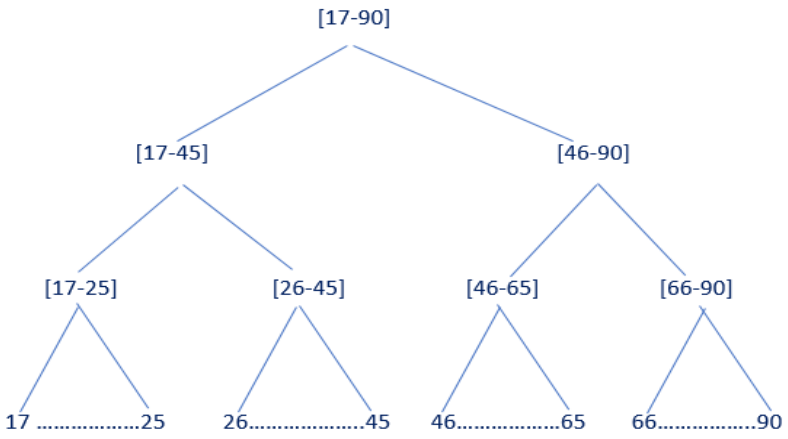

Figure III-6: La hiérarchie de généralisation de l'attribut "age".

III.6. Algorithme génétique

Nous avons choisi l'algorithme génétique (AG) pour implémenter la méthode k-anonymat afin de sélectionner la meilleure généralisation qui maximise la confidentialité des données et minimise la perte d'information.

Dans cette section, nous présenterons un aperçu de l'algorithme génétique et nous détaillerons les différents composants de l'algorithme génétique qui sont adaptés pour résoudre ce problème.

III.6.1. Aperçu globale sur AG :

AG est un algorithme évolutionnaire métaheuristique basé sur la méthode de population. Cette méthode fait référence à un algorithme qui traite plusieurs solutions à la fois. L'idée derrière les algorithmes génétiques est dérivée de la théorie de l'évolution de Darwin, elle utilise le concept de la sélection naturelle. C'est un algorithme de recherche globale itérative dont le but est de trouver la meilleure solution à des problèmes d'optimisation difficiles. **[44]**

Pour améliorer la qualité des individus, l'algorithme fait évoluer une population initiale pour un nombre prédéfini d'itérations. À chaque itération cet algorithme applique le processus de reproduction (sélection, croisement et mutation), certains individus seront supprimés de la population et d'autres seront insérés créant de plus en plus des individus plus adaptés pour leur environnement, donc des solutions optimales. **[45]**

Cet algorithme fonctionne de manière suivante : **[44]**

- Générer la population initiale ;
- Évaluer et sélectionner les individus de la population ;
- Application des croisements et des mutations ;
- Insérer des nouveaux individus dans la population ;
- Répétez la procédure jusqu'à ce que la condition d'arrêt soit satisfaite ;

III.6.2. Adapter AG au problème k-anonymat :

La Figure III-7 présente les principales étapes de notre implémentation de l'algorithme génétique afin de trouver la solution optimale pour la méthode k-anonymat.

