UNIVERSITÉ DU QUÉBEC À MONTRÉAL

UN MODÈLE DE CONCEPTION POUR LA PROTECTION DE LA CONFIDENTIALITÉ ET DE L'ANONYMAT DES UTILISATEURS D'UN SERVICE LIVRÉ PAR UNE PLATEFORME MOBILE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

LORRY JAMES ENCARNACION

OCTOBRE 2021

REMERCIEMENTS

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner toute ma gratitude.

Je voudrais remercier dans un premier temps mon directeur de mémoire, M. Guy Bégin pour son support, sa patience, sa disponibilité et surtout ses commentaires qui ont contribué à alimenter ma réflexion.

Je voudrais également remercier ma mère Maria Encarnacion, mon père Gérard Bros ainsi que ma soeur Taina Margarita Encarnacion qui ont placé toutes leurs confiances en ma réussite.

Je tiens à témoigner toute ma reconnaissance envers Patrick Pierre, Betchinie Gilles, les membres du laboratoire Grisq et Teamlab spécialement André Mondoux, Jonathan Bonneau, Simon Colin et Jude Jacob Nsiempba pour leurs soutiens constant et leurs encouragements, ainsi que toute la famille Enloja.

Et à Dieu, de m'avoir donné le courage et la santé.

Lorry James Encarnacion

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Au cours de la dernière décennie, les données créées à partir des sources comme les courriels, les tweets, et les messages Facebook, ont connu une croissance phénoménale dans la façon dont les gens interagissent avec les systèmes à travers le monde.

L'exploration et l'exploitation de données à grande échelle sont cruciales pour tirer des enseignements précieux de ce déluge de données. Par contre, les renseignements personnels des utilisateurs peuvent être divulgués. Par exemple, les données personnelles sur les employés et les clients peuvent être volées suite à des attaques dans des bases de données appartenant à des organismes publics et à des entreprises. De telles divulgations peuvent mener à de graves violations de la vie privée. Bien que le traitement des données personnelles des utilisateurs devienne incontournable, l'anonymisation, comme moyen d'assurer la confidentialité des renseignements personnels dans l'exploitation de gros volumes donnée, est de plus en plus envisagée. Cependant, d'après la littérature consultée, il n'existe pas de technique qui permet de protéger la confidentialité et l'anonymat des utilisateurs à 100%.

Dans ce mémoire nous avons exploré les différentes techniques de protection de données les mieux considérées à ce jour. Nous avons par la suite mis en place un modèle de conception que nous avons appelé PROCOM qui en cinq (5) étapes nous a permis de combiner ces techniques, dans le but de préserver l'anonymat et la confidentialité des utilisateurs d'un service livré par une plateforme mobile. Nous avons aussi appliqué PROCOM au projet de recherche USAGES MOBILES. Ce projet vise à étudier les pratiques et les perceptions des utilisateurs des médias socionumériques dans un contexte de mobilité.

Les résultats que nous avons obtenus nous ont permis de mieux anonymiser les données du projet USAGES MOBILES tout en minimisant la perte de qualité des données originales.

Mots clés : confidentialité des données, sécurité de l'information, respect de la vie privée, anonymisation, base de données.

ABSTRACT

Over the past decade, data created from sources such as emails, tweets, and Facebook messages has seen a phenomenal growth in the way people interact with systems around the world. However, the personal information of users can be shared. For example, personal data on employees and customers can be stolen as a result of attacks on databases belonging to public agencies and companies. Such disclosures can lead to serious breaches of privacy. Although the processing of personal data of users is becoming increasingly important, anonymization, as a means of ensuring the confidentiality of personal information in the exploitation of large volumes of data, is increasingly being considered. However, according to the literature consulted, there is no anonymization technique that can protect the confidentiality and anonymity of users 100%.

In this masters thesis we have explored the various anonymization techniques that are the best known to date, and we then set up an anonymization model called PROCOM which in five (5) steps allowed us to combine these anonymization techniques in order to preserve the anonymity and confidentiality of users of a mobile platform. We also applied PROCOM to the research project called USAGES MOBILES.

The results we obtained allowed us to better anonymize the data of the USAGES MOBILES project while minimizing the loss of original data.

Index Terms : data privacy, information security, privacy, anonymization, database.

ACRONYMES

ANS	Attribut non sensible
AS	Attribut sensible
EMD	Distance du terrassier (*Earth Mover Distance*)
GO	Giga octet
GPS	Système mondial de positionnement (*Global Positioning System*)
Grisq	Groupe de recherche sur l'information et la surveillance au quotidien
IE	Identifiant explicite
IPI	Informations personnellement identifiables
NAS	Numéro assurance sociale
PbD	Confidentialité par conception (*Privacy by Design*)
PROCOM	Protection de la confidentialité et de l'anonymat des utilisateurs d'un service livré par une application mobile.
PTSD	Trouble de stress post-traumatique (*Post-traumatic stress disorder*)
QI	Quasi identifiant

RGPD	Règlement Général pour la Protection des Données
SDC	Contrôle statistique de la divulgation (*Statistical Disclosure Control*)
UM	Usages mobiles
UUID	Identifiant universel unique (*Universally Unique IDentifier*)

CHAPITRE I

INTRODUCTION

1.1 Mise en contexte

En raison de l'explosion des données, de changement majeur des matériels informatiques et des plateformes, une variété de nouveaux algorithmes d'exploration de données ont été proposés. Ces algorithmes traitent entre autres des données confidentielles, telles que les transactions financières, les dossiers médicaux, le trafic de communication réseau, etc. L'exploration de données sensibles par rapport à la vie privée devient une grande préoccupation, car de plus en plus d'informations sur des utilisateurs appartenant à des organismes publics et à des entreprises peuvent être obtenues (Kataoka *et al.*, 2014). Prenons l'exemple d'Equifax, une agence d'évaluation du crédit, qui a admis le 7 septembre 2017 que des pirates avaient compromis les informations de plus de 140 millions de personnes entre mai et juillet de la même année. Les pirates informatiques on pu avoir accès aux numéros de sécurité sociale, aux dates de naissance, aux numéros de permis de conduire et aux informations de carte de crédit des clients (Rosati *et al.*, 2020). Des informations qui pourraient servir à des usurpations d'identité pour des demandes de prêts et de cartes de crédit.

Il est donc important de pouvoir faire l'extraction des données tout en protégeant la vie privée des personnes. En outre, l'un des aspects qui sont principalement pris

en compte serait de savoir comment garantir que les informations personnelles, telles que le numéro de carte d'identité, le nom, l'adresse, etc., ne seraient pas révélées dans le processus d'exploration de données (Okuno *et al.*, 2011).

1.2 Motivation

L'utilisation des applications mobiles génère une énorme collection de données spatio-temporelles, appelées données d'objets en mouvement ou données de mobilité. Ces données peuvent être utilisées à diverses fins d'analyse, telles que le contrôle du trafic urbain, la gestion de la mobilité, la planification urbaine et les services de publicités basées sur la localisation (Bonchi *et al.*, 2011).

