REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE AKLI MOAND OULHADJE-BOUIRA

Faculté des Sciences et des Sciences Appliquées
Département Math et informatique

Mémoire de fin d'étude

Présenté par :
Haboussi Omar
Guenoune Kheireddine

En vue de l'obtention du diplôme de **Master 02** en **:**

Filière : **INFORMATIQUE**

Option *: Génie des Systémes Informatiques*

Thème :

Protection de la vie privée (anonymat) dans les réseaux sociaux

Devant le jury composé de :

	UAMOB	Président
Mr.Amad.Mourad	UAMOB	Encadreur
	UAMOB	Examinateur
	UAMOB	Examinateur

Année Universitaire 2018/2019

Remerciements

Au terme de ce travail, je remercie avant tout Dieu le tout puissant qui a éclairémon chemin tout au long de mes études.

Nous remercions chaleureusement les membres de jury qui examineront notre travail. On tient à leur témoigner toute notre gratitude pour leurs remarques et leurs commentaires pertinents.

En particulier, nous remercions notre chére encadreur Mr.Mourad Amad qui était toujours à notre écoute sans réservation pendant la préparation de ce mémoire de fin de Cycles, ce sont ses conseils qui nous ont aidés à réaliser ce travail

Merci enfin à tous nos amis.

Dédicaces

Je dédie ce travail à mes chers parents. Qui n'ont jamais cessé de m'encourager. à tous mes chers frères, à chaque membre de ma famille, et à tous les camarades de ma promo.

Guenoune Kheir eddine.

Dédicaces

Je dédie ce PFE à mes chers parents qui m'ont aidé à comprendre la vie, à tous mes chers frères et sœurs, à chaque membre de ma famille, ainsi qu'à mes amis, et à tous les camarades de ma promo.

Haboussi Omar.

Table des matières

Table des figures

Liste des équations

Liste des Tableaux

Abreviations

PD : la pénalité de discernement.

NCP : Normalized Certainty Penalty.

Résumé

Les données personnelles et les informations sensibles dans les réseaux sociaux sont des problèmes importants. Ces dernières années, différentes approches ont été adoptées comme solutions pour protéger la vie privée dans les réseaux sociaux. Dans le même contexte, le présent travail introduit une nouvelle méthode d'anonymisation. L'évaluation des performances montre que la méthode proposée fournit des meilleurs résultats, car les données obtenues sont totalement différentes des données originales, avec une perte de données minimale.

Mots clés : vie privée dans les réseaux sociaux, privacy, protection des données personnelles.

Abstract

Personal data and sensitive information in social networks is important issues. In recent years, different approaches have been adopted as a solution to protect privacy in social networks. In the same context, the present work introduces a novel anonymization method. The performance evaluation of the proposed method provides better results, because the data obtained are completely different from the original data, with minimal data loss.

Keywords: privacy in social networks, privacy, protection of personal data.

ملخص

البيانات الشخصية والمعلومات الحساسة في الشبكات الاجتماعية هي قضايا مهمة. في السنوات الأخيرة ، تم اعتماد أساليب مختلفة كحل لحماية الخصوصية في الشبكات الاجتماعية. في اطار هذا السياق ، يقدم العمل الحالي طريقة لإخفاء الهوية ، توفر هذه الطريقة نتائج أفضل لأن البيانات التي تم الحصول عليها مختلفة تمامًا عن البيانات الأصلية. مع الحد الأدنى من فقدان البيانات.

الكلمات الرئيسية: الخصوصية في الشبكات الاجتماعية ، الخصوصية ، حماية البيانات الشخصية

Introduction générale

Les communautés électroniques existent depuis la création d'Internet. À l'origine, c'étaient des forums et des listes de diffusion qui fournissaient à des personnes du monde entier les moyens de communiquer et d'échanger des informations sur des sujets spécifiques. Actuellement, elles sont sous forme de réseaux sociaux.

La plupart des utilisateurs de réseaux sociaux partagent une grande partie de leurs informations privées dans leur espace de réseau social. Ces informations peuvent être des informations de contact, commentaires, images, vidéos etc. Les utilisateurs de réseau social ont des relations sociales parmi eux pour suivre d'autres utilisateurs et communiquent avec eux, comme sur Twitter où en établissant des amitiés, comme sur Facebook aussi où les utilisateurs communiquent entre eux en publiant leurs activités et leurs intérêts via des messages courts. Les utilisateurs peuvent lire les messages de leurs abonnés et amis, et ils peuvent répondre via des réactions telles que "J'aime" sur Facebook et « *Retweet* » ou *« intérêt* » sur Twitter.

De nombreux utilisateurs publier leurs informations publiquement sans un examen attentif. Donc les réseaux sociaux sont devenus un grand refuge de données sensibles à cause du stockage énorme d'informations et l'accessibilité simple, les réseaux sociaux sont devenus des nouvelles cibles qui attirent les attaquants. La sécurité des réseaux sociaux est liée principalement à la vie privée des utilisateurs. Donc, toute attaque sur un réseau social menace directement sa sécurité conduit indirectement à la violation de la vie privée des acteurs. Des méthodes d'anonymisation et des algorithmes ont été proposées pour protéger la vie privée des utilisateurs dans les réseaux sociaux tels que le K-anonymat et le l-diversité. La vie privée des utilisateurs doit être protégée contre toutes les menaces. C'est dans cette problématique que s'inscrit ce mémoire de fin de cycle Master.

Le présent document est structuré de la manière suivante : dans le premier chapitre, nous présentons des concepts généraux sur les réseaux sociaux ainsi que les menaces de sécurité dans ces réseaux. Nous introduisons dans le second chapitre la vie privée et l'anonymat dans les réseaux sociaux lorsque ces données sont publiées. Le troisième chapitre est consacré à la description de l'approche d'anonymisation proposée pour l'anonymat dans le cadre des réseaux sociaux et dans le quatrième chapitre nous évaluons la solution proposée.

Enfin on conclut le mémoire par un rappel sur les différents points abordés et nous présentons quelques perspectives que l'on souhaite réalisé dans l'avenir

Chapitre 1

1.Généralités sur les réseaux sociaux

1.1.Introduction

Les réseaux sociaux se sont élargis au cours des dernières décennies et sont considérés comme le point de repère de l'ère du Web 2.0. Près de la moitié des utilisateurs ayant accès à Internet sont membres d'au moins un réseau social en ligne. Des sites tels que Facebook, Myspace[1], Twitter[2], Google+[3], LinkedIn[4] et bien d'autres ont attiré des utilisateurs de tous âges et de tous horizons. De nombreux utilisateurs ont intégré les sites de réseaux sociaux dans leurs routine et pratiques quotidiens, les réseaux sociaux modernes offrant diverses possibilités. Outre l'aspect social évident qu'ils couvrent, c'est-à-dire aider les gens à entrer en contact avec les autres et à socialiser, ils sont passionnés par les nouvelles activités et l'auto-publicité, la publicité personnelle et la recherche de débouchés professionnels.

Avec les nouveaux sites de réseaux sociaux crées chaque année, il est difficile de choisir celui qui convient le mieux à notre entreprise, utilisation professionnelle ou personnelle. Il est donc impératif que nous sachions quels sites de réseaux sociaux répondent à nos besoins et notre stratégie de communication.

[1] www.myspace.com

[2] www.Twitter.com

[3] www.google.com

[4] www.linkedIn.com

1.2.L’Evolution Du Web

Le Web est caracterisé par des évolutions constantes du fond et de la forme des pages Web, parmis ces évolutions on trouve :

1. Le web 1.0 fait référence à la première étape du World Wide Web, il s'agissait d'un ensemble de sites Web statiques qui ne fournissaient pas encore de contenu interactif. Il se caracterise par des sites orientés produits, qui sollicitent peu l'intervention des utilisateurs.

2. Le web 2.0, ou web social, change totalement de perspective. Il privilegie la dimension de partage et d'échange d'informations et de contenus (*textes, videos, images ou autres*). Il est la cause d'emergence des réseaux sociaux, des smartphones et des blogs. Tim O ’Reilly est le créateur du terme « *Web 2.0* », c’est dans son article « *Qu’est-ce que le web 2.0 ?* » que le président d’O ‘Reilly Media[5] explique le terme en septembre 2005. Il définit alors internet comme une plateforme et non plus comme un média. L’utilisateur maintenant est au centre du Web 2.0, il peut contrôler son activité sur internet.

Figure 1.1 Web 2.0

[5] Maison d’édition américaine

Web 2.0 est le terme utilisé pour décrire le Web social, où les sites de réseautage social occupent une place de choix dans les activités en ligne des utilisateurs. Le passage à ce Web plus interactif à partir du Web 1.0 s'est généralement produit en raison des changements technologiques qui ont rendu Internet et la capacité de développer du contenu. Ces changements incluent l'Internet haut débit, des meilleurs navigateurs. Selon Tristan Nitot, président de Mozilla Europe, « *c'est le Web tel qu'il a été imaginé par son inventeur, Tim Berners-Lee, un Web où chacun peut publier et consommer de l'information. Un Web où l'on est consommateur, et acteur à la fois. Consomm'acteur, en quelque sorte.* » **[1]**. L'un des concepts les plus importants du Web 2.0 qui sont directement compatibles avec les réseaux sociaux est l'idée de contenu généré par l'utilisateur (*UGC*) ou contenu généré par l'utilisateur. En fait, l'internaute ne télécharge pas seulement du contenu depuis différents endroits du web, mais crée également son propre contenu et le distribue comme il le souhaite via de nombreux médias tels que les réseaux sociaux par exemple. Enfin, nous remarquerons également une idée très récente qui montre clairement que le sujet est au cœur du Web, c'est le Web social, Internet est considéré comme un lieu de rencontres sociales entre les gens **[2]**.

Il n'est pas possible de déterminer exactement où se termine le Web 2.0 car il s'agit d'un changement qui s'est produit progressivement au fil du temps à mesure qu'Internet devenait plus interactif.

1.3.Qu'est-ce qu'un réseau social ?

Réseau social est un terme qui vient des domaines scientifiques de la sociologie et son concept théorique a été utilisé pour la première fois dans les sciences sociales. John A. Barnes **[3]** les 'inventeurs du terme et le premier à le définir en 1954 en disant que le réseau social est une structure sociale, qui comprend des individus ou des groupes liés par le même type d'activité, d'intérêts communs, d'amitié ou de relations (*Barnes, 1954*). Depuis lors, il a progressé et a acquis une grande popularité en tant que concept conduisant finalement à l'emploi dans d'autres sciences également, telles que les sciences de réseau.

Un réseau social est une plate-forme en ligne qui permet aux utilisateurs de créer un profil public et d'interagir avec les autres utilisateurs du site. D'où les utilisateurs peuvent communiquer et partager des photos ou des vidéos entre eux.

1.4.Type de sites de réseaux sociaux

Le développement des technologies de l'information, avec la capacité de collecter, d'analyser et de diffuser des informations, constitue une menace importante pour la vie privée des utilisateurs des réseaux sociaux. Il est avancé que la technologie de l'information est en train de subir des dégâts considérables et qu'elle s'accélère si rapidement. Les problèmes de confidentialité préoccupent totalement les utilisateurs face aux invasions potentielles dans le monde de l'Internet.

Il existe de nombreuses catégories de sites de réseaux sociaux où les utilisateurs aiment s'inscrire. Partage de photos entre profils, sites de localisation, téléchargement de sites de vidéos, sites pouvant localiser notre position GPS, discussions sur les blogs, forums. Toutes ces activités sont exposées au danger de violation de la vie privée. De nombreux réseaux sociaux peuvent être divisés en plusieurs catégories et la plupart des réseaux appartiennent à plus d'une catégorie. Voici quelques statistiques **[4]**

• Internet compte 4,2 milliards d'utilisateurs,

• Il y a 3,03 milliards d'utilisateurs actifs de médias sociaux,

• En moyenne, les utilisateurs ont 5,54 comptes de médias sociaux,

• Facebook ajoute 500 000 nouveaux utilisateurs chaque jour ; 6 nouveaux profils chaque seconde,

• Google traite 100 milliards de recherches par mois, 40 000 requêtes de recherche / s

• 300 heures de vidéo sont téléchargées sur YouTube toutes les minutes,

• Plus de 95 millions de photos sont téléchargées chaque jour

Au cas où nous devrions les classer, voici les sept grandes catégories :

• Liens sociaux

Rester en contact avec les amis et les membres de la famille est l'un des plus grands avantages des réseaux sociaux

• Partage multimédia

Les réseaux sociaux facilitent le partage de contenu vidéo et photographique en ligne

• Professionnel

Les réseaux sociaux professionnels sont conçus pour offrir des possibilités de croissance liée à la carrière. Certains de ces types de réseaux constituent un forum général permettant aux professionnels de se connecter, tandis que d'autres sont axés sur des professions ou des intérêts spécifiques.