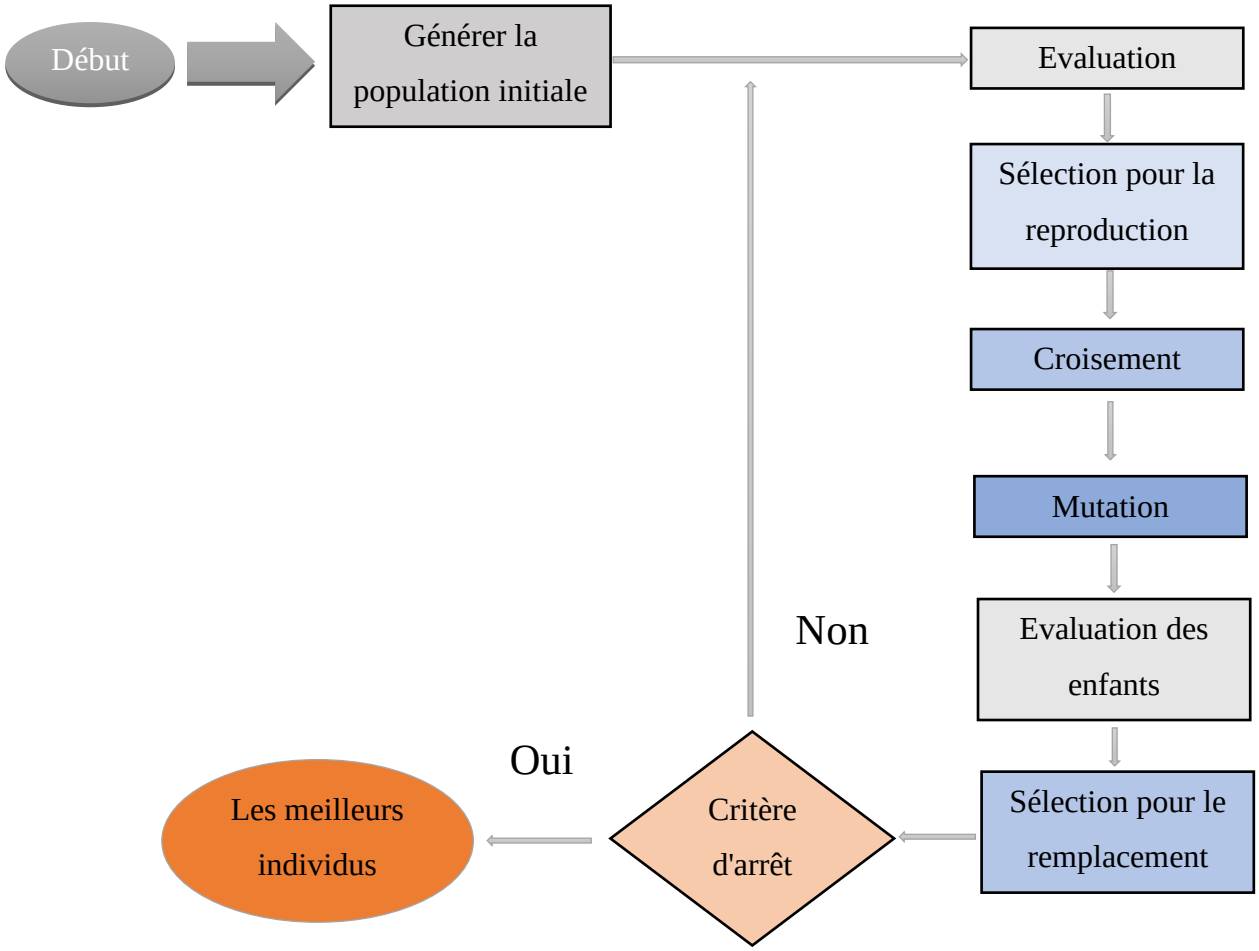

Figure III-7: les étapes de l'algorithme génétique.

Ce principe n'est pas très difficile à mettre en œuvre. En effet, la difficulté réside plutôt dans le choix des paramètres : les opérateurs génétiques (sélection, crossover et mutation), le nombre des itérations, la taille de population. **[53]**Nous présentons ci-dessous les différents paramètres adaptés pour obtenir des bons résultats.

III.6.2.1. Codage (représentation) d'un individu :

Il existe trois principaux types de codages utilisés par l'algorithme génétique : le codage binaire, qui est une représentation sous forme de chaîne binaire et le codage réel, qui représente directement la valeur réelle d'une variable et le gray. D'une manière générale, le codage réel est le plus utilisé. **[46]**

Notre approche utilise une représentation réelle pour encoder les éléments de la population afin d'offrir des solutions plus efficaces. Un individu (chromosome) de la population est représenté par un vecteur de généralisation de la table "Adult", il se compose de 8 gènes (Figure III-6) chaque gène indique le niveau de généralisation d'un attribut de QID de la table. (Figure III-8)

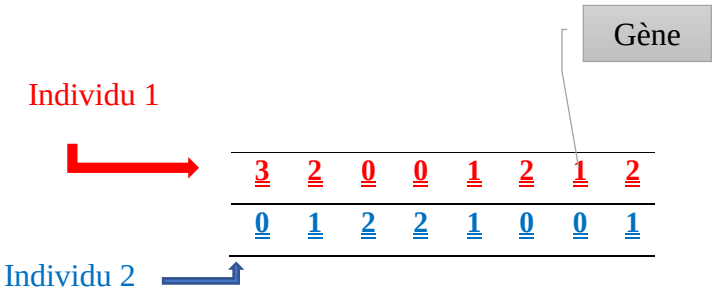

Figure III-8: représentation d'un individu.

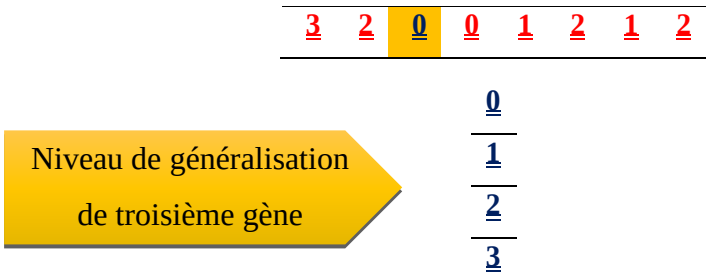

Figure III-9: un exemple de gène avec le niveau de généralisation.

III.6.2.2. La population initiale :

Le choix de la population initiale est important car il peut rendre plus ou moins rapide la convergence vers l'optimum globale**. [46]**Dans notre cas la position de l'optimum dans l'espace de recherche est totalement inconnue, donc la population initiale (de taille N individus) est générée de manière aléatoire afin d'explorer le plus largement possible l'espace de recherche. Dans notre cas nous avons généré que des individu valide c'est-à-dire la valeur d'un gène ne dépasse jamais le niveau de généralisation pour chaque attribut du QID, par exemple le niveau de généralisation de l'attribut "workclass" est limité entre 0 et 3.