Il est clair que les données spatio-temporelles ainsi collectées peuvent aider un attaquant à découvrir des informations personnelles et sensibles telles que les habitudes de l'utilisateur, coutumes sociales, préférences religieuses ou sexuelles des individus. Par conséquent, cela soulève de sérieuses préoccupations concernant la vie privée. Néanmoins, il existe l'anonymisation qui permet de protéger les données sensibles des utilisateurs. Et la pseudonymisation qui offre souvent seulement une apparence de protection.

Selon la norme ISO 29100 : 2011, l'anonymisation est un «processus par lequel des informations personnellement identifiables (IPI) sont irréversiblement altérées de telle façon que le sujet des IPI ne puisse plus être identifié directement ou indirectement, que ce soit par le responsable du traitement des IPI seul ou en collaboration avec une quelconque autre partie». En d'autres termes, elle rend ambiguë l'information de l'utilisateur, pour que l'information de cet individu soit indistinguable de celles d'autres individus. L'anonymisation est un processus complexe, notamment parce qu'il tente de satisfaire deux objectifs contradictoires que sont : l'utilité des données (c'est-à-dire leur qualité) et leur sécurité (c'est-à-dire leur confidentialité). Par conséquent, les détenteurs de données (concept expliqué

au chapitre 2) doivent mettre en œuvre un processus de protection qui réponde au mieux à la confidentialité et à l'utilité de leurs données. Pour sa part la pseudonymisation, permet de remplacer les identifiants réels des utilisateurs (nom, numéro assurance sociale (NAS), etc.) par des pseudonymes, dont la caractéristique est qu'ils doivent rendre impossible tout lien entre cette valeur et l'individu réel. Toutefois, la pseudonymisation est insuffisante pour garantir l'anonymat, car la combinaison d'autres champs peut permettre de retrouver l'individu concerné (Bonchi *et al.*, 2011). Ces deux concepts sont détaillés respectivement dans le chapitre 2 et 3.

1.3 Problématique

La plupart des téléphones intelligents sont équipés de nombreux capteurs tels que le GPS, le microphone, l'accéléromètre, et les capteurs de proximité. De ce fait, le développement d'applications mobiles devient populaire pour la collecte de données personnelles ou d'environnement. Selon (Isaac, 2016), dans le modèle d'affaires des applications gratuites, les données sont une «monnaie» et des stratégies d'acquisitions de données spécifiques sont déployées pour les collecter («capture»), elles proviennent des utilisateurs des applications, des partenaires, des usages, des interactions que génèrent les applications. Par contre, la plupart des systèmes d'exploitation des téléphones intelligents actuels ne parviennent souvent pas à fournir aux utilisateurs une visibilité sur la manière dont les applications tierces collectent et partagent leurs données privées. C'est pour cela que les téléphones intelligents sont souvent la source de fuite de données.

À partir des données collectées, on peut déduire des informations sur la santé, l'emplacement, le déplacement des utilisateurs, ainsi que leur environnement (par exemple, la pollution, le bruit, les conditions météorologiques). Prenons l'exemple du Centre national américain des anciens combattants pour le trouble de stress

post-traumatique (PTSD en anglais), qui a publié en 2011 PTSD Coach, une application mobile destinée à fournir des outils de psychoéducation et d'autogestion aux survivants de traumatismes présentant des symptômes de PTSD. Des recherches émergentes sur PTSD Coach démontrent une grande satisfaction des utilisateurs, une grande faisabilité et une amélioration des symptômes du PTSD et d'autres résultats psychosociaux (Kuhn *et al.*, 2018). Bien que la plupart de ces applications permettent d'améliorer la santé et la vie des personnes en utilisant les données recueillies, elles soulèvent cependant de graves problèmes de confidentialité pour les utilisateurs, car les données collectées pourraient contenir des informations personnelles, confidentielles et sensibles (Zhang *et al.*, 2016). À titre d'exemple, les informations sur la localisation des utilisateurs peuvent être divulguées à une application malveillante avec une intention criminelle, ce qui constitue une menace pour la sécurité des utilisateurs.

Selon une étude réalisée par (Enck *et al.*, 2014), les applications Android sont capables de signaler par exemple, des informations aux serveurs de publicité. Ces chercheurs ont sélectionné au hasard 30 applications populaires d'Android qui utilisent des données de localisation, de caméra ou de microphone, et ont montré que 15 de ces applications ont signalé des informations de localisation aux serveurs de publicité à distance, et que les deux tiers des applications utilisaient de manière suspecte des données sensibles. Cependant, il existe un certain nombre de modèles d'anonymisation tels que le k-anonymat et la confidentialité différentielle qui ont été suggérés ces dernières années afin de réaliser une exploration de données préservant la confidentialité. Toutefois, ces modèles à eux seuls ne permettent pas de pallier tous les problèmes de protection des données des utilisateurs, il y a donc un risque pour que l'anonymat d'une personne soit compromis par la collecte d'informations à partir de différentes sources (Aggarwal et Philip, 2008b).

Généralement, les entreprises collectent les données sans vraiment les filtrer, et

elles sont ensuite nettoyées. Alors, les entreprises sont encouragées à sélectionner les données les plus pertinentes qui répondent à leurs objectifs, c'est le principe de minimisation des données. Prenons l'exemple d'une société de location de véhicule qui met en place un dispositif de géolocalisation continue sur l'ensemble de sa flotte, la finalité de ce dispositif étant la lutte contre la non-restitution ou le vol de véhicule. Un tel dispositif permet de recueillir et traiter de nombreuses données dont notamment la position en temps réel de chacun des véhicules, le parcours emprunté, la durée de stationnement dans des lieux précis, etc. Ces données ne sont donc pas strictement limitées à ce qui est strictement nécessaire au regard de la finalité. Dans ce cas précis, pour limiter la collecte de données et respecter le principe de minimisation, le dispositif de géolocalisation pourrait n'être activé que dans une situation d'un retard ou de vol de véhicule. Par conséquent, le défi majeur est de faire en sorte que les applications fournissent les mêmes services de façon efficace et en même temps de pouvoir protéger la confidentialité et l'anonymat des utilisateurs en rendant leurs informations moins précises. Dans ce cas, il serait crucial d'avoir un mécanisme ou un modèle de conception qui permettrait de stocker uniquement les informations utiles au besoin des applications et qui protégerait l'anonymat et la confidentialité des utilisateurs.

1.4 Question de recherche

La problématique soulevée précédemment nous amène à soulever la question de recherche principale suivante : comment peut-on combiner plusieurs techniques de protection de données afin de concevoir un modèle de conception qui permettrait de protéger la confidentialité et l'anonymat des utilisateurs d'un service livré par une plateforme mobile ?