• Informatif

Les communautés d'information sont composées de personnes qui cherchent des réponses aux problèmes quotidiens. Par exemple, lorsque nous songeons à démarrer un projet de rénovation domiciliaire ou que nous souhaitons apprendre comment passer au vert chez nous, nous pouvons effectuer une recherche sur le Web et découvrir d'innombrables blogs, sites Web et forums remplis de personnes recherchant le même type d'information.

• Éducatif

Les réseaux éducatifs permettent aux étudiants de collaborer avec d'autres étudiants à des projets académiques, d'effectuer des recherches pour l'école ou d'interagir avec des professeurs et des enseignants via des blogs et des forums en classe. Les réseaux sociaux éducatifs sont devenus extrêmement populaires dans le système éducatif actuel.

• Loisirs

L'une des raisons les plus populaires pour lesquelles de nombreuses personnes utilisent Internet, correspondent principalement à des endroits où les utilisateurs partagent leurs hobbies et intérêts

•Académique

Les chercheurs universitaires qui souhaitent partager leurs recherches et examiner les

résultats obtenus par leurs collègues peuvent trouver les réseaux sociaux spécifiques aux universités très utiles.

Un réseau social académique se distingue par les autres réseaux sociaux par le fait qu'il est essentiellement orienté vers la recherche scientifique. Cet outil permet au chercheur de se créer un profil détaillé. Il lui permet également de publier ses résultats de recherche et de suivre le contenu publié par ses collaborateurs

Liste de réseaux sociaux académique :

- ResearchGate[6]
- Academia[7]
- MyScienceWork[8]

1.5. Caractéristiques des réseaux sociaux

Malgré que les sites de réseau social puissent avoir un aspect différent en ce qui concerne leur utilisation, mais leurs propriétés communes consistent en des profils visuels indiquant des informations personnelles détaillées ainsi que de nombreuses autres caractéristiques communes.

Après avoir rejoint un site de réseau social, le nouvel utilisateur doit créer un profil pour se présenter aux autres utilisateurs du réseau. L'utilisateur est tenu de remplir des formulaires avec des informations d'identification telles que le nom, le sexe, la date de naissance, des informations de contact (*adresse / numéros de téléphone*) et des informations de localisation, ainsi que des détails sur les centres d'intérêt, les antécédents scolaires et le travail. Les pages de profil dans les réseaux sociaux jouent le rôle d'identification dans la vie réelle et doivent par conséquent avoir une photo d'identité valide.

Une autre caractéristique commune des réseaux sociaux est la connexion des utilisateurs avec ceux qu'ils connaissent déjà dans la vie réelle ou avec des personnes qu'ils rencontrent

[6] www.ResearchGate.net

[7] www.academia.edu

[8] www.mysciencework.com

dans le réseau social pour la première fois. Amis / abonné / sont tous des termes utilisés pour décrire ces relations qui peuvent aller dans un sens ou dans les deux sens.

N'oubliez pas que le terme « amis » ne signifie pas nécessairement amitié dans la vie numérique, car les gens se connectent avec les autres pour de nombreuses raisons (« *amis du travail* »).

1.6.QUELQUES RÉSEAUX SOCIAUX

Il existe plusieurs réseaux sociaux utilisés dans la vie quotidienne par la plupart des de nous. Certains d'entre eux sont YouTube, Myspace et LinkedIn. Nous discuterons brièvement sur ces réseaux sociaux d'application.

- **LinkedIn**

LinkedIn[9] est un service de réseau social orienté vers les entreprises et l'emploi. Il a été fondé en 2002 et lancé en mai 2003. Il est principalement utilisé pour les réseaux professionnels, notamment les employeurs qui postent des emplois et les demandeurs d'emploi qui affichent leur CV. À partir d'avril En 2017, LinkedIn comptait 500 millions de membres dans 200 pays. LinkedIn permet membres (*travailleurs ou employeurs*) de créer des profils et des « *connexions* » les uns aux autres dans un réseau social en ligne qui peut représenter le monde réel relations professionnelles.

- **YouTube**

YouTube a été fondé par Chad Hurley, Steve Chen et Jawed Karim. YouTube offre aux utilisateurs la possibilité de visionner ses vidéos sur des pages Web extérieures à leur site Web. Chaque vidéo YouTube est accompagnée d'un fichier HTML qui peut être utilisé pour intégrez-le sur n'importe quelle page du Web. Cette fonctionnalité est souvent utilisée pour intégrer Vidéos YouTube dans les pages de réseaux sociaux et les blogs. Utilisateurs souhaitant poster une vidéo discutant, inspirée par ou liée à la vidéo d'un autre utilisateur peut faire une "*réponse vidéo*". À partir de 2013, intégrer, noter, commenter et la publication de la réponse peut être désactivée par le propriétaire de la vidéo sur YouTube.

[9] www.LinkedIn.com

YouTube ne propose généralement pas de lien de téléchargement pour ses vidéos pour qu'ils puissent être visionnés à travers son interface web seulement. Un petit nombre de vidéos peut être téléchargé en tant que fichiers MP4. Nombreux sites Web tiers, applications et les plug-ins de navigateur permettent aux utilisateurs de télécharger des vidéos YouTube.

- **Myspace**

Myspace[10] est un site web de réseautage social fondé aux Etats-Unis en aout 2003, qui met gratuitement à disposition de ses membres enregistrés un espace web personnalisé, permettant de présenter diverses informations personnelles et d'y faire un blog. Il héberge notamment de nombreuses pages internet C'est le plus grand site de réseautage social au monde et se concentre principalement sur musique et culture populaire.

1.6.1.Sites de réseaux sociaux les plus populaires

Les sites de réseaux sociaux les plus populaires sont :

Facebook[11] : Considéré comme synonyme de « *médias sociaux* » par certains, Facbook est le seul site sur lequel nous trouverons probablement des amis, des collègues et des membres de notre famille qui circulent. Bien que Facebook soit principalement axé sur le partage de photos, de liens et de réflexions personnelles, les utilisateurs peuvent également manifester leur soutien aux marques ou aux organisations en devenant fans.

La statistique ci-dessous **[5]** montre un calendrier indiquant le nombre mondial mensuel d'utilisateurs actifs sur Facebook de 2008 à 2018. Au troisième trimestre de 2018, Facebook comptait 2,27 milliards d'utilisateurs actifs par mois. Au troisième trimestre de 2012, le nombre d'utilisateurs actifs de Facebook avait dépassé le milliard, ce qui en faisait le premier réseau social à le faire les utilisateurs actifs sont ceux qui se sont connectés a facebook au cours des 30 derniers jours facebook le réseau social le plus populaire au monde

[10] www.myspace.com

[11] www.facebook.com

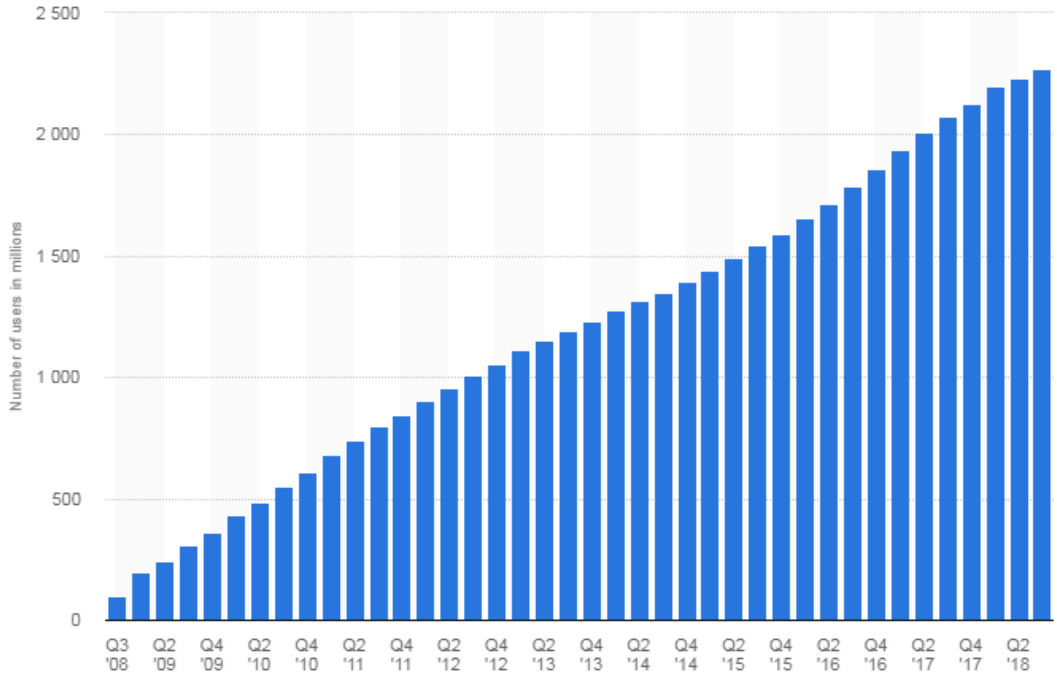


Figure 1.2- nombre de utilisateurs Facebook.

Twitter[12] : Peut-être la plus simple de toutes les plates-formes de médias sociaux, Twitter est également l'une des plus amusantes et des plus intéressantes. Les messages sont limités à 140 caractères ou moins, mais c'est amplement suffisant pour publier un lien, partager une image ou discuter de sujets de pensée avec notre célébrité ou notre influenceur préféré. L'interface de Twitter est facile à apprendre et à utiliser et la configuration d'un nouveau profil ne prend que quelques minutes.

Cette statistique **[5]** montre une chronologie avec le nombre d'utilisateurs actifs mensuels de Twitter dans le monde. Au troisième trimestre de 2018, le service de micro-blogging comptait en moyenne 326 millions d'utilisateurs actifs par mois.

[12] www.twitter.com

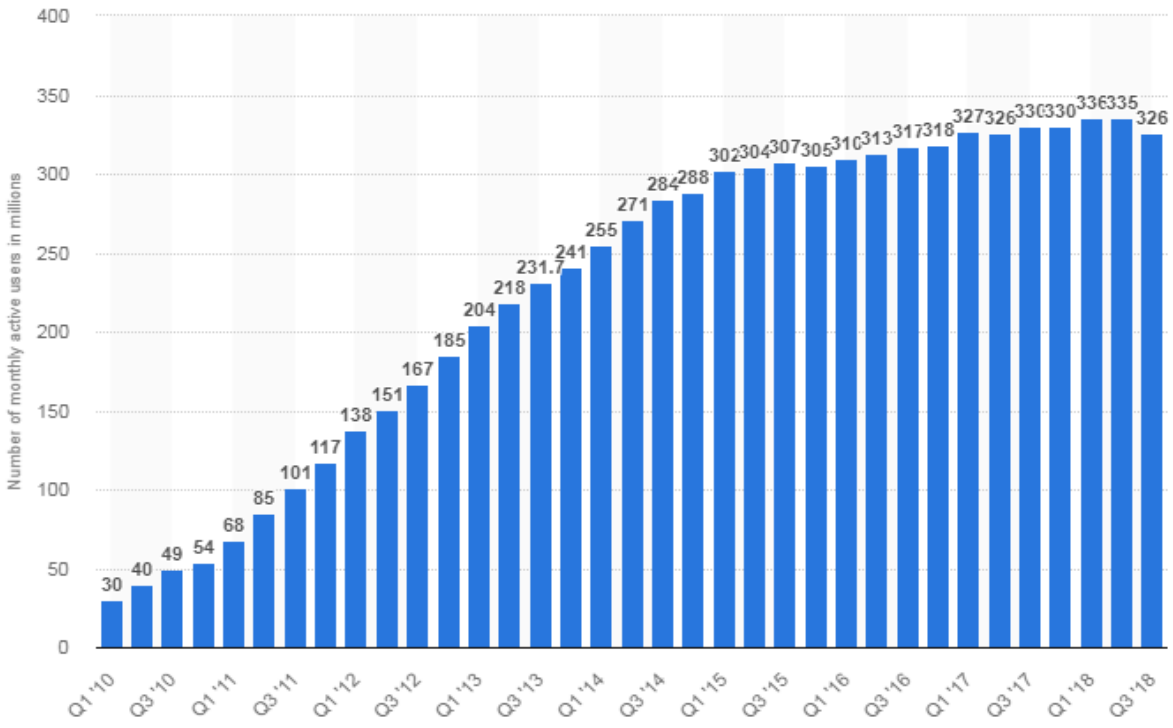


Figure 1.3- Nombre de utilisateurs "Twitter"

1.7.Théorie des graphes et les réseaux sociaux

Le réseau social peut être représenté sous forme de graphe non dirigé G (*V, E*) où V est l'ensemble des sommets représentant les utilisateurs et E l'ensemble d'arêtes représentant des liens sociaux tels que l'amitié, les commentaires et les mentions J'aime. Les graphiques sont utilisés pour représenter les réseaux sociaux et l'analyse de certains de leurs propriétés de base même il peut aider à évaluer et à améliorer la performance des réseaux sociaux. Nous citons quelque définition de base concernant la théorie des graphes pour bien comprendre la Modélisation :

- **Un sommet** : est acteur de base dans un graphe G, Dans un réseau social représente un utilisateur **[6].** Chaque sommet dans un graphe a certaines caractéristiques nommées des attributs représente le nom, Salair, Maladie etc. exemple ci-dessus représente un graphe contient deux sommets avec attributs nom.