III.6.2.3. La fonction de fitness (objective ou évaluation) :

Le choix de le fonction objective est une étape importante pour obtenir une bonne solution. Cette fonction associe une valeur de performance à chaque individu, ce qui offre la possibilité de déterminer les meilleurs individus de la population dans un sens bénéfique pour la recherche de la meilleure solution. Dans le contexte de notre modèle pour déterminer le k optimum, plus le K augmente plus la confidentialité des données est forte, et plus le niveau de généralisation de chaque attribut du QID est diminué plus l'utilité des données augmente. De manière plus formelle, la fonction fitness est défini comme suit :

$$Fit = 0.5 * \frac{1}{K} + 0.5 * \frac{\sum_{j=1}^{|QID|} \frac{hAj}{|VGHAj|}}{|\mathrm{QID}|}$$

Pour calculer l'utilité, nous avons utilisé la métrique de précision :

$$\mathrm{PREC} = 1 - \frac{\sum_{j=1}^{|QID|} \sum_{i=1}^{|T|} \frac{h}{|VGHAj|}}{|\mathrm{T}| \ * \ |\mathrm{QID}|}$$

Où :

- K est le seuil de k-anonymat
- |QID | : est le nombre des attributs du QID
- hAj représente la hauteur de la hiérarchie de généralisation de l'attribut Aj
- |VGHAj| est la hauteur totale de la hiérarchie de généralisation de l'attribut Aj.

Selon la fonction de fitness, un individu est meilleur s'il a la plus petite valeur de la fonction fitness, Donc on est face à un problème de minimisation.

III.6.2.4. Les opérateurs génétiques :

C'est un ensemble de processus effectué pour générer de nouveaux chromosomes, à partir de la population précédente. On distingue trois opérateurs différents : la sélection, le croisement et la mutation. **[55]**

- **La sélection :**

La sélection est un opérateur utilisé pour identifier les meilleurs individus d'une population et d'éliminer les mauvais. **[55]** Dans la littérature, on trouve plusieurs techniques de sélection : sélection par rang, Sélection par roulette, sélection par tournoi et sélection uniforme. Dans notre problème sujet d'étude, on a choisi d'utilisé la sélection par rang. Cette technique consiste à choisir les individus possédant les meilleurs scores d'adaptation, **[55]**exemple si la taille de population est N, cette sélection permet à conserver les M meilleurs individus au sens de la fonction fitness.

- **Le croisement (Crossover) :**

L'opérateur de Crossover a pour but d'enrichir la diversité de la population afin de générer de nouveaux individus par l'échange d'information entre les chromosomes. Il utilise deux parents (individus) pour former un ou deux enfants appelés aussi descendants **[47]**. Généralement, on trouve plusieurs types de croisements : croisement à k point et croisement uniforme. Nous avons utilisé le croisement à 2 points car c'est une solution largement utilisée, comme montre la Figure III-10.

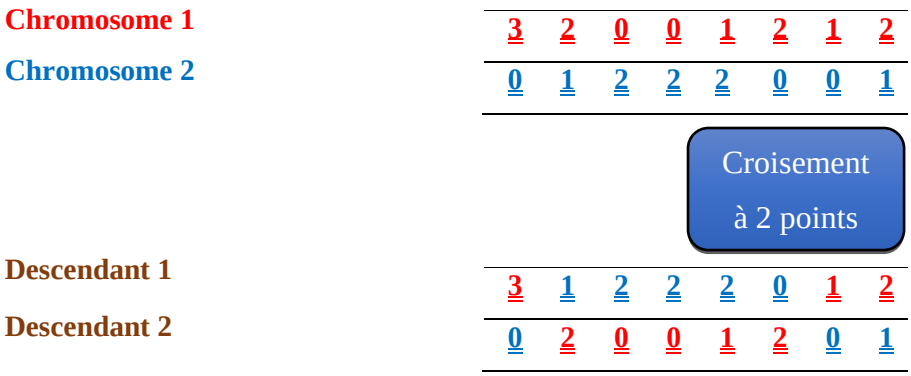

Figure III-10: l'opérateur de croisement.

- **La mutation :**

La mutation consiste à modifier aléatoirement, avec une certaine probabilité, la valeur d'un composant de l'individu**.** Cet opérateur a pour but de garantir l'exploration de l'espace d'état et pour améliorer la qualité de la nouvelle génération et éviter la convergence vers une solution optimale locale**.** Dans le contexte de notre approche, la mutation consiste à tirer aléatoirement un gène d'un individu et à le remplacer par la valeur le plus générale. (Figure III-11)

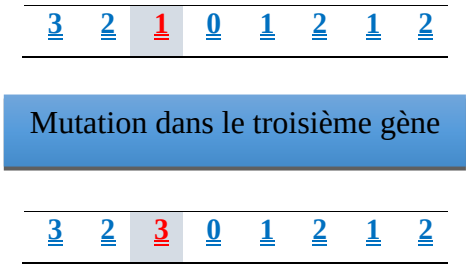

Figure III-11: l'opérateur de mutation.