Pour répondre à la question de recherche principale, on a relevé les questions de recherche secondaires suivantes :

— Comment peut-on préserver la confidentialité et l'anonymat des utilisateurs tout en maximisant l'utilité des données ?

— Comment peut-on utiliser le principe de minimisation des données afin d'éliminer ou de réduire certaines données que les applications collectent, et de conserver seulement le strict nécessaire, sans nuire aux besoins de l'application ?

1.5 Objectifs

Dans le cadre de notre projet, afin de répondre à notre question de recherche, nous avons visé les objectifs spécifiques suivants :

— L'identification d'un ensemble de techniques de protection de données qui peuvent contribuer à protéger la confidentialité des données sensibles des utilisateurs.

— L'élaboration d'une démarche de conception qui permettrait de combiner les techniques de protection de données que nous avons identifiées et ensuite de voir a priori et a posteriori dans quelle mesure on pourrait éliminer ou assouplir certaines contraintes ou exigences de l'application et de quand même être capable de fournir les mêmes services, avec potentiellement un certain niveau de dégradation prévisible et gérable.

— La mise sur pied d'un prototype qui va nous permettre de mettre à l'épreuve les idées, les méthodes et les techniques envisagées pour notre modèle de conception.

1.6 Méthodologie et contributions

Dans le cadre de notre projet, nous avons mis en oeuvre une démarche de conception pour pouvoir éliminer ou assouplir certaines contraintes ou exigences des

applications mobiles dans le but de protéger la confidentialité et l'anonymat des utilisateurs.

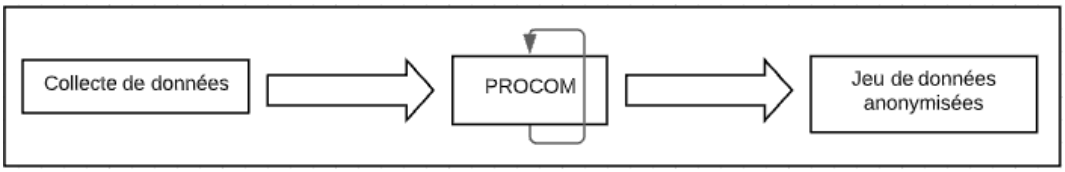


Figure 1.1 Phases de la résolution du problème

Dans le but d'aider les chercheurs dans le choix d'une ou plusieurs techniques de protection de données, nous avons présenté dans la figure 1.1 les différentes phases qui nous mèneront à tenter de renforcer l'anonymat. La phase 1 correspond à la collecte des données non anonymes ou faiblement anonymes d'une application mobile en appliquant le principe de minimisation des données. La phase 2 correspond à notre démarche de conception PROCOM (Protection et Confidentialité Mobile), pour la protection de la confidentialité et de l'anonymat des utilisateurs d'un service livré par une application mobile.

PROCOM compte cinq (5) étapes principales à suivre pour pouvoir résoudre le problème d'anonymisation des données. À l'étape 1 nous avons évalué les besoins d'une application mobile dont les données ont été collectées. À l'étape 2, nous avons vérifié la protection accordée aux données, en utilisant le moteur de recherche Google. À l'étape 3, nous avons identifié les données à anonymiser c'est-à-dire identifié les données à supprimer, masquer ou conserver. À l'étape 4, nous avons identifié les différents types d'anonymisation, à savoir si nous devons utiliser une anonymisation réversible ou irréversible. Et à l'étape 5 nous avons proposé plusieurs combinaisons possible des méthodes et techniques de protection de données qui s'adaptent le mieux entre elles. Toutes ces étapes sont présentées de manière détaillée dans le chapitre 4. La dernière phase de notre solution corres-

pond à la validation des données anonymes produites par PROCOM (voir figure 1.2).

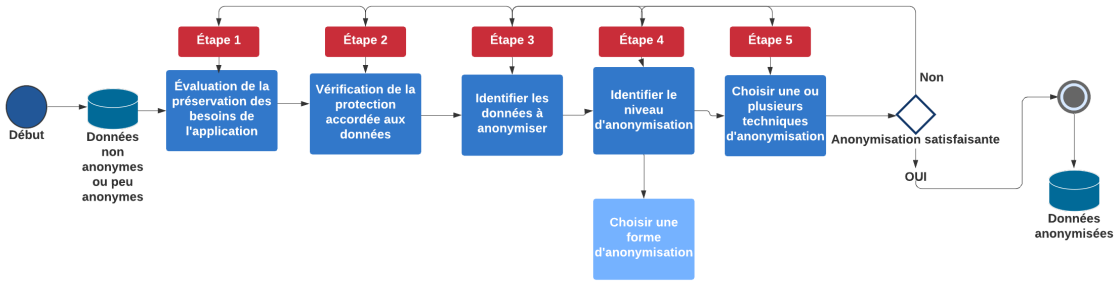


Figure 1.2 Les étapes du processus de PROCOM.

Pour éviter que le processus d'anonymisation ne dégrade trop la précision des données et leur utilité, il existe des métriques de qualité ou d'évaluation appelées « data metrics » permettant de mesurer la qualité des données d'une table anonyme en la comparant à la qualité des données de la table originale. Dans le chapitre 6 nous avons appliqué la métrique de complétude des données et la métrique de discernabilité pour quantifier le gain d'anonymat et la perte d'utilité des données produites par PROCOM.

La principale contribution de notre projet de recherche est la combinaison de plusieurs techniques et méthodes de protection de données les plus utilisées à ce jour (elles sont détaillées dans le chapitre 3). Pourquoi combiner plusieurs techniques ? Suite à l'examen des différentes techniques et méthodes, nous avons pu constater les avantages et les inconvénients de chacune d'entre elles. Nous avons vu qu'il n'y avait pas de technique meilleure ou moins bonne qu'une autre, mais que chacune d'elles trouvait son utilité en fonction du contexte et des données à anonymiser. Le fait de combiner plusieurs techniques va nous permettre de concrétiser le processus d'anonymisation des données tout en offrant des aides à la prise de décision.

1.7 Organisation du mémoire

Le reste du mémoire est organisé de la manière suivante :

— Le chapitre 2 présente une vue globale des concepts clés liés à la notion d'anonymisation ainsi que quelques modèles d'attaques à la vie privée que peuvent subir les données.

— Le chapitre 3 présente une revue de la littérature sur les méthodes de protection ainsi que les modèles d'anonymisation les plus connus à ce jour, et pour chacune des méthodes et modèles, nous présentons les avantages et les inconvénients.

— Le chapitre 4 présente l'analyse et la conception de notre démarche de conception PROCOM conçue dans le but de protéger la confidentialité et l'anonymat des utilisateurs.

— Le chapitre 5 présente l'application de notre démarche PROCOM à un cas concret.

— Le chapitre 6 présente les résultats produits par notre démarche PROCOM ainsi que les métriques de qualité utilisées pour valider les données produites par PROCOM. Par la suite, nous présentons nos conclusions sur le travail réalisé ainsi que quelques recommandations pour avoir une meilleure anonymisation.