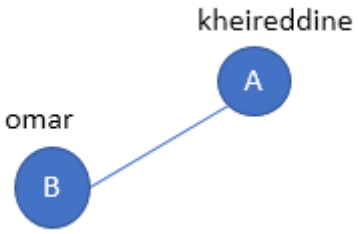


Figure 1.4 Graphe represente deux sommets

- **Degré d'un sommet d'un graphe** : c'est le nombre d'aretes incidentes à ce sommet. Exemple ci-dessus le degré de sommet A est 2

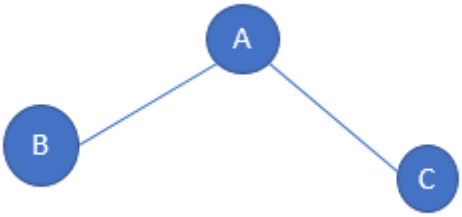


Figure 1.5 exemple de Graphe pour expliquer la degré de sommet

- **L'arête d'un graphe** : c'est Ligne qui joint deux sommets consécutifs, d'un graphe non orienté. Dans les réseaux sociaux l'arête représente des relations entre les utilisateurs par exemple la relation d'amitié.

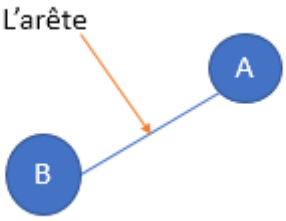


Figure 1.6 l'arête d'un graphe

- **La taille d'un graphe :** c'est le nombre d'arête d'un graphe.
  On va utiliser les graphes non orientés pour représenter un réseau social. Exemple ci-dessous montre un simple réseau social avec 6 sommets représente les utilisateurs avec un attribut sensible (*Salair*).

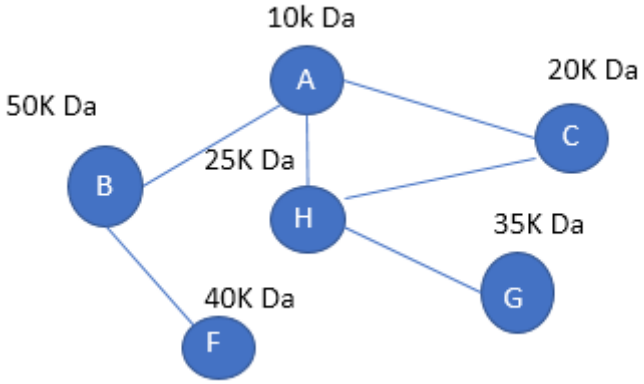


Figure 1.7 une structure simple d'un réseau social

1.8.Sécurité des réseaux sociaux

Les réseaux sociaux font désormais partie de la vie humaine, sont des outils d'échanger des informations telles que des textes, des photos, des messages.
En raison de son utilisation par les internautes de toutes les manières possibles, Les informations des utilisateurs doivent être confidentielles.

1.8.1.Menaces de sécurité dans les réseaux sociaux

Les réseaux sociaux n'échappent pas aux activités ou attaque malveillantes. Comme usurpation d'identité, Malware, les spam, phishing.

- **Phishing**

Le phishing est un type d'attaque dans lequel un attaquant crée des sites falsifiés pour tromper les utilisateurs et collecter des informations sensibles telles que des données de carte de crédit, des noms d'utilisateur, des mots de passe, etc.

Le phishing a une relation avec la psychologie car il s'appuie sur une défaillance humaine plutôt que sur du matériel ou des logiciels, il s'agit d'une sorte d'attaque d'ingénierie sociale. Les attaques de phishing utilisent de faux emails qui convainquent un utilisateur de saisir des informations sensibles sur un faux site Web. Ces messages demandent généralement à l'utilisateur de réinitialiser son mot de passe ou de confirmer les données de sa carte de crédit et de le transférer sur un faux site Web très similaire à l'original. Les principaux types de phishing sont **[7]** :

- **Pharming** : un attaquant falsifié les enregistrements DNS pour rediriger les visiteurs vers un site frauduleux qui est similaire de site officiel.
- **The Watering Hole** : les attaquants catégorisent les utilisateurs et identifient les sites qu'ils visitent. Après Ils analysent ces sites pour trouver des vulnérabilités, si oui il existe, il essaye de saisir des scripts malveillants pour tromper les utilisateurs lors de leur prochaine visite sur ce site.
- **Utilisation des Annonces** : les annonces payées sont une autre méthode de phishing. Ces annonces utilisent des sites malveillants créés par des attaquants, lorsqu'un utilisateur fait un recherche Ces sites peuvent apparaître au-dessus du résultat de la recherche, Ces sites sont souvent utilisés comme moyens de phishing.
- **Clone Phishing** : un attaquant utilise un courrier électronique déjà envoyé et copie son contenu dans un contenu similaire contenant un lien malveillant. L'attaquant peut alors dire qu'il s'agit d'un lien nouveau et que l'ancien lien est expiré.

Figure 1.8 exemple de phishing dans les réseaux sociaux

- **Malwares :**

Le développement de réseaux sociaux, et l'augmentation de nombre d'utilisateurs permet aux réseaux sociaux d'être un refuge pour la distribution des logiciels malveillants. Malware est une version abrégée du terme logiciel malveillant. Il existe plusieurs types de malware, nous citons :

- **Un spayware** est un **logiciel malveillant** qui s'installe dans un ordinateur dans le but de collecter et transférer des informations sur l'ordinateur infecté sans connaissance d'utilisateur.
- **Adware** : un adware est un **logiciel malveillant** utilisé souvent pour des fins publicitaires, il essaye de modifier la page de démarrage des navigateurs Internet pour afficher des liens publicitaires bandeaux, pop-ups.
- **Downloader** : des applications utilisées pour infecter un utilisateur, puis télécharger d'autres logiciels malveillants à partir d'un emplacement déjà prédéfini sans connaissance de la victime.
- **Crimware** : Tout programme informatique utilisé pour faciliter les activités illégales en ligne.

- **Vol d'identité**

Le vol d'identité survient lorsqu'une personne ou une société utilise nos informations personnelles qui peuvent être notre nom, notre adresse e-mail, nos identifiants de connexion de manière illégal. Les réseaux sociaux contiennent les informations personnelles de des millions d'utilisateurs. Donc l'attaquant essaye d'obtenir nos informations personnelles via des applications et des sites (*sites et applications d'éducation ou loisir*).il demande la permission d'accéder aux informations de notre compte avant de pouvoir les utiliser. C'est une façon pour les pirates de voler nos informations

- **Applications tierces ou Third-party applications**

Certains services de réseaux sociaux peuvent nous permettre d'ajouter des applications tierces, y compris des jeux et des quiz, qui offrent des fonctionnalités supplémentaires. Des réseaux sociaux sont des applications Web avancées, car leur utilisation nécessite un niveau action et capacités. Les utilisateurs doivent être prudents en utilisant ces applications, même si une application ne contient pas de code malveillant, il peut accéder aux informations de notre profil sans nos connaissances. Ces informations pourraient ensuite être utilisées de différentes manières, comme la personnalisation de publicités, la réalisation d'études de marché, l'envoi de spam, ou accéder à nos contacts. Les applications Facebook et d'autres réseaux sociaux sont écrits par des tiers-développeurs de parties et ils ont souvent des contrôles de sécurité minimaux

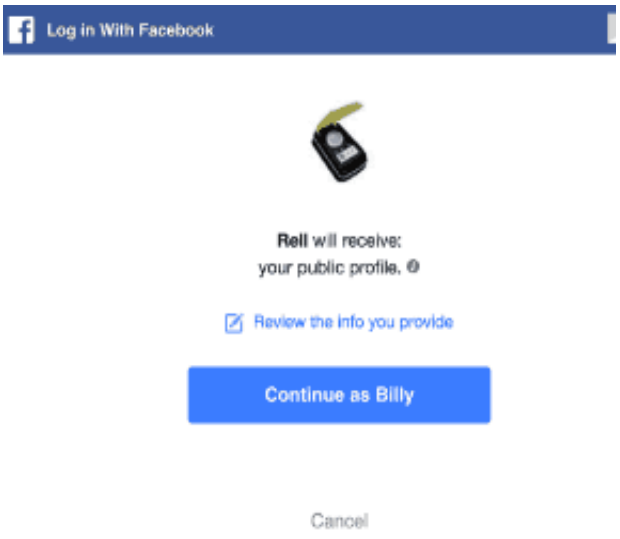


Figure 1.9. exemple d'application tierce

- **Spams**

Le SPAM est l'envoi des messages à un très grand nombre d'utilisateurs dans le but de faire de la publicité, de phishing ou de propager des logiciels malveillants. Il est classé comme une menace dans les réseaux sociaux.

Les réseaux sociaux en ligne essayent de faciliter la procédure d'inscription des utilisateurs pour attirer plus d'utilisateurs. Par exemple, par le processus simple qui demande seulement un nom, une adresse email et un mot de passe. Cette procédure encourage les utilisateurs à s'inscrire plus facilement. Cependant, elle facilite également le processus de création de faux comptes. Par exemple, Facebook a publié une statistique qui révèle environ 83 millions de ses utilisateurs sont des faux utilisateurs. Un faux utilisateur reste sans danger pour les utilisateurs légitimes à moins qu'il ne soit connecté avec eux. Dans ce type d'attaque, l'attaquant crée un faux compte avec des informations d'apparence légitimes comme un faux nom, ville, date de naissance. Ensuite, il tente de devenir ami avec la victime pour avoir ses informations personnelles. Toute personne sur un réseau social peut créer un compte ou une page sous le nom d'une marque, une société en toute illégalité pour des fins illégales.

Figure 1.10 spam

- **Commentaires**

Les commentaires du public peuvent contenir des liens vers des sites frauduleux et comme nous le savons, la plupart des personnels utilisent les réseaux sociaux.

- **URL malveillantes**

Pour partager des informations intéressantes dans les réseaux sociaux, l'URL du site Web est le meilleur moyen de diffuser le contenu. L'utilisateur transmet des informations à ses amis en utilisant les URL de sites Web. Dans les réseaux sociaux les URL malveillantes se propagent sous la forme des publications ou des commentaires. Un adversaire partage des liens sous la forme d'un site Web malveillant en utilisant diverses méthodes et installations telles que bit.ly. Les URL courtes peuvent masquer les URL d'origine et placer le site Web derrière les shorts URL. En conséquence, il peut propager des menaces de spam et de phishing.

1.8.2. Quelques solutions à adopter pour notre sécurité sur les réseaux sociaux

Nous n'invitons pas à cesser d'utiliser les sites des réseaux sociaux, mais nous demandons de prendre les mesures de sécurité nécessaires pour pouvoir utiliser ces sites sans risque.

- **Outil pour la sécurité des URLS**

Pour vérifier si une URL est sécurisée, nous pouvons utiliser des outils de sécurité de sites Web, Ces outils contient des 'URL malveillant et des sites Web non sécurisés. Il suffit de copier-coller n'importe quelle URL dans le champ de saisie de ces outils et vérifier si l'url est sécurisée ou non.

- **Paramètres de confidentialité**

Des nombreux utilisateurs de réseaux sociaux souffrent parfois d'un manque de « *sécurité* », il est donc devenu important de maintenir leur confidentialité.

Voici les 3 paramètres les plus importants que les utilisateurs doivent vérifier sur leurs comptes sociaux pour connaître leur niveau de sécurité :

1 - Changer périodiquement le mot de passe : L'une des choses dont les utilisateurs doivent s'occuper est de changer périodiquement le mot de passe pour différents comptes, et de le remplacer par un mot de passe plus puissant et sophistiqué prenant en compte certains points tels que l'utilisation de lettres, de chiffres et de symboles, afin d'obtenir une protection plus importante.

2- Activer l'authentification en deux étapes : Il est important, en particulier que de nombreux services sociaux tels que Google, Facebook et autres autorisent la fonctionnalité d'authentification en deux étapes, qui offre aux utilisateurs une plus grande sécurité, car personne ne pourra accéder à ce compte, sauf ceux qui ont un téléphone avec un numéro associé Compte.

3- Vérification des applications liées aux comptes : De nombreuses applications et services permettent aux utilisateurs de créer des comptes avec eux via les réseaux sociaux et pour certaines applications nous pouvons être enregistrées via Google ou un compte Facebook, en utilisant les informations personnelles déjà enregistrées sur ces sites. Donc nous pouvons Revenir aux paramètres des réseaux sociaux et supprimer les applications liées.