Une fois les processus génétiques (sélection, croisement et mutation) sont terminés, il nous faut sélectionner les meilleurs individus N pour les insérer dans la nouvelle population. Dans cette étape, nous avons conservé la taille initiale de la

population pour éviter le risque de voir exploser le temps de traitement puisque la population de chaque génération sera plus grande. **[56]**

III.6.2.5. Critère d'arrêt :

AG est itératif, donc on a besoin de définir un critère d'arrêt, il existe plusieurs critères. Dans notre approche nous avons fixé a priori le nombre d'itération pour trouver une solution dans un temps limité**.**

III.7. Expérimentations

Pour évaluer les performances de notre implémentation, nous avons appliqué les tests sur les 5000 enregistrements un échantillonnage aléatoire du dataset "Adult".

Nous avons fixé les paramètres métaheuristiques liés à l'algorithme génétique (TableauIII-5) pour analyser et interpréter les résultats obtenus.

Tableau III-5: les paramètres métaheuristiques.

La taille de population	**5**
La taille d'individus	8
Probabilité de croisement (Pc)	0.9
Probabilité de mutation (Pm)	0.1

Dans cette expérimentation, nous étudions la qualité des résultats obtenus après 20 itérations sur un population de taille 5.À cet effet, nous avons généré dans le premier cas la population initiale de manière aléatoire, et dans le deuxième cas nous avons inséré dans la population initiale le vecteur le plus générale et le vecteur le moins générale et le reste de la population est généré de manière aléatoire. La Figure III-12 présente les résultats obtenus de cette expérimentation.

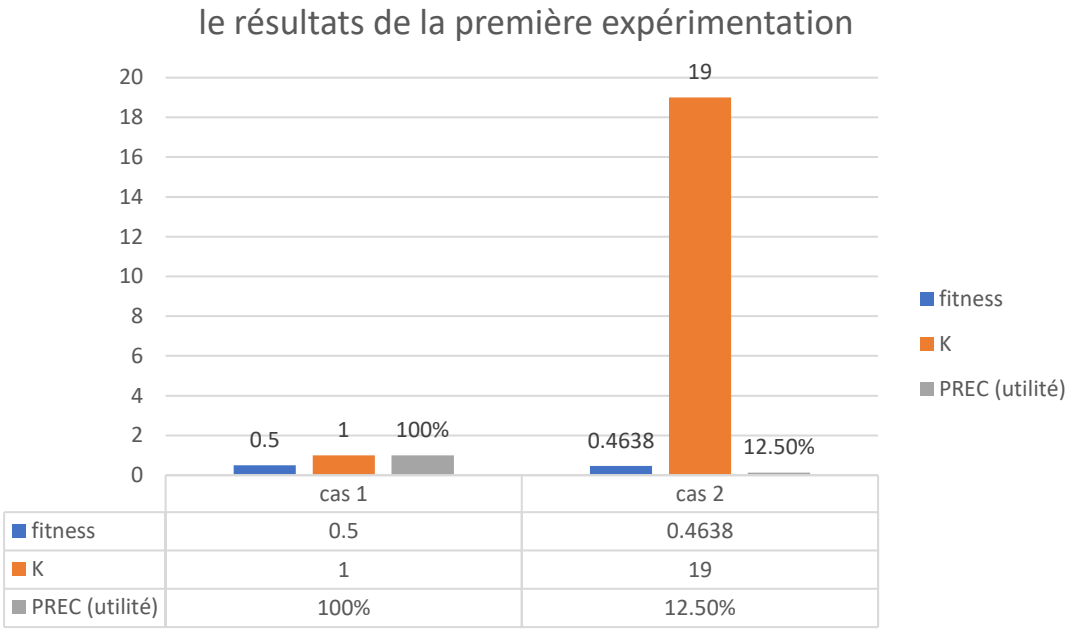

Figure III-12: le résultat de la première expérimentation.

L'expérimentation montre que l'insertion des individus : vecteur le plus générale et le vecteur le moins générale (le deuxième cas) à beaucoup améliorer les résultats par rapport au premier cas, ou la population est complètement générée aléatoire. Dans ce dernier cas, la meilleure valeur de k est 1, ce qui implique un taux de protection de 0%. Pour cela nous avons utilisé le deuxième cas pour obtenir des meilleurs résultats dans la deuxième expérimentation.

Dans la deuxième expérimentation, nous avons fixé les paramètres (la taille de la population, Pm, Pc…) (voir Tableau III-5), et nous avons varié le nombre des itérations. La Figure III-13, la Figure III-14 et la Figure III-15 montrent respectivement les résultats obtenus de fitness, de K et de l'utilité par rapport au nombre des itérations.