CHAPITRE II

ANONYMISATION DES DONNÉES : MODÈLES D'ATTAQUES

2.1 Introduction

Les données massives sont devenues incontournables (ou une réalité) ces dernières années. Elles sont collectées par une multitude de sources indépendantes, puis elles sont fusionnées et analysées pour générer des connaissances. Bien que les données massives constituent une ressource précieuse dans de nombreux domaines, elles ont un effet secondaire important, car la vie privée des personnes dont les données sont collectées et analysées (souvent à leur insu) est de plus en plus menacée (Soria-Comas et Domingo-Ferrer, 2016). L'anonymisation est un moyen pour surmonter les conflits entre les principes de confidentialité et les analyses pouvant être effectuées sur ces données. Dans ce chapitre nous allons présenter les concepts clés liés à l'anonymisation ainsi que les différents modèles d'attaques que peuvent subir les données, tels que les modèles d'attaques par liens d'enregistrements, les modèles d'attaques par inférence d'attributs, par appartenance et probabilistes.

2.2 Notions préliminaires

Pour exploiter pleinement l'utilité analytique des données, elles doivent être rendues disponibles aux chercheurs. Par contre, ces données sont susceptibles de contenir des informations confidentielles. De ce fait, la publication des données

doit maintenir l'équilibre entre l'utilité des données et le droit au respect de la vie privée. L'anonymisation est une réponse à ce besoin d'équilibre, car elle est l'approche la plus courante pour préserver la confidentialité des données lors de leur publication. Elle consiste à modifier le contenu des enregistrements (Kiran et Kavya, 2012). Le contrôle statistique de la divulgation (SDC pour *Statistical Disclosure Control*) est une discipline qui concerne l'anonymisation des données statistiques qui contiennent des informations confidentielles sur des entités individuelles telles que des personnes ou des entreprises. Le but du SDC est d'empêcher les tiers (décideurs, chercheurs universitaires et grand public par exemple) de s'appuyer sur des données pour identifier des personnes et divulguer des informations confidentielles les concernant (Solé *et al.*, 2012).

Selon (Singh et Parihar, 2013), lors de la publication des données personnelles, il existe au moins trois catégories d'acteurs : le détenteur des données, la personne concernée par la donnée elle-même, et le destinataire des données. Par exemple, un hôpital collecte des données sur les patients et publie les dossiers des patients dans un centre médical externe. Dans cet exemple, l'hôpital est le détenteur des données, les patients sont les personnes dont les données sont collectées et le centre médical est le destinataire des données.

Définition 2.1. Un identifiant explicite (IE) est un ensemble d'attributs qui contient des informations qui peuvent explicitement identifier les personnes concernées par les enregistrements tels que le nom ou le numéro d'assurance sociale (NAS).

Définition 2.2. Un quasi identifiant (QI) est un ensemble d'attributs tels que le code postal, l'âge et le sexe et, dont les valeurs, prises ensemble, peuvent potentiellement identifier les personnes concernées par les données.

Définition 2.3. Les attributs sensibles (AS) comprennent des informations sensibles propres à une personne, telles que l'état de santé ou le salaire.

Définition 2.4. Les attributs non sensibles (ANS) contiennent tous les attributs qui ne font pas partie des trois catégories précédentes.

Les données sont en général représentées à partir d'une table relationnelle [1] représentant des individus de la forme : T (IE, QI, AS, ANS) (voir tableau 2.1).

IE	QI		AS
Nom	**Age**	**Sexe**	**Casier Judiciaire**
Andy	28	M	Vol
James	23	M	Meurtre au premier degré
Patrick	42	M	Vol
Pierre	39	M	Vol
Gilles	19	F	Meurtre au premier degré
Flore	19	F	Possession illégale d'arme à feu
Mike	26	M	Extorsion
Sarah	28	F	Voies de fait grave

Tableau 2.1 Exemple de table originale non anonyme

Dans le tableau 2.1, l'attribut «nom» est un IE, les attributs «âge» et «sexe» constituent des QI et l'attribut «casier judiciaire» est un AS. Chaque ligne de la table correspond aux personnes concernées par les données, c'est-à-dire la personne concernée par la donnée elle-même. Il est possible de faire des traitements sur les

1. Une table relationnelle est un ensemble de données organisées sous forme d'un tableau où les colonnes correspondent à des catégories d'information (une colonne peut stocker des numéros de téléphone, des noms, des prénoms,...) et les lignes à des enregistrements, également appelés entrées.

attributs QI et AS. Ces attributs peuvent être continus, quand leurs valeurs sont numériques et qu'ils peuvent faire l'objet d'opérations arithmétiques, par exemple nous pouvons ajouter ou multiplier une valeur aléatoire à l'attribut «âge», afin de masquer sa valeur réelle, et catégoriels c'est-à-dire qu'on ne peut leur appliquer d'opération arithmétique, par exemple l'attribut «casier judiciaire» (Fienberg et McIntyre, 2004).

Dans un contexte de données massives les résultats rapportés dans des travaux récents donnent à penser que les distinctions entre QI, AS et ANS pourraient devenir moins claires. Par exemple, (Ke *et al.*, 2018) montrent que certains attributs peuvent être à la fois des AS et des QI dans la pratique. Et ils considèrent que les AS sont plutôt considérés comme des QI sensibles.

2.3 Modèles d'attaques

Une définition stricte de la préservation de la vie privée a été donnée par (Dalenius, 1977). Selon lui, « L'accès aux données publiées ne devrait pas permettre à l'attaquant d'en apprendre davantage sur une victime cible par rapport à l'absence d'accès à la base de données, même avec la présence des connaissances de base de l'attaquant obtenues à partir d'autres sources ». Cependant, si la publication des données publiées enseigne quoi que ce soit à l'attaquant, la notion de vie privée est irréalisable. Par exemple, si une base de données médicales nous apprend que le tabagisme provoque le cancer, l'attaquant qui sait qu'un individu fume, pourrait déduire que ce dernier a des chances de développer un cancer, indépendamment de sa présence ou de son absence dans cette base de données (Dwork *et al.*, 2014). Selon (Fung *et al.*, 2010b), une protection absolue de la vie privée est impossible, en raison de la connaissance de base de l'attaquant. Ils considèrent qu'il y a deux catégories de menaces à la vie privée ou attaques à la vie privée, à savoir, les modèles d'attaques par liens et les modèles d'attaques probabilistes ou inférences

probabilistes.

2.4 Modèle d'attaque par liens

Une attaque par liens se produit lorsqu'un adversaire est capable de lier la personnes concernée par la donnée à un enregistrement dans une table de données, à un attribut sensible dans une table de données ou à la table de données publiée elle-même. Il existe les attaques par lien d'enregistrements, par inférence d'attributs et par lien de tables (Fung *et al.*, 2010b).