Figure 1.11 paramétres de confidentialité

- **Réfléchir à deux fois**

Soyons toujours attentif avant de publier toute information nous concernant sur le site de réseautage social, il faut réfléchir deux fois avant de publier des informations personnelles car nous ne savons plus où se trouve le message et qui peut y accéder sur le réseau. Cela pourrait nous hanter à l'avenir.

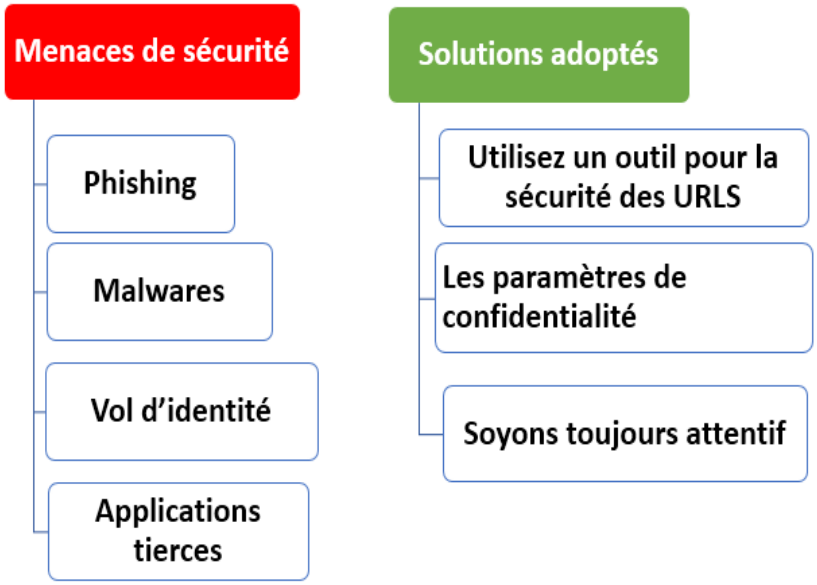


Figure 1.12 la sécurité dans les réseaux sociaux

1.9.Conclusion

Dans nos jours, les réseaux sociaux sont devenus des outils de communication incontournables. Dans ce chapitre, nous avons présenté les réseaux sociaux, nous avons commencé par une introduction sur les réseaux sociaux, ensuite nous avons cité quelques définitions (*web 2.0, réseaux sociaux*) et nous avons terminé par la sécurité dans les réseaux sociaux. Le prochain chapitre se porte sur l'anonymat dans les réseaux sociaux ou on va essayer de faire un état de l'art critique sur les différents problèmes et solutions

Chapitre 2

2.L’anonymat dans les réseaux sociaux

2.1.Introduction

L’anonymat est une notion très importante car il est le garant de notre vie privée. Avec l’essor des réseaux sociaux, l’anonymat est devenu pour certains une nécessité. En effet, les réseaux sociaux sont une véritable mine d’information sur nous. Une simple saisie de notre prénom et notre nom et nous trouverons des photos de nous ou des informations dont nous n’avons peut-être jamais autorisé la diffusion. Cet anonymat est lié aux informations laissées par les utilisateurs sur les réseaux sociaux ou les forums. Effectivement lorsque nous naviguons sur internet nous semons des informations comme notre adresse IP qui se comporte comme un identifiant de notre connexion. Grâce à cette adresse IP, les sites de réseaux sociaux ont la possibilité de savoir si nous sommes déjà rendus sur leurs pages. De plus ils récupèrent diverses informations comme notre IP, notre historique de navigation, les cookies enregistrés, les éventuelles données- saisies dans des formulaires. Les scriptes Intelligents de Facebook peuvent récupérer nos messages privés ou encore nous reconnaître sur certaines photos.

2.2.Le vié privé

Les personnes de tous âges se soucient beaucoup de la vie privée. Et ils se soucient juste autant sur la vie privée en ligne qu’ils le font. Mais ce que la vie privée signifie peut ne pas être ce que nous pensons. Fondamentalement, la vie privée consiste à contrôler la manière dont l'information circule. C'est être capable de comprendre le contexte social afin de se comporter de manière appropriée.

La vie privée est l’impossibilité de connaître les informations privées de l’utilisateur, le respect de la confidentialité des utilisateurs et non-utilisation à des fins illégales.

2.3.Vie privée dans les réseaux sociaux

Ces dernières années, l'utilisation des réseaux sociaux a connu une croissance exponentielle. Un réseau social tels que Facebook, Twitter, LinkedIn, YouTube servent quotidiennement des millions d'utilisateurs. Avec cette utilisation accrue, des nouveaux problèmes de confidentialité ont été soulevés. Les utilisateurs de réseaux sociaux publient des informations personnelles sur eux-mêmes et sur leurs amis. Toutes ces informations peuvent être collectées par une tierce personne. Acquisti et Gross **[8]** dans le domaine de la confidentialité des réseaux sociaux tentent d'évaluer la quantité d'informations personnelles qui ont été révélées par les utilisateurs sur Facebook. Ils ont conclu que de nombreux utilisateurs de Facebook divulguent des informations personnelles les concernant, y compris les dates de naissance, les adresses e-mail, les relations statuts et même des numéros de téléphone. Un autre fait intéressant est qu'environ 55% des utilisateurs acceptent les demandes d'amis des gens qu'ils ne connaissent pas. Les utilisateurs divulguent leur information privée à des étrangers, de plus, des études sur les niveaux de confiance dans les réseaux ont montré que 27,5% des utilisateurs de Facebook qui ont participé à l'étude ont rencontré en personne les personnes qui ils ont rencontrées déja via Facebook **[9].** Boshmaf et al. **[10]** ont réussi à récolter des données sur les utilisateurs en utilisant des demandes d'amis provenant de faux profils Facebook en utilisant Social bot[13].

2.4. Les méthodes de préservation de la vie privée dans les réseaux sociaux

De nombreux site de réseaux sociaux tels que Facebook et Twitter, publiez régulièrement les données lors des activités en ligne de leurs utilisateurs à des tiers, tels que les sociétés commerciales. Ces tiers exploiter ces données sensibles et extraire des informations à des fins spécifiques Ce partage de données suscite les inquiétudes des utilisateurs concernant la divulgation de leur vie privée.il existe des models de la préservation de la vie privée notamment k- l'anonymat et la l-diversité, qui ont été développés à l'origine pour protéger la confidentialité des données afin de sécuriser la confidentialité des utilisateurs dans les réseaux soicaux.

[13] Un social bot est un type particulier de dialogueur utilisé sur les réseaux sociaux afin de générer des messages automatiques

La méthode d'anonymisation vise à rendre l'enregistrement individuel impossible à distinguer d'un enregistrement de groupe en utilisant des techniques de généralisation et de suppression. La croissance rapide des bases de données des réseaux sociaux suscités des préoccupations concernant la protection de la vie privée des personnes.

2.4.1. **Modèle k-anonymat**

Le K-anonymat est un concept clé qui a été introduit pour faire une protection au risque de réidentification des données anonymisées via la liaison à d'autres données. Le modèle de confidentialité de k-anonymat a été proposé pour la première fois en 1998 par Latanya Sweeney **[11]**. Le but de k-Anonymat est que chaque enregistrement ne puisse pas être distingué d'au moins k-1 autres enregistrements et que ces k enregistrements forment une classe d'équivalence. L'idée derrière le k-anonymat est de rendre difficile la liaison d'attributs sensibles et insensibles.
Quasi-identifiant : Un quasi-identifiant (QI) est un ensemble d'attributs dont la sélectivité est telle qu'ils présentent un risque de réidentification, Par exemple {sexe, code postal, date de naissance}
Pour réaliser le k-Anonymat, il y a deux façons principales de le faire :

Généralisation : remplacez les quasi-identifiants par des valeurs moins spécifiques, mais cohérentes sur le plan sémantique, jusqu'à obtenir k valeurs identiques et partitionner les domaines de valeurs ordonnées en intervalles.

Suppression : lorsque la généralisation entraîne une grande perte d'informations, le quasi-identifiant peut être supprimé. Ci-dessous (*tables 2.1 et 2.2*) un exemple d'une table k-Anonyme avec k =2 et les quasi-identificateurs = {*code postale, âge*} et attribut salaire comme donnée sensible.

Code postale	Age	Salaire
1200	28	10K Da
1250	32	40K Da
1450	24	30K Da
1430	26	60K Da

Tableau 2.1 Table Originale

Code postale	Age	Salaire
12**	[28,32]	10K Da
12**	[28,32]	40K Da
14**	[24,26]	60K Da
14**	[24,26]	60K Da

Tableau 2.2 Table - k-anonymat

Donc pour le tableau 2.2 les informations concernant chaque individu contenu dans la table ne peut pas être distinguée d'au moins k-1 autres individus qui apparaissent également dans la table

2.4.2.Points faibles de k-anonymat

On pourrait dire que le k-Anonymat nous offre la protection, mais la vérité est que la méthode est vulnérable à de nombreuses attaques et elle a des inconvénients. Nous citons à titre d'exemple :

- Perte d'information,
- Les enregistrements non triés : le problème est que les enregistrements apparaissent dans le même ordre dans la table publiée que dans la table d'origine. Mais la solution est simple, randomisez l'ordre avant de le libérer
- Attaque d'homogénéité : cette attaque exploite le cas où toutes les valeurs d'une valeur sensible dans un ensemble de k enregistrements sont identiques. Dans de tels cas, même si les données ont été anonymisées, la valeur sensible pour l'ensemble des k enregistrements peut être prédire. Ci-dessous un exemple d'attaque d'homogénéité, si l'attaquant a des connaissances sur la victime que son code postale 1423 et son âge 26, il peut directement déduire que le salaire est 60K Da

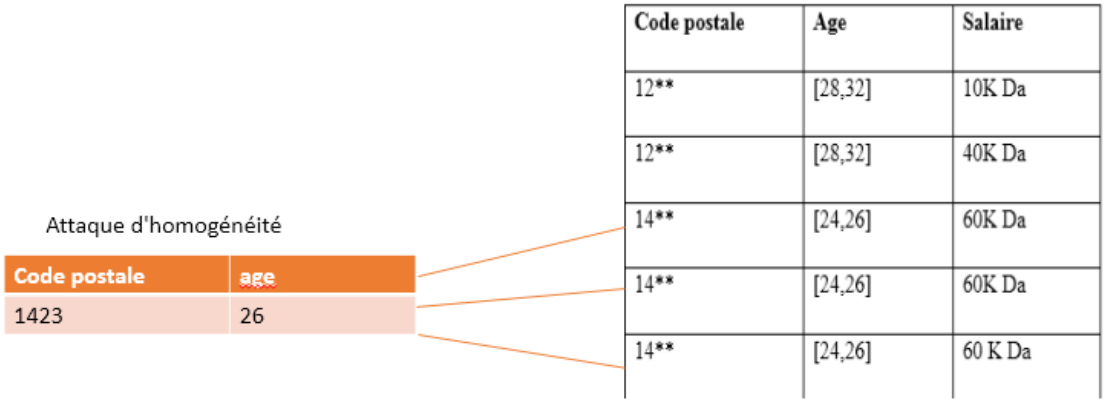


Figure 2.1 Attaque d'homogénéité

2.4.3.Modèle l-diversité

Le modèle de L-diversité **[12]** a été proposé à l'origine pour préserver la confidentialité des données sous forme de tableau. Les chercheurs ont pris conscience de la faiblesse du modèle k-anonymat. Plus précisément, même si le k-anonymat peut garantir que tout objet sera indiscernable à partir d'au moins k-1 autres enregistrements basés sur les valeurs d'attributs de quasi-identifiants, Mais les valeurs d'un attribut sensible pour les objets d'un groupe peuvent être le même. Dans ce cas, bien que l'attaquant ne puisse pas identifier un objet cible, il peut toujours avec succès déduire la valeur de l'attribut sensible. La l-diversité est une extension du modèle k-anonymat mais garanti aussi la diversité pour les valeurs sensibles dans le mécanisme d'anonymisation.

Une définition formelle du modèle l-diversité est la suivante : une classe d'équivalence est dite à l-diversité s'il existe au moins l valeurs distinctes pour l'attribut sensible. Une table est dite à l-diversité si chaque classe d'équivalence de la table a une l-diversité **[12]**.

Code postale	Age	Maladie
12**	[28,32]	Diabète

12**	[28,32]	VH+
12**	[28,32]	Grip
14**	[24,26]	Brûlures gastriques
14**	[24,26]	Cancers de l'estomac

Tableau 2.3 Table anonymisé avec l-diversité

Pour table 2.3 on a chaque classe d'équivalence (groupe) de la table a au moins l valeurs distinctes pour l'attribut sensible

2.4.4.Points faibles de l-diversité

• Le L-diversité est vulnérable de l'attaque par similarité. Par exemple, nous considérons les données de la table 2.3. Un attaquant pourrait supposer que la victime est un patient présent dans la table 2.3 et, il pourrait découvrir que l'âge de victime est compris entre (*24,26*), il peut même savoir qu'il souffre d'une maladie de l'estomac, car le l-diversité ne considère pas les significations sémantiques des données sensibles.