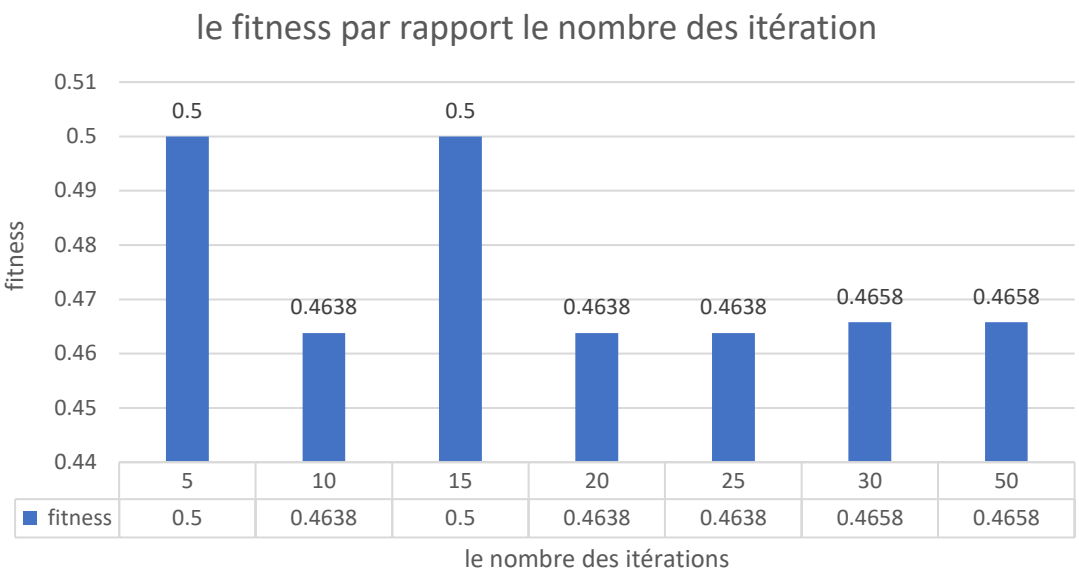

Figure III-13:l'influence du nombre des itérations sur la fonction fitness.

Dans le graphe de la Figure III-13 nous remarquons un comportement typique de notre algorithme, c'est-à-dire, d'une manière générale, la fonction fitness s'améliore (un problème de minimisation dans notre cas) en augmentant le nombre d'itérations. Des exceptions sont dû à l'aspect aléatoire de l'algorithme génétique. Nous remarquons aussi le début de convergence de l'algorithme à partir de la 20eme itération. Au-delà de l'itération 50, les résultats restent stables.

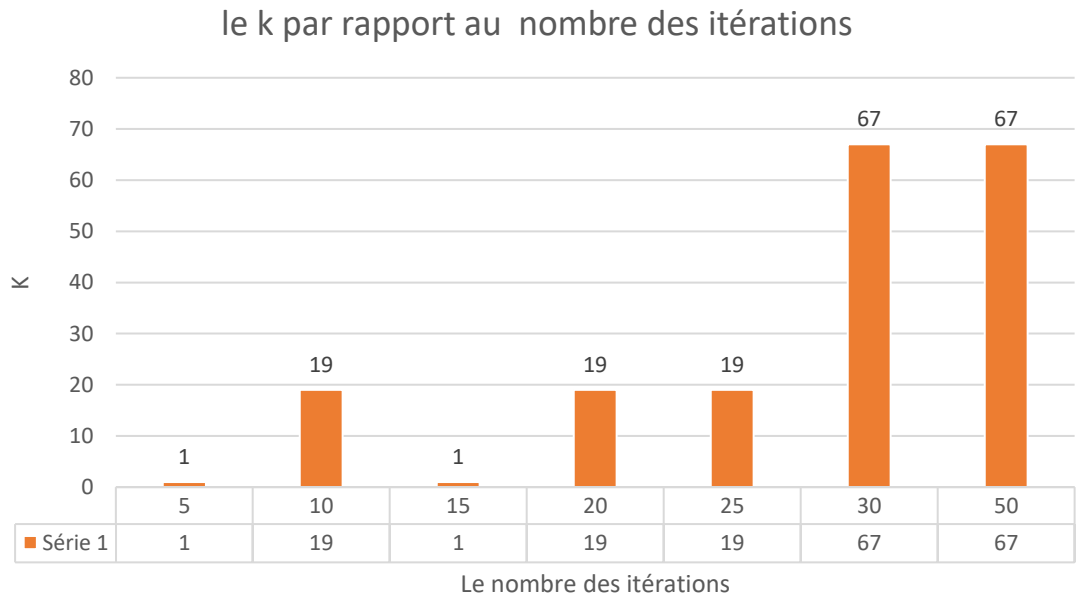

Figure III-14: L'influence du nombre des itérations sur la valeur de K.