2.4.1 Attaque par lien d'enregistrements

L'attaque par lien d'enregistrements est possible si l'attaquant connaît une valeur QI de la victime dont les informations ont été publiées et si cette table contient très peu d'enregistrements ayant la même valeur QI. Grâce aux connaissances supplémentaires de l'attaquant, il est possible qu'il puisse identifier de manière unique l'enregistrement de la victime dans le groupe (Fung *et al.*, 2010b).

Example 2.4.1. Dans le tableau 2.2, si l'attaquant sait que la victime a un casier judiciaire et fait partie des données publiées, s'il connaît la valeur des QI «âge» et «sexe» de la victime il peut facilement lier l'identité de la victime à son casier judiciaire. Par exemple, un homme de 23 ans sera identifié comme une personne qui a commis un meurtre au premier degré grâce à des connaissances supplémentaires de l'attaquant.

2.4.2 Attaque par inférence d'attributs

Dans le cas de l'attaque par inférence d'attributs, l'attaquant peut ne pas identifier avec précision l'enregistrement de la victime cible, mais peut déduire ses valeurs sensibles des données publiées, en fonction de l'ensemble de valeurs sensibles associées au groupe auquel appartient la victime (Fung *et al.*, 2010b).

Age	Sexe	Casier Judiciaire
28	M	Vol
23	M	Meurtre au premier degré
42	M	Vol
39	M	Vol
19	F	Meurtre au premier degré
19	F	Possession illégale d'arme à feu

Tableau 2.2 Exemple de table publiée de manière anonyme

Example 2.4.2. Dans le tableau 2.2, supposons que l'attaquant sache que Andy est un homme qui a un casier judiciaire. L'attaquant peut déduire que Andy a 75% de chances d'avoir commis un vol car trois (3) des quatre (4) hommes ont commis un vol. Quelle que soit l'exactitude de l'attaque, la vie privée de Andy a été compromise.

Dans la littérature consultée, il existe plusieurs types d'attaques par inférence d'attributs, telles que les attaques par homogénéité, les attaques par similarité et les attaques fondées sur les connaissances de base.

— L'attaque par homogénéité est possible quand toutes les valeurs du groupe QI partagent la même valeur d'AS (Truta *et al.*, 2007). Par exemple, si l'attaquant sait que la victime est âgée de plus de 25 ans, et dans le tableau 2.2, tous les gens qui sont âgés de plus de 25 ans ont commis un vol, on peut donc déduire que notre victime a commis un vol.

— Avec l'attaque par similarité, lorsque les valeurs des attributs sensibles du groupe sont distinctes, mais sémantiquement similaires, un attaquant peut apprendre des informations importantes (Li *et al.*, 2007). Par exemple si dans le tableau 2.2, la victime se trouve dans un groupe où tout le monde a commis soit un meurtre au premier degré ou au second degré on peut déduire que la victime a commis un meurtre.

— On dit qu'il y a attaque fondée sur les connaissances de base quand l'attaquant utilise des connaissances de base qui lui permettent d'éliminer certaines valeurs possibles des attributs sensibles pour des individus spécifiques. Prenons l'exemple de (Machanavajjhala *et al.*, 2006), qui suppose qu'un attaquant connaisse Emako, une jeune japonaise de 21 ans qui habite au 13068. Sur la base de cette information l'attaquant peut conclure que les informations concernant Emako figurent parmi les dossiers numéro 1, 2, 3 ou 4 dans le tableau 2.3. Sans informations supplémentaires, l'attaquant n'est pas sûr qu'Emako ait attrapé un virus ou a une maladie cardiaque. Cependant, il est bien connu que les japonais ont une incidence extrêmement faible de maladies cardiaques. Cette connaissance de base permet à l'attaquant de conclure qu'Emako est très probablement atteinte d'une infection virale. Selon (Truta *et al.*, 2007), la protection de la vie privée de la victime contre les attaques de connaissance de base est plus difficile car le détenteur des données n'est pas au courant du type de connaissances de base que l'attaquant peut posséder.

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Tableau 2.3 Tableau original des patients (Machanavajjhala *et al.*, 2006)

2.4.3 Attaque par appartenance

Les deux modèles présentés précédemment supposent que l'attaquant sait déjà que l'enregistrement de la victime se trouve dans la table qui a été publiée. Cependant, dans certains cas, la présence (ou l'absence) de l'enregistrement de la victime dans la table révèle déjà des informations sensibles sur la victime.

Example 2.4.3. Supposons qu'une table anonyme T a été publiée (tableau 2.4). Pour lancer une attaque par appartenance sur une victime cible, par exemple, Gilles, sachant que l'attaquant ait accès également à une table publique externe E (tableau 2.5) avec $T \subseteq E$. La probabilité que Gilles soit soit présent dans T est $\frac{4}{5} = 0,8$ parce qu'il y a 4 enregistrements dans T et 5 enregistrements dans E contenant [Infirmière, F, [30, 40)]. L'attaque par appartenance se produit si un attaquant peut déduire avec confiance la présence ou l'absence de l'enregistrement de la victime dans la table publiée (Fung *et al.*, 2010b).

Age	Sexe	Poste	Casier Judiciaire
[20,30)	M	Préposé	Meurtre au premier degré
[20,30)	M	Préposé	Meurtre au premier degré
[20,30)	M	Préposé	Vol
[30,40)	F	Infirmière	Voies de fait grave
[30,40)	F	Infirmière	Vol
[30,40)	F	Infirmière	Vol
[30,40)	F	Infirmière	Vol

Tableau 2.4 Table publiée de manière anonyme

Nom	Age	Sexe	Poste
Gilles	[30,40)	F	Infirmière
Andy	[20,30)	M	Préposé
Flore	[30,40)	F	Infirmière
Patrick	[20,30)	M	Préposé
Anne	[30,40)	F	Infirmière
Mike	[20,30)	M	Préposé
Emako	[30,40)	F	Infirmière
Tommy	[20,30)	M	Préposé
Sarah	[30,40)	F	Infirmière

Tableau 2.5 Table publique externe

En effet, en publiant plusieurs tables anonymes, on ne peut exclure la possibilité de rapprochement entre elles dès lors qu'elles partagent des valeurs de QI. Certains rapprochements peuvent mener à la divulgation de données sensibles. A titre d'exemple, supposons que le Tableau 2.7 sur les maladies a été publié au même titre que le Tableau 2.6 sur les catégories professionnelles. Le Tableau 2.6 révèle la tranche d'âge et le niveau d'éducation des individus. A titre d'exemple, Alice, la victime de l'attaquant, a un âge compris entre 19 et 23 ans. Elle a un niveau supérieur d'études. En rapprochant ces deux tables, l'attaquant peut déduire qu'Alice a une probabilité de $3/4 = 75\%$ de se trouver dans le Tableau 2.7 (le chiffre 3 correspond à la taille de la classe d'équivalence du QI « [19, 23], Supérieur » dans le Tableau 2.6 et le chiffre 4 correspond à celle du même QI dans le Tableau 2.7). Sachant que les individus de QI « [19, 23], Supérieur » sont tous atteints d'un cancer et qu'Alice a ce même QI, on peut déduire que la probabilité qu'Alice soit atteinte d'un cancer est de 75%.