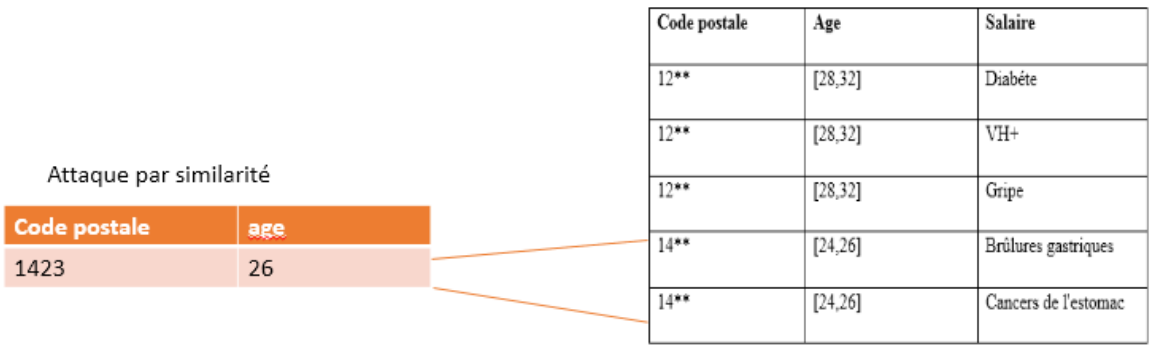


Figure 2.2 l'attaque par similarité

• Perte d'information : à cause de la généralisation qui est utilisé de manière globale

2.4.5.Méthode de R. Mahesh and T. Meyyappan

La méthode de R. Mahesh and T. Meyyappan **[16]** tente d'améliorer des techniques de préservation de la vie privée existantes. Elle adopte une technique de généralisation et préserve l'intimité d'une nouvelle manière.

Des techniques de suppression sont appliquées sur un quasi-identifiant Qi sélectionné. Après le processus de suppression, les enregistrements de la table 2.4 sont triés et organisés en n groupes. Ensuite on va utiliser la technique de généralisation avec une nouvelle manière.

```
Function Anonymize(T)
Array q , w;
Int, nextminimum,nextmaximum,i;
T=functiongroup(T[B])
T=functionsort(T[Q],G);
q=T[Q];
w=T[S];
if(G.count==1)
q0='<='+q0
Else
While (u=0 to G.count)
While(i=0 to T.rowcount in u )
Nextminimum=Findnextminimum(q[i],T[Q]))
Nextminimum=q[i];
Nextmaximum=FindNextmaximum(q[i].T[Q])
q[i]=Nextminimum+"<="+Nextmaximum
End while
End While
T*=Arrange(T[Q],q)
End function
```

Figure 2.3 Pseudo algorithme de généralisation [16]

Exemple illustratif de la méthode : Une table T de données contient des quasi identifiants Qi (i = 1,2… n) et l'attribut sensible S.

Code postale	Age	Salaire (attribut sensible)
4520	23	22k DA
4528	25	28k Da
4523	22	42k Da
4560	30	40k Da
4569	32	60k Da

4562	29	26k Da

Tableau 2.4 Table t originale

Dans un premier temps, la technique de suppression est appliquée sur l'attribut code postale pour transformer l'ensemble de données

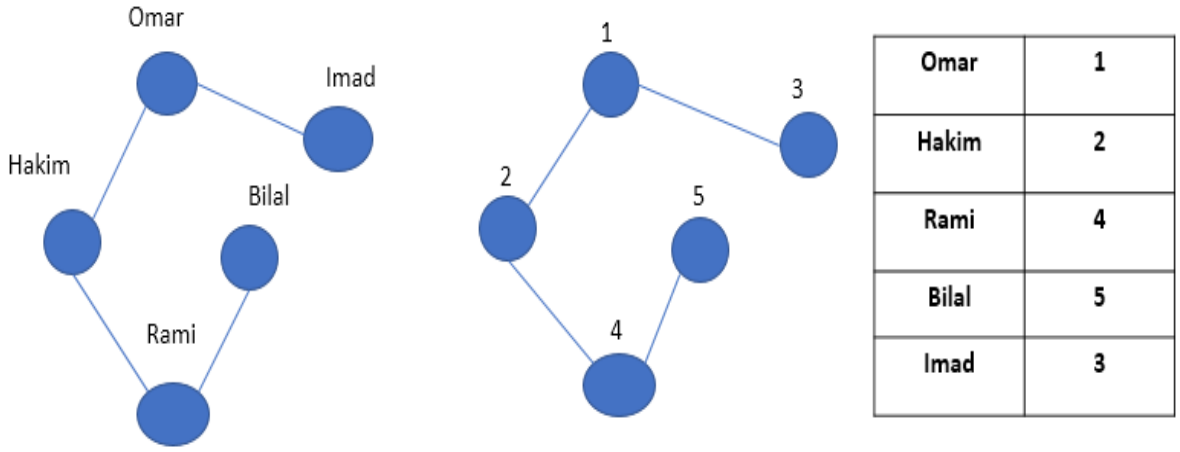
Code postale	Age	Salaire (Attribut sensible)	Groupe
452*	22<=25	22k DA	1
452*	23<=25	28k Da	1
452*	22<=23	42k Da	1
456*	29<=32	40k Da	2
456*	30<=32	60k Da	2
456*	29<=30	26k Da	2

Tableau 2.5 table après utilisation de la méthode

La méthode d'anonymisation proposée est appliquée sur le tableau 2.4 pour transformer l'ensemble de données a des données anonymisé

2.4.6.**Modele d'anonymisation naive**

C'est une façon plus simple d'anonymisation, soit le graphe social suivant avec son graphe anonymisé et la table de correspondance entre les nœuds de chaque graphe



Omar	1
Hakim	2
Rami	4
Bilal	5
Imad	3

Figure 2.4 Anonymisation naive

L'anonymisation naïve d'un réseau social **[13]**, consiste à remplacer les noms des utilisateurs (*c'est-à-dire les étiquettes sur les nœuds*) par des identifiants synthétiques générés aléatoirement afin de préserver la vie privée.

2.4.7.Point faible d'anonymisation naive

- **L'attaque de voisinage**

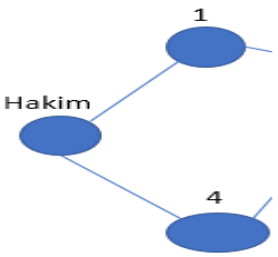


Figure 2.5 Voisinage du nœud

Habituellement, l'adversaire recueille quelques connaissances de base sur victime avant de lancer des attaques, telles que les informations sur les voisins. Une fois que l'adversaire utilise un moyen pour obtenir les informations sur les relations entre le neoud et ses voisins, comme le montre la figure 2.5 (*voisinage du nœud ciblé par exemple, Hakim*), l'adversaire est capable de reconnaître la position du nœud cible dans la topologie anonyme de la figure 2.4 (2) car le graphe de relations de hakim est unique dans la topologie. Il s'agit de l'attaque de voisinage proposée par Zhou **[17]**. Lorsque l'adversaire obtient de nombreux graphiques de voisinage, il peut fusionner ces graphiques pour déduire une partie de la topologie sociale. Il

peut utiliser ces informations pour lancer d'autres attaques. Pour éviter cette attaque, l'ajout des liens dans la topologie sociale est l'un des moyens possibles.

2.4.8.Model k-degee en utilisant la théoré de graphe

Liu et Terzi **[14]** sont les premiers qui as utilisé le model k-degee pour garder l'identité des utilisateurs privée dans l'attaquant utilise la connaissance préalable du degré d'un sommet de la victime pour appliquer l'attaques de ré-identification. Dans cet exemple, ajoutez un lient entre H et B et entre C et F peut préservée la divulgation d'identité, car Tous les nœuds ont le même degré dans le graphe.

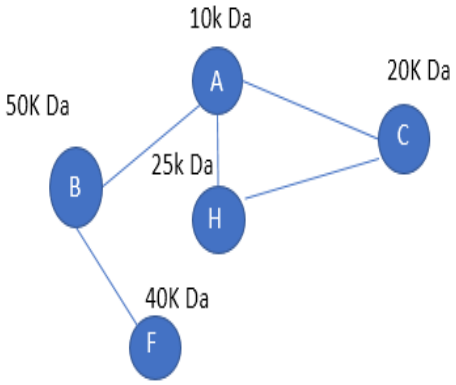


Figure 2.6 Graphe originale

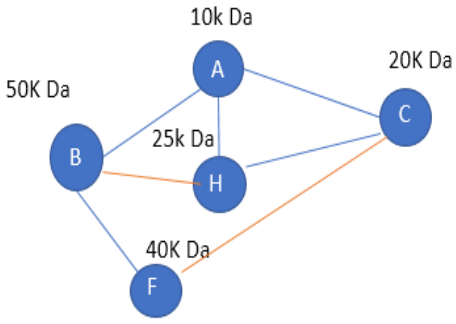


Figure 2.7 Graphe 3-degee

Un graphe G (*V, E*) est anonyme à k-degrés si chaque sommet a le même degré avec au moins k-1 autres sommets. Cette méthode basée sur la modéfication de graphe en ajoutant et en supprimant des arêtes du graphe.

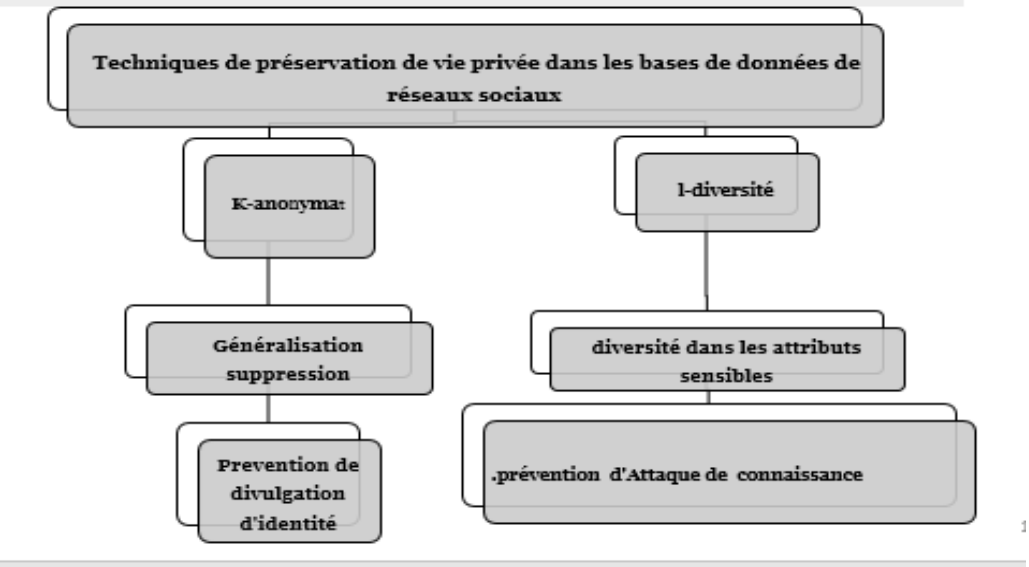


Figure 2.8 les méthodes de préservations des données dans les réseaux sociaux

La Figure 2.8 représente les déferent méthodes de préservation de vie privée dans réseaux sociaux, le principe de fonctionment et les avantages de ces méthodes

2.5.Travaux connexes

Les méthodes existantes trouvent une solution au problème de confidentialité à un certain dégrée. K-anonymat peut empêcher l'attaque par divulgation d'identité mais pas l'attaque de divulgation d'attribut. Une autre méthode l-diversité préserver la confidentialité contre l'attaque par divulgation d'attribut mais, elle est faible en cas ou l'attaque par similarité.

Qian Wang **[15]** a proposé un modèle pour combler la pénurie de k-anonymat dans la protection de la divulgation des attributs. Ça peut empêcher la divulgation d'attributs en contrôlent la déférence des valeurs d'attributs sensible.

Mahesh, Meyyappan **[16]** a proposé une nouvelle méthode pour anonymiser le jeu de données en définissant des valeurs d'intervalle dans les Quasi-identifiants. Si le Quasi-identifiant est constitué du même attribut dans toutes les classes, cette méthode ne permet pas préserver la confidentialité privée. La méthode proposée dans le chapitre suivant résout ce problème et assure la confidentialité privée.

2.6.L’avantage de l’anonymat dans les reseaux sociaux

Tout n’est pas authentique en ligne. Juste parce que nous n’avons pas envie de révéler tout ce qui se passe actuellement dans notre vie, nous ne sommes pas obligés de donner tous ces détails. Les personnes réagissent différemment en fonction des rôles qui leur sont assignés. Nous sommes différents au travail, en famille, avec nos amis. Donc il est normal de vouloir garder l’anonymat en ligne et de créer des profiles différents sur Twitter, Facebook, YouTube et sites de blog en montrant différentes facettes de nous, dans divers secteurs.