La même remarque est faite pour la valeur de K (Figure III-14). Il y a une corrélation positive entre k et les nombres des itérations. Dans cette expérimentation la convergence est claire à partir de l'itération 30. Après la 50ème itération les résultats restent invariés. Notons aussi, que la meilleure valeur obtenue pour le k, est la valeur 67. Cette valeur représente un taux de protection considérable, c'est-à-dire que la probabilité pour identifier une personne particulière dans la base est de 0.014 (1/67).

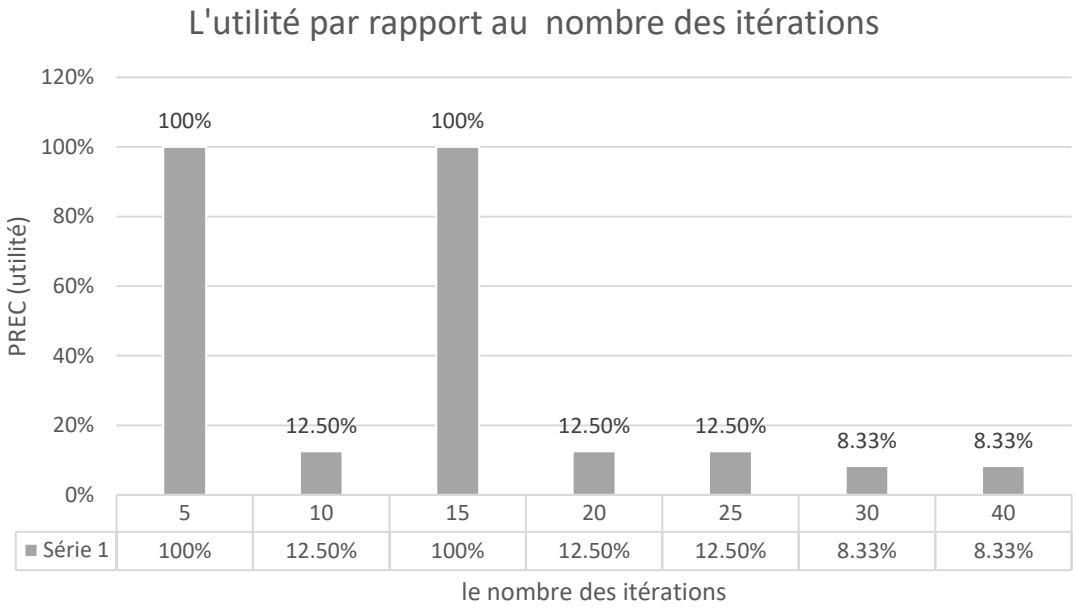

Figure III-15: L'influence du nombre des itérations sur l'utilité.

Pour les résultats de l'utilité (Figure III-14), nous remarquons un effet inverse, c'est-à-dire, le nombre des itérations influence négativement sur l'utilité des données. En terme général, si le nombre des itérations augmente, l'utilité diminue. Cela, a aussi une relation avec le taux de protection (la valeur de K). Nous remarquons dans le cas où le nombre des itérations est égale à 5, l'utilité a pris une valeur de 100%, c'est-à-dire qu'on est face à une base non anonymisée avec un taux de protection de 0%, par conséquent une valeur de k égale à 1.

À la fin, on peut dire qu'il est difficile de dire qu'une solution est bonne ou mauvaise (sauf pour les cas extrêmes), du fait qu'il faut toujours faire un compromis

entre le taux de protection et l'utilité des données. Cela peut varier selon plusieurs paramètres, comme le contexte de l'utilisation des données ou la nature des individus de la base. L'avis d'un expert est toujours apprécié dans ce type de situation.

III.8. Conclusion

Dans ce chapitre nous avons implémenté la méthode k-anonymat à la base de l'algorithme génétique. L'objectif était de déterminer la meilleure généralisation qui maximise à la fois la confidentialité et l'utilité des données. Nous avons commencé par les outils et les données utilisées, ainsi une vue générale sur les AG et à la fin une présentation des résultats obtenus a été discuté. Nous pouvons dire que l'implémentation proposée a donné des résultats satisfaisants.

Conclusion Générale

Avec l'augmentation rapide de la quantité de données, ces données sont devenues de plus en plus importantes dans le développement de divers secteurs et services. Cela, augmenté la nécessité de développent des approches de protection et de donner des garanties d'anonymisation.

Dans ce travail, nous avons étudié le modèle de protection « k-anonymat ». L'idée derrière ce modèle, consiste à rendre difficile d'établir d'un lien entre l'identité d'un individu particulier et l'une de ces données (ou attributs) privées et sensibles. Du fait, nous avons fait une implémentation de ce modèle en adaptant un algorithme génétique pour ce type de problème. Notre objectif était de maximiser à la fois la confidentialité et l'utilité des données. Les différents tests réalisés sur le jeu de données nous ont permis de dire que l'implémentation proposée donne des résultats satisfaisants. Loin d'être parfaits, ces résultats peuvent être améliorés pour obtenir de meilleures solutions.