Age	Education	Nom
[19,23]	Supérieur	Malik
[19,23]	Supérieur	George
[27,30]	Supérieur	Fred
[27,30]	Supérieur	Jean
[19,23]	Supérieur	Pierre
[27,30]	Supérieur	Paul
[19,23]	Supérieur	Alice

Tableau 2.6 Table publiée de manière anonyme

Age	Education	Maladie
[19,23]	Secondaire	Maladie cardiaque
[19,23]	Secondaire	Cancer
[27,30]	Secondaire	Grippe
[27,30]	Secondaire	Grippe
[19,23]	Supérieur	Cancer
[19,23]	Supérieur	Cancer
[19,23]	Supérieur	Cancer

Tableau 2.7 Table publique externe

2.5 Modèle d'attaque probabiliste ou inférences probabilistes

Contrairement au modèle d'attaque par liens, les modèles d'attaques probabilistes ne se concentrent pas sur les enregistrements, les valeurs des attributs, et les tables que l'attaquant peut associer à une victime cible, mais l'attaquant exploite sa croyance probabiliste sur les renseignements sensibles d'une victime après avoir consulté les données publiées. Le scénario d'attaque, fréquemment mentionné dans la littérature pour ce type de modèle, est l'attaque par dissymétrie « skewness attack » (Machanavajjhala *et al.*, 2006).

Dans cette attaque, l'adversaire déduit la valeur d'une donnée sensible de sa victime en comparant la distribution globale des valeurs de l'attribut sensible (croyance probabiliste avant analyse des données publiées) avec la distribution des valeurs de ce même attribut sensible au sein d'un groupe d'individus de même QI (croyance probabiliste après analyse des données publiées).

À titre d'exemple, dans le tableau 2.8, nous avons 10 000 enregistrements sur un virus qui n'affecte que 1% de la population. Pour les lignes 1 à 4, des mesures de confidentialité strictes ne sont probablement pas nécessaires, car les personnes

qui ne sont pas atteintes de la maladie ne se soucient pas de savoir si leur identité est découverte. Les lignes 5 à 8 ont un nombre égal d'enregistrements positifs et négatifs. Cela donne à tout le monde dans ce groupe 50% de chance d'avoir le virus, ce qui est beaucoup plus élevé que la distribution réelle. Enfin, les ligne 9 à 12 affichent un risque encore plus élevé d'attaque par inférence.

	Code postal	**Age**	**Salaire**	**Maladie**
1	476**	2*	3k	negative
2	476**	2*	4k	negative
3	476**	2*	5k	negative
4	476**	2*	6k	negative
5	4790*	>=40	7k	negative
6	4790*	>=40	8k	positive
7	4790*	>=40	9k	negative
8	4790*	>=40	10k	positive
9	476**	3*	11k	positive
10	476**	3*	12k	positive
11	476**	3*	13k	positive
12	476**	3*	14k	negative
13	4760*	4*	15k	negative
...	...	...	...	...
10000	488**	>=60	16k	negative

Tableau 2.8 Table de 10 000 enregistrements des patients d'un hôpital

2.6 Synthèse

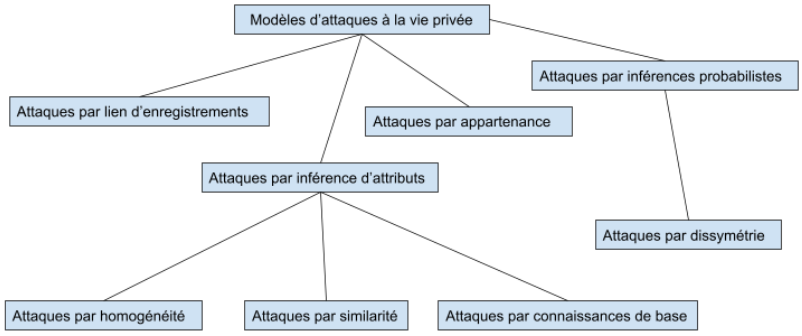


Figure 2.1 Modèles d'attaques à la vie privée

Tout au long de ce chapitre, nous avons détaillé les différents concepts liés à l'anonymisation. Nous avons présenté quelques modèles d'attaques à la vie privée, comme le montre la figure 2.1. Ces modèles sont divisés en quatre types à savoir, les attaques par lien d'enregistrements qui se produisent lorsqu'un attaquant est capable de lier la personne concernée par la donnée à un enregistrement dans une table, les attaques par inférence d'attributs quand il est capable de le lier à un attribut sensible, les attaques par appartenance quand il est capable de le lier à la table elle-même et les attaques par inférences probabilistes quand l'attaquant s'appuie sur sa croyance probabiliste. Nous avons aussi présenté un type particulier d'attaques par inférences probabilistes qui est l'attaque par dissymétrie et trois types d'attaques par inférence d'attributs : les attaques par homogénéité, les attaques par similarité et les attaques par connaissances de base.

Dans le prochain chapitre, nous allons présenter brièvement quelques méthodes et modèles de protection de la vie privée qui ont été proposés dans la littérature pour contrer ces différents types d'attaques.

CHAPITRE III

ÉTAT DE L'ART

3.1 Introduction

Plusieurs méthodes et modèles ont été proposés dans le but de préserver la confidentialité et l'anonymat des utilisateurs. Dans ce chapitre nous allons présenter les différentes méthodes de protection de données qui sont classées en deux catégories : les méthodes non perturbatrices (la généralisation, la suppression, etc.) qui réduisent la quantité d'information mise à la disposition des utilisateurs et les méthodes perturbatrices (la randomisation, la permutation, etc.) qui altèrent les données initiales mais sans en réduire la quantité. Ces deux familles de méthodes seront présentées dans les deux premières sections. Par la suite, nous présenterons une revue des modèles de protection de données les plus récents en matière de confidentialité, telles que le k-anonymat, la l-diversité, la t-proximité, etc. Pour chacune des méthodes, nous identifierons les avantages et les inconvénients. Finalement, nous ferons une synthèse résumant les deux premières sections.

3.2 Méthodes non perturbatrices

3.2.1 La pseudonymisation

Dans la méthode de pseudonymisation, les identités réelles (nominatives) des personnes sont remplacées par des pseudo-identifiants qui ne peuvent pas être liés

directement aux identités nominatives correspondantes. Pour des raisons de confidentialité, il faut éviter de stocker des informations personnelles dans le jeu de données pseudonymisées (Neubauer et Heurix, 2011).