Rester anonyme aide les personnes à s’exprimer librement sur les blogs, sans craindre leurs parents, leurs superviseurs et leur tuteur. Cela aide les gens à avoir la liberté d’expression sur les sujets relatifs à la politique, à la religion, aux minorités et autres sujets à controverse. Nous pouvons nous caricaturer comme nous le souhaitons, créer un nouveau personnage, traiter de n’importe quel sujet qui nous intéresse.

Les cours en ligne ou les salles de classe en ligne, aident à avoir plus confiance en soi, car les étudiants n’ont pas besoin d’établir des contacts personnels et l’anonymat les préserve des problèmes liés à l’apparence et des autres malaises liés à la peur de sentir rejeté. L’anonymat aide à rester naturel et à faire des remarques et des critiques honnêtes.

Une personne peut rejoindre un groupe spécial en ligne, partager son travail artistique et même aider d’autres personnes à cacher leur véritable identité. L’anonymat protégé contre les hackers et les criminels qui tentent de voler notre réelle identité, nos données vitales et personnelles et nous garantissons notre sécurité personnelle et notre bien-être.

2.7.Conclusion

La sécurité et la confidentialité des utilisateurs et leurs données sont les principaux problèmes liés aux réseaux sociaux. Ces problèmes peuvent survenir auprès de fournisseurs de services de réseaux sociaux,dans ce chapitre nous avons présenté l’anonymat dans les réseaux sociaux nous avons commencé par une introduction sur anonymat,la vie privée ensuite nous avons cité les méthodes de préservation de la vie privée dans les réseaux sociaux et nous avons terminé par L’avantage de l’anonymat dans les réseaux sociaux Dans le prochain chapitre, nous allons présenter notre nouvelle approche d’anonymisation que nous proposons pour proteger les données dans les réseaux sociaux.

Chapitre 3

3. Nouvelle méthode de la protection de la vie privée dans les réseaux sociaux

3.1.Introduction

Beaucoup de réseaux sociaux en ligne comme Facebook, LinkedIn, Google, et Twitter offrent aux internautes des nouveaux moyens intéressants pour communiquer et interagir. Au cours des dernières années, les réseaux sociaux en ligne ont connu une croissance exponentielle dans le nombre de leurs utilisateurs et dans l'énorme quantité d'informations disponibles. Les informations disponibles sur ces réseaux sociaux décrivent généralement les personnes, leurs informations personnelles (*ex. salaires, maladies, adresse, ...*) et les interactions (*ex. avec qui ils échangent des messages, quels commentaires qu'ils publient, ...*). Dans le domaine de la santé, on a par exemple le réseau social PatientsLikeMe[14], Un réseau social avec plus de 150 000 utilisateurs en juillet 2012. Dans ce réseau, chaque patient puisse échanger des informations sur sa maladie, ses causes et comment la traiter. Par conséquent, il est important de protéger la grande quantité d'informations contre tous les types d'attaques qui menace la confidentialité des utilisateurs, compromettre leur sécurité et exposer leurs données à des tiers non autorisés. Des mécanismes et des méthodes doivent être fournis pour protéger les informations personnelles et la vie privée des utilisateurs dans les réseaux sociaux.

Tous les réseaux sociaux (ex. *facebook*, *twitter*, …) conservent de gros volumes de données. Lorsque ces réseaux sociaux publient des données, elles contiennent beaucoup d'informations sensibles, de sorte qu'ils aimeraient préserver la vie privée des individus. Pour protéger la confidentialité des données individuelles, le fournisseur de données supprime les attributs clés

[14] http://www.patientslikeme.com

identifiés tels que le nom, adresse, ID, etc. Cependant, malgré la suppression des attributs de clé, il n'existe aucune garantie de l'anonymat. L'information qui est publié contient souvent d'autres données appelées *quasi-identificateurs* (ou *QI*) tels que la date de naissance, le sexe et le code postal, qui peut être lié à des informations accessibles au public pour identifier la personne, laissant ainsi filtrer des informations non destinées à la divulgation.

Dans ce chapitre, nous allons proposer une méthode de protection de la vie privée dans les réseaux sociaux avec une perte d'information minimale.

3.2.Méthode proposée

La méthode proposée fournit une nouvelle technique d'anonymisation comprenant l'élimination des enregistrements et la généralisation et catégorisation.

3.2.1. Définition

Soit T *(A_1, A_2, A_3, ….A_n)* un tableau de donnée. A_1,A_2,A_3…An sont des ensembles finis d'attributs A dans la table T. Chaque tuple représente les informations d'une personne ou un objet. Il existe deux types d'attributs : les attributs Quasi-identificateurs (Q) et les attributs sensibles (S) . Les Quasi-indentifiants pourraient être connus par les adversaires. Les adversaires peuvent trouver les informations sensibles individuelles des Quasi-identifiants. Par conséquent, les attributs sensibles doivent être protégés. Soit QT est un Quasi-identifiant de la table T où $\{Q_i \dots Q_j\} \subset \{A_1 \dots . A_n\}$.

Soit T^* la table anonymisée et QT_{T^*} est les quasi-identifiants qui lui est associé.

Code postale	Age	Sex	Maladé
13056	28	H	l'hépatite A
13084	30	H	Cancer
14823	25	F	Diabet
25023	40	H	l'hépatite A
25069	42	F	Maladie cardiaque

14845	25	F	Diabet
13024	32	H	Cancer
14869	43	F	Fièvre
25073	44	F	Diabet
60220	48	F	Grippe
13020	30	H	Maladie cardiaque
60253	51	H	Fièvre
13034	31	F	Diabet
14828	37	F	Grippe
13047	30	H	Cancer
13015	30	H	Cancer

Tableau 3.1 table des donneés

Soit le tableau 3.1 noté T qui représente des données médicales contenant des quasi-identifiants {*code postale, âge, sexe, maladie*} $Q_i\{i = 1,2,3, \dots n\}$ et S l'attribut sensible.

3.2.2. Catégorisation

Catégoriser les valeurs des attributs sensibles en deux classes :

- **High sensitive class :** Soit A un ensemble des attributs sensibles. H= {S1, S2, S3 …Sn} est l'ensemble des attributs qui sont très sensibles par exemple le cancer.
- **Low sensitive class :** Soit A un ensemble des attributs sensibles. L= *{S1, S2, ..Sn}* est l'ensemble des attributs qui sont peu sensibles par exemple la grippe.

Basant sur un dictionnaire de données D qui contient les maladies plus sensibles, dans l'exemple suivant D= {*cancer, diabète, maladie cardiaque*}. On divise le tableau T en deux tableaux distincts (T1, T2). Le tableau T1 contient les attributs plus sensibles *(H),* et T2 contient les attributs peu sensibles *(L).*

Code postale	Age	Sex	Maladie
13084	30	H	Cancer
13024	32	H	Cancer
13047	30	H	Cancer
13020	30	H	Maladie cardiaque
13034	31	F	Diabete
13015	30	H	Cancer
14823	25	F	Diabete
14845	25	F	Diabete
25069	42	F	Maladie cardiaque
25073	44	F	Diabete

Tableau 3.2 :T1 Hight sensitive class

Code postale	Age	Sex	Maladie
13056	28	H	L'hépatite A
25023	40	H	L'hépatite A
14869	43	F	Fièvre
60220	48	F	Grippe
14828	37	F	Gripe
60253	51	H	Fièvre

Tableau 3.3: T2 low sensitive class

3.2.3. **Anonymisation et généralisation**

L'anonymisation est effectuée uniquement sur les attributs sensibles appartenant à la classe H. La généralisation est utilisée pour effectuer l'anonymisation, les données sont

généralisées en construisant le domaine hiérarchie de généralisation (*DGH*). Pour les quasi-identifiants par exemple, la DGH des codes postaux sont illustrés surla Figure 3.1.

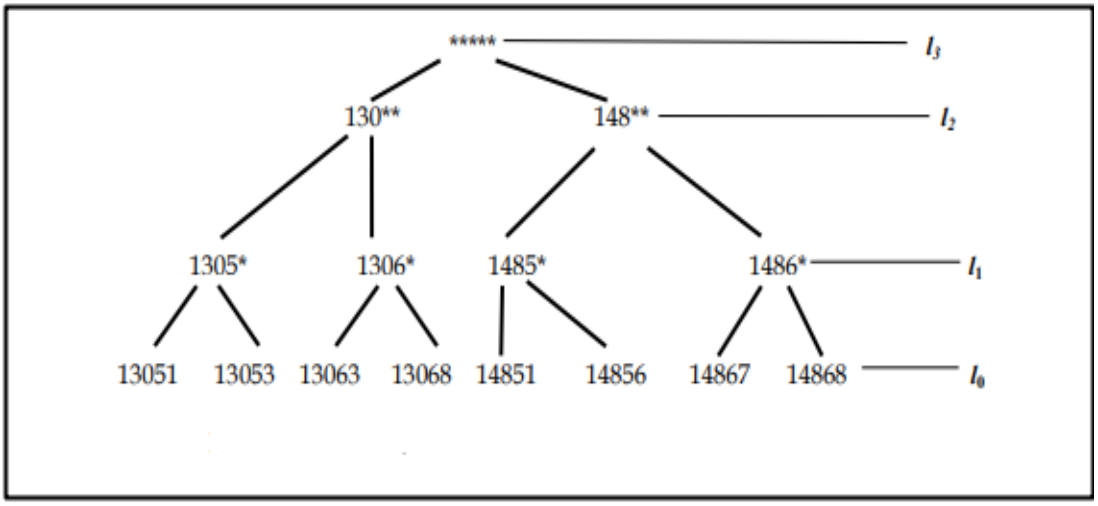


Figure 3.1: DGH des codes postaux

- Les données sont groupées en utilisant la DGH

Code postale	Age	Sex	Maladie
130**	30	H	Cancer
130**	32	H	Cancer
130**	30	H	Cancer
130**	30	H	Maladie cardiaque
130**	31	F	Diabete
130**	30	H	Cancer
148**	25	F	Diabete
148**	25	F	Diabete
250**	42	F	Maladie cardiaque
250**	44	F	Diabete

Tableau 3.4: T1 après la généralisation

- Après la généralisation, s'il existe au moins un enregistrement qui a la même valeur d'attribut sensible et le même quasi-identifiant dans le même groupe, on conserve un seul enregistrement et on élimine les autres. Donc on obtient le tableau suivant :

Code postale	Age	Sex	Maladié
130**	30	H	Cancer
130**	32	H	Cancer
130**	30	H	Maladie cardiaque
130**	31	F	Diabete
148**	25	F	Diabete
250**	42	F	Maladie cardiaque
250**	44	F	Diabete

Tableau 3.5:T1 après la suppression

Le tableau ci-dessous contient les éléments supprimés

130**	30	H	Cancer
130**	30	H	Cancer
148**	25	F	Diabete

Tableau 3.6: exemple du suppression

- Le processus de généralisation suit le processus d'élimination, dans chaque groupe Gi, la prochaine valeur entière minimale L_i, et ensuite la plus grande valeur entière M_i est trouvée pour chaque valeur d'attribut. Les valeurs d'attribut du quasi-identifiant Qi dans le groupe Gi sont réécrites comme une valeur de plage $L_i <= M_i$. Ce processus est répété jusqu'à ce que toutes les valeurs Qi dans chaque groupe Gi sont supprimées jusqu'à l'obtention d'un tableau généralisé (*voir table 3.7)*

Code postal	Age	Sex	Maladie
130**	30<=31	H	Cancer
130**	31<=32	H	Cancer
130**	30<=31	H	Maladie cardiaque
130**	30<=32	F	Diabete
148**	24<=26	F	Diabete
250**	42<=44	F	Maladie cardiaque
250**	42<=44	F	Diabete

Tableau 3.7:T1 Généralisé

- Output : T^*= T1 + T2 : l'étape suivante consiste à regrouper le tableau généralisé T1 avec le tableau qui contient les données non sensible T2 avec un mélange aléatoire.

Code postal	Age	Sex	Maladié
130**	30<=31	H	Cancer
130**	30<=31	H	Maladie cardiaque
130**	30<=32	F	Diabete
148**	24<=26	F	Diabete
250**	42<=44	F	Maladie cardiaque
13056	28	H	L'hépatite A
25023	40	H	L'hépatite A
130**	31<=32	H	Cancer
14869	43	F	Fièvre

60220	48	F	Gripe
14828	37	F	Gripe
250**	42<=44	F	Diabete
60253	51	H	Fièvre

Tableau 3.8: T* table anonymisée

L'algorithme de la méthode proposée est décrit ci-dessous. Une table de base de données non anonymisée T avec l'attribut sensible S et le Quasi identifiant Q est entré dans l'algorithme. Une opération de généralisation est effectuée pour convertir la table T en T *. Jeu de données dans une table T avec n nombre de tuples, n nombre de quasi-identificateurs et l'attribut sensible S sont choisis.