Comme perspectives nous envisageons d'améliorer notre implémentation à travers des tests plus poussés, pour définir d'une manière plus exacte les paramètres de notre algorithme génétique, comme la taille de la population initiale, et les taux de croisement et de mutation. D'autres perspectives comme le test de notre implémentation sur d'autres bases et la comparaison de notre travail avec d'autres travaux sont aussi envisageables. A long terme nous pourrons explorer d'autres facettes de la protection de la vie privée soit avec d'autres algorithmes ou avec d'autres modèles de protection.

Bibliographie

[1] SAMARATI, Pierangela et SWEENEY, Latanya. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. 1998.

[2] MEYERSON, Adam et WILLIAMS, Ryan. On the complexity of optimal k-anonymity. In : Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. 2004. p. 223-228.

[3] LEBRUN, Pierre-Brice. La vie privée. Empan, 2015, no 4, p. 168-172.

[4] UNIES, Nations et PUYBARET, Eric. Déclaration universelle des droits de l'homme. Département de l'information de l'ONU, 2008.

[5] Une vie privée est-elle possible sur Internet ? - Tendance Droit.

[6] BELABED, Amine. La protection de la vie privée sur Internet. Thèse de doctorat. Université de Tlemcen-Abou Bekr Belkaid.

[7] P. J.-P. S. e. O. PREVOST, «QU'EST CE QUE LE PRINCIPE DE MINIMISATION AU SEIN DU RGPD ?,» 24 6 2022.

[8] LIFRANGE, Marie, et al. Protection des données à caractère personnel: le consentement à l'épreuve de l'ère numérique. 2018.

[9] KHELIFI, Lazhar, TOURÉ, Fodé, et AÏMEUR, Esma. Protection de la Vie Privée dans les Réseaux Sociaux d'Entreprise.

[10] DESWARTE, Yves. Sécurité et protection de la vie privée sur Internet. 2002.

[11] Guide-Non-a-la-fraude_2014.pdf.

[12] WALTER, Jean-Philippe. Le profilage des individus à l'heure du cyberespace: un défi pour le respect du droit à la protection des données. Datenverknüpfung: Problematik und rechtlicher Rahmen, Zürich: Schulthess, 2011, p. 87-114.

[13] GARRIE, Daniel B., BLAKLEY, Alan F., et AMSTRONG, Matthew J. Legal Status of Spyware. Fed. Comm. LJ, 2006, vol. 59, p. 157.

[14] TRUDEL, Pierre, ABRAN, France, et DUPUIS, Gabriel. Analyse du cadre réglementaire québécois et étranger à l'égard du pourriel, de l'hameçonnage et des logiciels espions. 2007.

[15] Crim. 14 mars 2006, Bull. crim., n.69.

[16] HUET, J., DREYER, E., Droit de la communication numérique, coll. "Manuel", Paris, éd. LG.D.J Lextenso, 2011, p.328.

[17] Les Cookies disponibles sur :https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1445194-cookies-sur-internet-definition-technique-normes-en-cours-et-role, consulter sur 14 avril 2022,

[18] L'attaque par Phishing , disponible:https://www.arobase.org/phishing/phishing.htm,consulté le 15avril 2022.

[19] DEL PRADO CORTEZ, Miguel Nuñez. Attaques d'inférence sur des bases de données géolocalisées. 2013. Thèse de doctorat. INSA de Toulouse.

[20] SWEENEY, Latanya. k-anonymity: A model for protecting privacy. International journal of uncertainty, fuzziness and knowledge-based systems, 2002, vol. 10, no 05, p. 557-570.

[21] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression..

[22] BENKHELIF, Tarek. Publication de données individuelles respectueuse de la vie privée: une démarche fondée sur le co-clustering. 2018. Thèse de doctorat.

Université de Nantes.

[23] LI, Ninghui, LI, Tiancheng, et VENKATASUBRAMANIAN, Suresh. t-closeness: Privacy beyond k-anonymity and l-diversity. In : 2007 IEEE 23rd international conference on data engineering. IEEE, 2006. p. 106-115.

[24] FUNG, Benjamin CM, WANG, Ke, CHEN, Rui, et al. Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys (Csur), 2010, vol. 42, no 4, p. 1-53.

[25] CETINTAS, Suleyman et AKOVA, Ferit. Generalizations with Probability Distributions for Data Anonymization. 2008.

[26] CIRIANI, Valentina, CAPITANI DI VIMERCATI, S. De, FORESTI, Sara, et al. Microdata protection. In : Secure data management in decentralized systems. Springer, Boston, MA, 2007. p. 291-321.

[27] DOMINGO-FERRER, Josep, SÁNCHEZ, David, et SORIA-COMAS, Jordi. Database anonymization: privacy models, data utility, and microaggregation-based inter-model connections. Synthesis Lectures on Information Security, Privacy, & Trust, 2016, vol. 8, no 1, p. 1.

[28] SAKHAR, A. et GANAR, S. Anonymization: A method to protect sensitive data in cloud. International Journal of Scientific & Engineering Research, 2013, vol. 4, no 5.