Selon (Claerhout et DeMoor, 2005), la pseudonymisation est utile dans les scénarios de collecte de données où de grandes quantités de données provenant de sources différentes sont rassemblées pour le traitement statistique et l'exploration de données (par exemple, des études de recherche). Cependant, une utilisation négligente de la technologie de pseudonymisation pourrait créer un faux sentiment de protection de la vie privée. Le danger réside principalement dans la non-séparation des identifiants et de la charge utile. Il convient de s'assurer que les données utiles ne contiennent aucun champ susceptible d'entraîner une ré-identification indirecte, c'est-à-dire une nouvelle identification basée sur le contenu (information), et non sur les identifiants.

La pseudonymisation ne donne pas un niveau de protection suffisamment élevé, car la combinaison d'autres champs peut permettre de retrouver les individus concernés. En effet, même avec des attributs dit non sensibles, une combinaison d'attributs va probablement permettre de retrouver quelqu'un, dès lors qu'on a des connaissances annexes sur l'individu. C'est ce qu'on appelle le problème de quasi-identifiant, présenté dans le chapitre précédent.

3.2.2 La généralisation

Dans la méthode de généralisation, les valeurs des QI sont remplacées par des valeurs ou des plages de valeurs moins spécifiques, mais cohérentes sur le plan sémantique (Ghinita *et al.*, 2010). Par exemple, à partir du tableau original 3.1, nous pouvons généraliser la colonne «âge» par des plages de valeur, afin de réduire le risque d'identification (voir tableau 3.2).

Nom	Age	Sexe	Code Postal	Casier Judiciaire
Andy	28	M	12000	Vol
James	23	M	14000	Meurtre au premier degré
Patrick	42	M	18000	Vol
Pierre	39	M	19000	Vol
Gilles	19	F	17000	Meurtre au premier degré
Flore	19	F	22000	Possession illégale d'arme à feu
Mike	26	M	24000	Extorsion
Sarah	28	F	36000	Voies de fait grave

Tableau 3.1 Tableau original

Nom	Age	Sexe	Code Postal	Casier Judiciaire
Andy	[26, 30]	M	12000	Vol
James	[21, 25]	M	14000	Meurtre au premier degré
Patrick	[41, 45]	M	18000	Vol
Pierre	[36, 40]	M	19000	Vol
Gilles	[15, 19]	F	17000	Meurtre au premier degré
Flore	[15, 19]	F	22000	Possession illégale d'arme à feu
Mike	[26, 30]	M	24000	Extorsion
Sarah	[26, 30]	F	36000	Voies de fait grave

Tableau 3.2 Exemple montrant la généralisation de l'attribut âge

La généralisation est l'une des méthodes d'anonymisation les plus utilisées et l'un des principaux avantages de cette approche est qu'elle préserve la «véracité» de l'information. Cependant, des travaux récents ont montré qu'en raison du nombre d'éléments possibles, la généralisation perd une quantité considérable d'informations ce qui risque de rendre les données inutiles (Lee *et al.*, 2017). Aussi l'un des inconvénient de la généralisation c'est qu'elle nécessite la définition d'une hiérarchie pour chaque attribut composant le QI, afin de pouvoir remplacer une valeur par son ancêtre direct dans la hiérarchie de généralisation, et cela à chaque étape de la généralisation.

3.2.3 La suppression

Dans la méthode de suppression, la valeur de l'attribut est complètement supprimée. Cette méthode supprime certains attributs QI appelés aussi valeurs aberrantes des micro-données. La suppression peut se faire de deux manières : la suppression est dite globale quand on supprime une ou plusieurs lignes dans leur totalité, sinon, la suppression est locale, quand on remplace un attribut dans une ligne dans la table originale par une marque, soit par des «*» ou des « ?» (Ped-

dapunnaiahp et Kiran, 2016). Par exemple à partir du tableau 3.3, la valeur de l'attribut QI «nom» a été supprimée pour chaque personne, il s'agit d'une suppression globale de l'attribut «nom» du tableau original 2.1.

Age	**Sexe**	**Casier Judiciaire**
[26, 30]	M	Vol
[21, 25]	M	Meurtre au premier degré
[41, 45]	M	Vol
[36, 40]	M	Vol
[15, 19]	F	Meurtre au premier degré
[15, 19]	F	Possession illégale d'arme à feu
[26, 30]	M	Extorsion
[26, 30]	F	Voies de fait grave

Tableau 3.3 Exemple montrant le résultat de la suppression globale de l'attribut nom

La méthode de suppression réduit le risque de ré-identification par l'utilisation d'enregistrements publics, tout en réduisant la précision des applications sur les données transformées (Samarati et Sweeney, 1998). Cependant, selon (Jia *et al.*, 2014), la suppression induit une perte d'information qui provoque une distorsion significative dans la structure des enregistrements dégradant l'utilité des données.

3.3 Méthodes perturbatrices

3.3.1 La randomisation

La méthode de randomisation est une technique d'exploration de données préservant la confidentialité dans laquelle un bruit est ajouté aux données afin de

masquer la valeur des attributs des enregistrements. Le bruit ajouté est suffisamment important pour que des valeurs d'enregistrement individuelles ne puissent pas être récupérées (Aggarwal et Philip, 2008a). Le bruit peut être introduit soit en ajoutant ou en multipliant des valeurs aléatoires aux enregistrements numériques soit en supprimant des items réels et en ajoutant des fausses valeurs à l'ensemble des attributs (Agrawal et Srikant, 2000).

L'un des principaux avantages de la méthode de randomisation est qu'elle est relativement simple, ne nécessite pas de connaître la distribution des autres enregistrements dans les données. La méthode de randomisation peut être implémentée au moment de la collecte des données (Aggarwal et Philip, 2008a). Cependant, cet avantage entraîne certaines faiblesses, car la randomisation traite tous les enregistrements de la même manière. De ce fait, il faut faire preuve de plus de précaution en ajoutant du bruit à tous les enregistrements des données, parce qu'en ajoutant trop de bruit nous ne serons plus certain de la véracité de ces données, et cela va réduire l'utilité des données à des fins d'exploration.

3.3.2 L'échange, permutation ou la technique de «Swapping»

Nous notons que l'ajout de bruit dans les données n'est pas la seule technique pouvant être utilisée pour perturber les données. Une méthode apparentée est celle de l'échange de données, qui a été proposée pour la première fois par Tore Dalenius et Steven Reiss (1978) comme une méthode de préservation de la confidentialité dans un ensemble de données qui contient des variables catégorielles c'est-à-dire des variables sur lesquelles aucune opération arithmétique ne peut être appliquée. L'idée de base de la méthode est de transformer une base de données en interchangeant des valeurs de variables sensibles entre des enregistrements individuels (Fienberg et McIntyre, 2004).

Selon (Gunawan et Mambo, 2019), l'un des avantages de cette technique est qu'elle

ne réduit pas le nombre d'éléments et ne provoque pas de perte d'éléments dans une base de données. Bien que les enregistrements individuels affectés soient modifiés, le processus d'échange préserve la distribution globale des valeurs comprises dans la base de données. Par conséquent, certains types de calculs d'agrégats peuvent être effectués exactement sans porter atteinte à la confidentialité des données (Fienberg et McIntyre, 2004). Il est important de savoir que cette technique ne suit pas le principe général de la randomisation qui permet de perturber la valeur d'un enregistrement indépendamment des autres enregistrements. Par contre, selon (Fienberg et McIntyre, 2004), il est difficile d'identifier les échanges de données appropriés dans des bases de données volumineuses.