- **Entrée** : Table T avec chaque tuple contenant quasi-identifiant (Q), attribut sensible S, où i = 1,2, 3…n.
- **Etape 1** : Catégoriser les valeurs d'attributs sensibles en deux classes, plus sensible H et peu sensible L. Pour chaque tuple dont la valeur sensible appartient à la classe H, c'est-à-dire si $t[S] \subset H$: Déplacez ces tuples dans la table T1 pour appliquer la généralisation. Pour chaque tuple dont la valeur sensible appartient à la classe L, c'est-à-dire si $t[S] \subset L$: Déplacez ces tuples dans la table T2.
- **Etape 2** : Grouper les tuples de T1 par rapport à une valeur numérique.
- **Etape 3** : La suppression : éliminer les enregistrements identiques dans chaque groupe Gi.
- **Etape 4** : Appliquez la généralisation sur des attributs quasi-identidiant pour les anonymiser.
    1. Répétez les sous-étapes de 2 à 4 jusqu'à trouver le minimum et le maximum dans chaque groupe
    2. Soit $L_j = t_j$, trouver l'entier minimal le plus proche suivant $L_i \; de \; t_i$ dans le groupe Gi ou i = 1,2, …. N, et Si trouvé, $L_i = t_j$.
    3. Soit $M_j = t_j$, trouver l'entier maximal le plus proche $M_i \; de \; t_i$ de groupe Gi ou i = 1,2, …N, et Si trouvé, $M_i = t_j$.

4. Les valeurs de tuple t_i du quasi-identifiant Q_i dans le méme groupe G_i sont réécrit sous forme de plage $L_j <= M_j$.

- **Etape 5** : T* = T1 + T2.

3.3.Conclusion

La protection de la vie privée est un domaine de recherche très actif. La publication de données (*microdonnées*) de tout individu sans révéler leur vie privée pose un problème très important. Plusieurs réseaux sociaux publient des données mais ils souhaitent préserver les données sensibles avant d'afficher. Il existe plusieurs Techniques qui ont été proposé pour protéger la vie privée, mais elles considèrent toutes les attributs sensibles au même niveau et appliquent la généralisation à tous. Donc plusieurs problèmes sont posés :

• Perte d'informations.

• Mesure de confidentialité.

Par conséquent, une amélioration supplémentaire est apportée au système d'anonymat existant, qui assure la confidentialité des données. Perte d'information minimale et maximum de données utilitaire. Ce chapitre a présenté une nouvelle approche basée sur la catégorisation des valeurs d'attributs sensibles dans différentes classes, seuls les tuples les plus sensibles sont anonymisés. La confidentialité d'individu est également préservée. Le chapitre suivant sera consacré à l'évaluation des performances de la solution proposée.

Chapitre 4

4.Evaluation des Performances

4.1.Introduction

Après avoir décrit les fondements théoriques de notre proposition, nous abordons l'aspect simulation afin de mettre en œuvre cette solution d'anonymisation, L'objectif étant d'évaluer notre approche de préservation des données. Nous évaluons notre algorithme proposé en comparant les résultats de l'algorithme avec la solution de R. Mahesh and Dr.T. Meyyappan **[16]** dont nous sommes inspirés, et que nous l'avons aussi implémenté.

4.2.Mesures de qualité

4.2.1.Perte d'information

La perte d'informations causée par les solutions de l'anonymisation peut être mesurée par les tuples généralisés qui se rapprochent de l'original. Après la généralisation, certaines valeurs d'attributs d'un tuple sont généralisées à un intervalle pour mesurer l'utilité des attributs dans l'anonymisation. La pénalité normalisée de la certitude ou en anglais « *the normalized certainty penalty* » ou NCP pour un attribut numérique c'est la taille de son intervalle de normalisation après la généralisation et pour une valeur catégorique, Le NCP mesure son nombre normalisé de descendants dans l'arbre de la hiérarchie après généralisation. La pénalité normalisée de la certitude devrait être minimisée. Pour l'attribut numérique, considérons une table T avec quasi identifiant (A_1 …… .A_n). Supposons qu'un tuple t = (x_1 … .x_n) soit généralisé à un tuple t′=([y_1, z_1], … [y_n, z_n]) tel que $y_i \leq x_i \leq z_i$ ($1 \leq i \leq n$). Ensuite, nous définissons la pénalité normalisée de la certitude (NCP) du tuple t sur un attribut Ai, tel que :

$$NCP_{A_i}(t) = \frac{z_i - y_i}{|A_i|}$$

Équation 4.1 NCP des données numériques

$$\text{Où} \quad |\text{Ai}| = \max \text{t.Ai} - \min \text{t.Ai}$$

L'arbre hiérarchique est utilisé pour la généralisation en attribut hiérarchique. Supposons qu'un tuple t ait la valeur v on attribut catégorique a_i, qu'il soit généralisé a un ensemble de valeurs v_1, …… .v_m c'est-à-dire le nombre de nœuds feuilles qui sont les descendants de l'ancêtre *(v_1, ……, v_m)* sont utilisés pour mesurer la généralisation en attribut catégorique. Il est défini par :

$$NCP_{A_i}(t) = \frac{|\text{descendants}(\text{v1}, \ldots\ldots, \text{vm})|}{|A_i|}$$

Équation 4.2 NCP des données non numérique

Où $|A_i|$: est le nombre de valeurs distinctes sur l'attribut A_i

4.2.2. **Discernement**

Discernement ou en anglais « *Déscernibility* » est la mesure de la perte d'information basée sur le nombre des enregistrements impossibles à distinguer les uns des autres, également connu sous le nom d'équivalence des classes. L'idée à la base de la mesure de pénalité de discernement *PD* est que plus les enregistrements sont impossibles à distinguer les uns des autres, plus la perte d'informations est importante. C'est parce que plus de généralisation est nécessaire pour rendre les enregistrements impossibles à distinguer les uns des autres. Par conséquent, l'algorithme idéal devrait réduire la pénalité de discernement *PD* en réduisant la taille des classes d'équivalence. Plus la taille des classes d'équivalence entraîne un PD moins élevé on a une perte d'informations moindre.

$$\text{PD} = \sum_{\text{equivalance class}} \text{E}^2$$

Équation 4.3 Discernibility

4.2.3.Temps d’exécutions

Un aspect extrêmement important d'un algorithme, à savoir sa performance. Il est donc important pour un informaticien de choisir l'algorithme le plus rapide. Nous avons donc envisagé le temps d’exécution pour évaluer et comparer la rapidité de notre approche avec l’approche de R. Mahesh et Dr.T. Meyyappan.

4.3. Tests et Expérimentations

Nous présentons et nous justifions tout d'abord les outils utilisés.

4.3.1.Outils utilisés

- Windows 7 comme système d’exploitation
- Langage JAVA (*JDK 1.8*) comme langage de programmation

Langage de programmation « JAVA »

Java est un langage de programmation orienté objet et un environnement d'exécution, développé par Sun Microsystems. Il fut présenté officiellement en 1995. Le Java était à la base un langage pour Internet, pour pouvoir rendre plus dynamiques les pages (*tout comme le JavaScript aujourd'hui*). Mais le Java a beaucoup évolué et est devenu un langage de programmation très puissant permettant de presque tout faire.

Les avantages de java sont:

- Langage puissant,
- Portabilité excellente,
- Langage orienté objet,
- Langage de haut niveau,
- JDK très riche,
- Nombreuses librairies tierces,
- Très grande productivité,
- Applications plus sûres et stables,
- Nombreuses implémentations,

- IDE de très bonne qualité et libres : Eclipse par exemple.

```
package me;
import java.io.Console;

public class main {

    public static void main(String[] args) throws FileNotFoundException {
        // TODO Auto-generated method stub
        Scanner s = new Scanner(new File("C:/Users/maman/Desktop/j.txt"));
        ArrayList<ArrayList<String>> list = new ArrayList<ArrayList<String>>();

        while (s.hasNextLine()){
            String [] listtable=s.nextLine().split(",");
            ArrayList<String> l=new ArrayList<>();
            Collections.addAll(l,listtable);
            list.add(l);
        }
        s.close();
        System.out.println(list.size());
        /*..... ancien methode */
        System.out.println("ancien methode");
        catégorisation c= new catégorisation();
        ArrayList<ArrayList<String>> ar = c.all(list);
            c.trié(ar);
            c.groupe(ar,3);
        généralisation g=new généralisation ();
```

Figure 4.1 Portion de code source de la méthode

```
public ArrayList<ArrayList<String>> generalisé(ArrayList<ArrayList<String>> g){
for (ArrayList<String> arrayList : g) {
    arrayList.set(1,arrayList.get(1).substring(0, arrayList.get(1).length() - 3) + "***");

    int p=0;
    int age= Integer.parseInt(arrayList.get(0));
    ArrayList<Integer> min = new ArrayList<>();
    ArrayList<Integer> max = new ArrayList<>();
    ArrayList<Integer> egal = new ArrayList<>();

      while(p < g.size()){
          if(arrayList.get(4)== g.get(p).get(4)){
              if(Integer.parseInt(arrayList.get(0)) >  Integer.parseInt(g.get(p).get(0))){
                 min.add(Integer.parseInt(g.get(p).get(0)));
              }
              else if(Integer.parseInt(arrayList.get(0)) <  Integer.parseInt(g.get(p).get(0))){
                  max.add(Integer.parseInt(g.get(p).get(0)));

              }
              else{
                  egal.add(Integer.parseInt(g.get(p).get(0)));
```

Figure 4.2 portion 2 de code de la méthode

4.4.Résultats

Un ensemble de données a été utilisé **[18]**. Les données accessibles au public pour vérifier les performances.

```
27,Private,174419,Bachelors,13,Never-married,Prof-specialty,Not-in-family,White,
21,Private,157916,7th-8th,4,Never-married,Machine-op-inspct,Own-child,White,Male
46,Private,283384,HS-grad,9,Married-civ-spouse,Prof-specialty,Husband,White,Male
45,Private,214223,Bachelors,13,Married-civ-spouse,Sales,Husband,White,Male,0,190
41,Self-emp-not-inc,264006,Assoc-voc,11,Married-civ-spouse,Sales,Not-in-family,W
20,Private,166313,11th,7,Never-married,Tech-support,Other-relative,White,Male,0,
28,Private,263614,Bachelors,13,Married-civ-spouse,Exec-managerial,Husband,White,
33,Private,196128,11th,7,Never-married,Craft-repair,Own-child,White,Male,0,0,40,
```

Figure 4.3 Exemple de données avant le traitement

Notre méthode est implémentée avec Java et est exécutée sur un processeur Intel Core i3 de 1,8 GHz avec 4 Go de RAM. Un ensemble de données public qui est ensemble de données standard pour vérifier la performance de l'algorithme d'anonymat. Le jeu de données contient 30061 lignes. Nous ne considérons que 1100 tuples à des fins expérimentales. Cet ensemble de données contient 11 attributs et nous ne retenons que 4 attributs qui sont l'âge, le code, le sexe et l'occupation. L'âge, le code et le sexe sont considérés comme quasi-identifiant et occupation est considérée comme un attribut sensible. Parmi tous les attributs sensibles « *Tech-support* » et « *Sales* » sont considérées comme des valeurs les plus sensibles qui doivent être protégées. Ces valeurs comprennent les classes très sensibles et autres valeurs incluses dans la classe de basse sensibilité.