[29] SADHWANI, Divya, SILAKARI, Sanjay, et CHOURASIA, Uday. Preserving Privacy during Big Data Publishing using K-Anonymity Model--A Survey. International Journal of Advanced Research in Computer Science, 2017, vol. 8, no 5.

[30] DOMINGO-FERRER, Josep, SÁNCHEZ, David, et HAJIAN, Sara. Database privacy. In : Privacy in a Digital, Networked World. Springer, Cham, 2015. p. 9-35.

[31] SAMARATI, Pierangela. Protecting respondents identities in microdata release. IEEE transactions on Knowledge and Data Engineering, 2001, vol. 13, no 6, p.

1010-1027.

[32] SWEENEY, Latanya. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, vol. 10, no 05, p. 571-588.

[33] BEN FREDJ, Feten. Méthode et outil d'anonymisation des données sensibles. 2017. Thèse de doctorat. Paris, CNAM.

[34] COX, Lawrence H. Suppression methodology and statistical disclosure control. Journal of the American Statistical Association, 1980, vol. 75, no 370, p. 377-385.

[35] LI, Tiancheng et LI, Ninghui. On the tradeoff between privacy and utility in data publishing. In : Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. 2009. p. 517-526.

[36] MAKHDOUMI, Ali et FAWAZ, Nadia. Privacy-utility tradeoff under statistical uncertainty. In : 2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton). IEEE, 2013. p. 1627-1634.

[37] LOUKIDES, Grigorios et SHAO, Jianhua. Data utility and privacy protection trade-off in k-anonymisation. In : Proceedings of the 2008 international workshop on Privacy and anonymity in information society. 2008. p. 36-45.

[38] KARR, Alan F., KOHNEN, Christine N., OGANIAN, Anna, et al. A framework for evaluating the utility of data altered to protect confidentiality. The American Statistician, 2006, vol. 60, no 3, p. 224-232.

[39] IYENGAR, Vijay S. Transforming data to satisfy privacy constraints. In : Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. 2002. p. 279-288.

[40] LEFEVRE, Kristen, DEWITT, David J., et RAMAKRISHNAN, Raghu. Mondrian multidimensional k-anonymity. In : 22nd International conference on data engineering (ICDE'06). IEEE, 2006. p. 25-25.

[41] BAYARDO, Roberto J. et AGRAWAL, Rakesh. Data privacy through optimal k-anonymization. In : 21st International conference on data engineering (ICDE'05).

IEEE, 2005. p. 217-228.

[42] LUNACEK, Monte, WHITLEY, Darrell, et RAY, Indrakshi. A crossover operator for the k-anonymity problem. In : Proceedings of the 8th annual conference on Genetic and evolutionary computation. 2006. p. 1713-1720.

[43] DEWRI, Rinku, RAY, Indrajit, RAY, Indrakshi, et al. On the Optimal Selection of k in the k-Anonymity Problem. In : 2008 IEEE 24th International Conference on Data Engineering. IEEE, 2008. p. 1364-1366.

[44] *python disponible sur:* Jupyter — Wikipédia (wikipedia.org),consulter le 18 juin 2022.

[45] Apercu globale sur AG disponible sur : https://www.lebigdata.fr/python-langage-definition, consulter le 18 juin 2022.

[46] Anaconda disponible https://fr.wikipedia.org/wiki/Anaconda_(distribution_Python), consulter le 18 juin 2022.

[47] Jupyter disponible sur :https://fr.wikipedia.org/wiki/Jupyter, consulter le 18 juin 2022.

[48] Jupyter disponible sur : https://jupyter.org/about, consulter le 15 juin 2022.

[49] Google Colab : https://ledatascientist.com/google-colab-le-guide-ultime/, consulter le 20 juin 2022.

[50] Adult Data Set disponible sur : https://archive.ics.uci.edu/ml/datasets/adult,consulter le 15 juin 2022.

[51] HAJLAOUI, Rejab. Résolution à base d'heuristiques du problème de routage dans les réseaux ad hoc de vehicules. 2018. Thèse de doctorat. Bourgogne Franche-Comté.

[52] AMINE, LADJICI Ahmed. Calcul Evolutionnaire Application sur l'Optimisation de la Planification de la Puissance Réactive. 2005. Thèse de doctorat. ECOLE NATIONALE POLYTECHNIQUE.

[53] VALLÉE, Thomas et YILDIZOĞLU, Murat. Présentation des algorithmes génétiques et de leurs applications en économie. Revue d'économie politique, 2004, p. 711-745.

[54] DURAND, Nicolas. Algorithmes Génétiques et autres méthodes d'optimisation appliqués à la gestion de trafic aérien. 2004. Thèse de doctorat. INPT.

[55] La sélection disponible sur : http://igm.univ-mlv.fr/~dr/XPOSE2013/tleroux_genetic_algorithm/fonctionnement.html,consulter le 20 juin 2022.

[56] la mutation disponible sur : https://khayyam.developpez.com/articles/algo/genetic/#LIII-C, consulter le 20 juin 2022.