La méthode proposée est comparée avec la méthode de Mahesh R, Meyyappan **[16]**. La comparaison est faite sur la base de la perte d'information et la discernibilité ainsi que le temps d'exécution

Les figures 4.4 et 4.5 présentent la perte d'information par rapport à la taille de base de donnée et par rapport au nombre de groupe. La méthode proposée montre une amélioration significative en réduisant la perte d'information par rapport à l'approche comparée car, on à pas utiliser la généralisation de manière globale

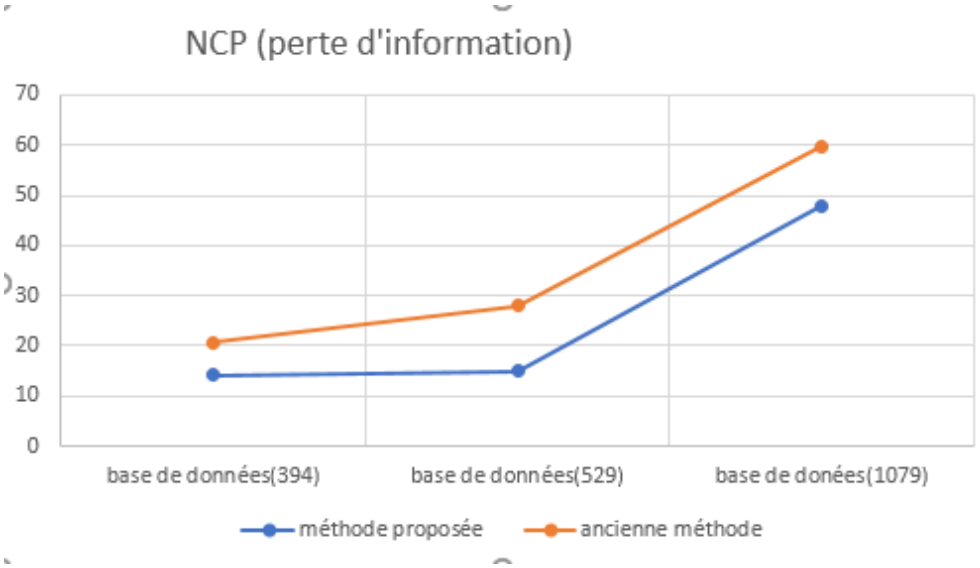


Figure 4.4 NCP méthode proposé et ancienne méthode

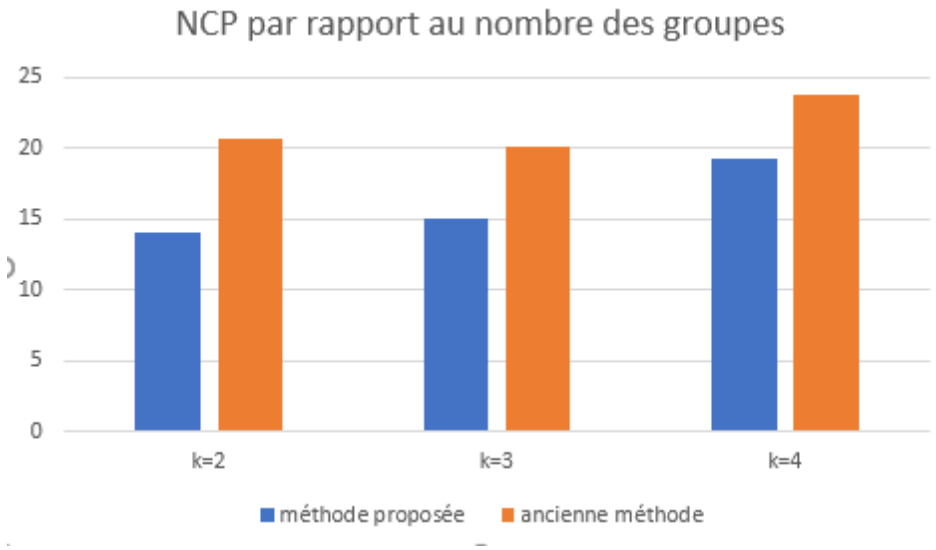


Figure 4.5 NCP par rapport au nombre des groupes

La figure 4.6 montre que la métrique de discernabilité (*DM*) par rapport à la taille de la base de données défère à la fois de la méthode ancienne et de la méthode proposée. L'algorithme idéal devrait réduire le discernement en réduisant la taille des classes d'équivalence. Donc la discernabilité devrait être minimisée. La méthode proposée montre une amélioration significative en réduisant la perte d'information par rapport à approche comparée.

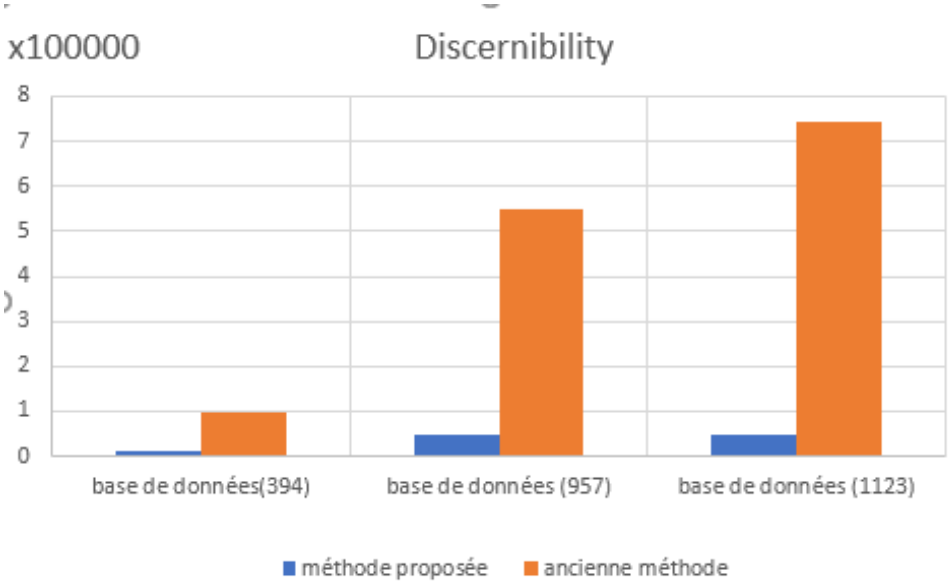


Figure 4.6 Discernibility

Pour figures 4.7, et 4.8, nous pouvons conclure que notre approche prend moins de temps d'exécution que l'approche de R. Mahesh et Dr.T. Meyyappan en fonction de la taille de la base de données et nombre des groupes.

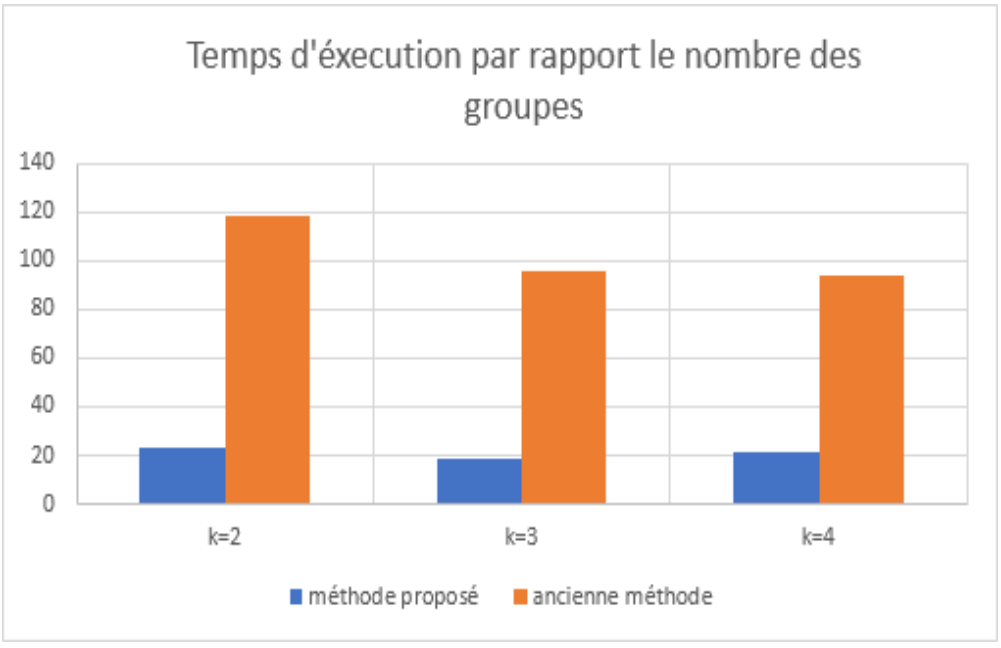


Figure 4.7 Temps d'exécution par rapport au nombre des groupes

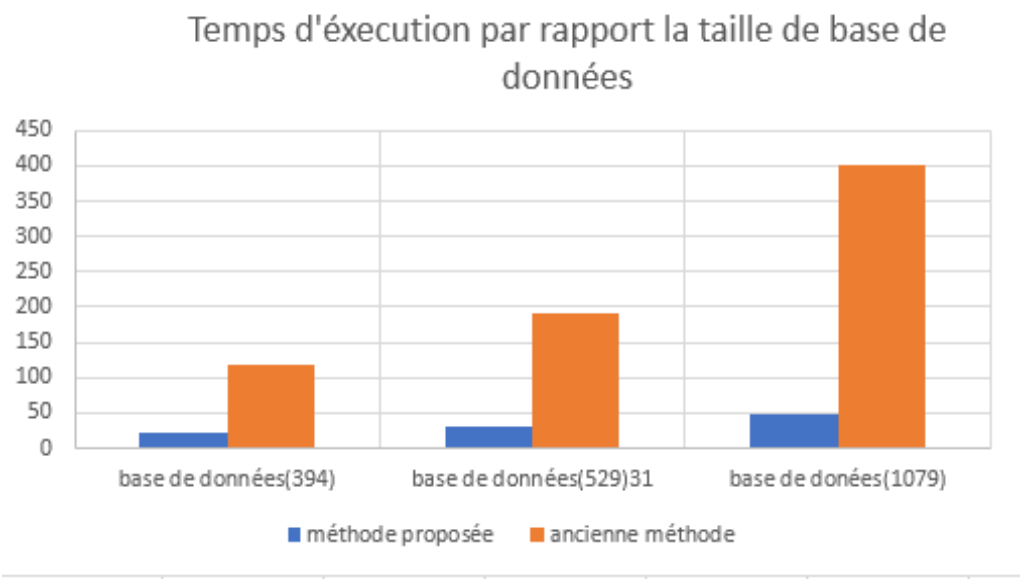


Figure 4.8 Temps d'exécution par rapport à la taille de base de données

4.5.Conclusion

Dans ce chapitre, nous avons implémenté notre algorithme d'anonymisation pour protéger la vie privée dans les réseaux sociaux, nous avons commencé par des mesures de qualité, des outils utilisés et à la fin les résultats obtenus. Nous avons remarqué que l'algorithme proposé donne une amélioration satisfaisante par rapport à l'approche comparée.

Conclusion générale et Perspectives

Les travaux présentés dans cette thèse s'inscrivent dans le cadre de la protection de la vie privée dans les réseaux sociaux.

La sécurité des informations est très importante de nos jours pour quiconque utilise les réseaux sociaux. La sécurité de l'information devrait être au premier plan de l'esprit de chacun, car une grande partie de nos informations personnelles sont disponibles sur réseaux sociaux. Déclare que la sécurité des informations est nécessaire en raison du risque généré lorsque la technologie est utilisée pour traiter les informations, car les informations peuvent être divulguées de la mauvaise manière ou à la mauvaise personne.

Nous avons proposé une nouvelle méthode qui anonymisé les données à partir des données originales basée sur l'algorithme de Mahesh R, Meyyappan **[16]**. Les nouvelles données obtenues diffèrent complètement des données d'origine, ce qui implique une grande protection de la vie privée des utilisateurs dans les réseaux sociaux.

Dans le but de comparer l'approche proposée avec celle de Mahesh R, Meyyappan, il était nécessaire d'implémenter l'algorithme de Mahesh R, Meyyappan. Les différents tests réalisés sur une variété de données réelles nous ont permis de conclure que l'algorithme proposé permet d'obtenir des meilleurs résultats et présente des améliorations significatives.

En guise de perspective, nous souhaitons réaliser les points suivants :

- Utilisation des tables de hachage pour les données sensibles,
- Lié la méthode proposée avec les méthodes de confidentialité différentielle.
- Travailler beaucoup plus avec les attributs catégoriques.

Bibliographie

[1] www.JournalduNet.com/solutions/chat/retrans/070326-nitot-mozilla-europe.shtml mars,2007

[2] Cousin C., Tout sur le web 2.0 et 3.0,edition DUNOD, 2010, page 5

[3] Barnes, John (1954). "Class and Committees in a Norwegian Island Parish". *Human Relations*

[4]: https://www.brandwatch.com/.(le deriner acces 12/2019)

[5]: fr.statista.com

[6]: https://www.apprendre-en-ligne.net/graphes/graphes.pdf

[7]: https://www.binance.vision/fr/security (le dernier acces le12/2019)

[8]: Acquisti A, Gross R (2006) Imagined communities: awareness, information sharing, and privacy on the Facebook, privacy enhancing technologies. Springer, Berlin/New York

[9]: Dwyer C, Hiltz SR, Passerini K (2007) Trust and privacy concern within social networking sites: a comparison of Facebook and MySpace. In: Proceedings of the Americas conference on information systems, Keystone

[10]: Boshmaf Y, Muslukhov I, Beznosov K, Ripeanu M (2011) The socialbot network: when bots socialize for fame and money. ACM Int Conf Proc Ser 93–102

[11]: 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

[12]: A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity.

[13]: Hay, Michael, Miklau, Gerome, Jensen, David, et al. Anonymizing social networks. Computer science department faculty publication series, 2007

[14] : Liu, Kun, Das, Kamalika, Grandison, Tyrone, et al. Privacy-preserving data analysis on graphs and social networks. book 2008.

[15] : Qiang Wang, Zhiwei Xu and Shengzhi Qu, “An Enhanced K-Anonymity Model against Homogeneity Attack”, Journal of software, 2011

[16]: Mahesh R, Meyyappan T, “A New Method for Preserving Privacy in Data Publishing”, International workshop on cryptography and Information Security, CS&IT proceedings,2012

[17]: Zhou IEEE 24th International Conference on Data Engineering Preserving, Privacy in Social. Networks Against Neighborhood. Attacks ,2008

[18] : http://archive.ics.uci.edu/ml/datasets (le dernier acces 12/2019)