3.3.3 La micro-agrégation

La micro-agrégation est une famille de techniques de SDC pour les micro-données en continu c'est-à-dire quand leurs valeurs sont numériques et qu'ils peuvent faire l'objet d'opérations arithmétiques. La micro-agrégation repose sur le fait que les règles de confidentialité en vigueur autorisent la publication de jeux de micro-données si les enregistrements correspondent à des groupes de k personnes ou plus.

L'application stricte de ces règles de confidentialité conduit à remplacer les valeurs individuelles par des valeurs calculées sur de petits agrégats (micro-agrégats) avant la publication. C'est le principe de base de la micro-agrégation. Pour obtenir des micro-agrégats dans un ensemble de micro-données comportant n enregistrements, ceux-ci sont combinés pour former des groupes de taille égale à k, k étant utilisé plus loin pour le k anonymat. Pour chaque attribut, la valeur moyenne sur chaque groupe est calculée et sert à remplacer chacune des valeurs d'origine. Enfin si le nombre total d'enregistrements n n'est pas un multiple de k, on reste avec un groupe incomplet à la fin (Domingo-Ferrer, 2008).

Example 3.3.1. Nous allons à titre d'exemple appliquer le principe de base de la micro-agrégation à l'attribut «âge». Premièrement à partir du tableau original 2.1 nous avons trié les enregistrements selon l'attribut «âge» en formant des groupes, chacun devant contenir au moins trois enregistrements comme le montre le tableau 3.1 . Ensuite nous allons remplacer la valeur de l'attribut âge pour chaque enregistrement par la valeur de la moyenne calculée pour chaque groupe comme le montre le tableau 3.2.

Nom	Age	Sexe	Code Postal
Andy Ken Mike	5 6 7	M M M	12000 18000 22000
Nash Bill Joe	8 9 12	M M M	19000 17000 14000
Sam Jane Sarah	19 26 28	M F F	24000 36000 37000

Figure 3.1 Exemple montrant la division des enregistrements selon l'attribut âge

Nom	Age	Sexe	Code Postal
Andy Ken Mike	6 6 6	M M M	12000 18000 22000
Nash Bill Joe	9 9 9	M M M	19000 17000 14000
Sam Jane Sarah	24 24 24	M F F	24000 36000 37000

Figure 3.2 Exemple montrant le remplacement de chaque valeur de l'attribut âge par la valeur de la moyenne du groupe

La micro-agrégation est l'une des méthodes les plus utilisées, car elle offre un bon compromis entre simplicité et qualité et elle peut être utilisée comme alternative à la généralisation (Soria-Comas et Domingo-Ferrer, 2016). Cependant, la plupart des algorithmes de micro-agrégation actuellement disponibles ont été conçus pour fonctionner avec de petits ensembles de données, alors que la taille des bases de données actuelles augmente constamment. La manière habituelle de résoudre

ce problème consiste à partitionner de gros volumes de données en fragments plus petits qui peuvent être traités en un temps raisonnable par les algorithmes disponibles. Cette solution est appliquée au détriment de la qualité (Solé *et al.*, 2012).

3.3.4 Anonymisation irréversible ou réversible

L'anonymisation irréversible rend anonymes les données sans permettre de revenir aux données originales. L'anonymisation irréversible se distingue de l'anonymisation réversible sur deux points : le caractère définitif et l'impossibilité de retrouver les données originales (Arfaoui *et al.*, 2020). Considérons par exemple un centre de recherche dans un hôpital qui souhaite mettre en place une application informatique pour la surveillance épidémiologique des maladies infectieuses à déclaration obligatoire comme le SIDA. Le but de cette application serait de disposer à tout moment des informations sur les maladies qui nécessitent une intervention locale urgente. Dans ce cas si les membres du centre de recherche effectuent une anonymisation irréversible, même s'il s'agit de la technique la plus utilisée, elle peut être à l'origine de problèmes non décelés immédiatement. Par exemple, si Pierre décide de donner des échantillons de tissu et qu'il anonymise son nom de manière irréversible, Pierre n'a plus aucun moyen de retrouver des informations sur ses tissus. Plus grave, Pierre n'a aucun moyen de contrôle sur ses données et ne sait donc pas à quelle fin elles seront utilisées. Ceci pose un problème d'éthique. En effet, les personnes qui mettent leurs échantillons et données au service de la science, spécifient dans quel cadre ils doivent être utilisés ou non. Malheureusement, comme ce procédé anonymise les échantillons de manière irréversible, nous n'avons aucune certitude quant à l'utilisation conforme de ces échantillons. Un autre problème que soulève l'irréversibilité, c'est l'incapacité d'être au courant des résultats des recherches et donc, d'en profiter. C'est pourquoi, à des fins épidémiologiques, une autre méthode a été proposée, spécialement dans le domaine

médical : l'anonymisation réversible (Kushida *et al.*, 2012).

L'anonymisation réversible signifie que nous pouvons, après avoir anonymisé une donnée, la récupérer et la désanonymiser. Cette méthode peut s'avérer utile dans le cadre de recherches sur des maladies à risque ou n'ayant actuellement pas de traitement efficace. En effet, les données de personnes atteintes du cancer peuvent être anonymisées afin d'effectuer une recherche sur cette maladie. En cas de résultat positif, il faudra alors désanonymiser les données afin d'administrer le traitement trouvé aux différents patients atteints de cette maladie. Cette méthode peut donc s'avérer fort utile, notamment dans le domaine de la santé. À titre d'exemple, (Tinabo *et al.*, 2009), utilisent la technique de pseudonymisation réversible dans le processus de masquage de l'identité des patients dans une base de données médicales, en utilisant de faux noms afin que les informations relatives à ces individus puissent être traitées sans savoir à qui se rapportent les informations. Cela garantit que l'utilisateur qui agit sous un ou plusieurs pseudonymes peut utiliser une ressource ou un service sans divulguer son identité en raison d'utilisation de faux noms.

3.4 Modèles de protection des données

Un modèle de protection des données permet, dans un contexte donné, de formaliser les garanties qui sont offertes en matière de protection de la vie privée. Les chercheurs ont consacrés beaucoup d'efforts à la protection des données, cela a donné naissance à plusieurs modèles et variantes de modèles. D'après la littérature consultée, il existe plusieurs modèles qui utilisent les méthodes de protection perturbatrices et non perturbatrices pour pouvoir contrecarrer les attaques que peuvent subir les données, que nous avons présentés dans le chapitre 2. Dans cette section, nous allons présenter les modèles les plus cités dans la littérature, à savoir, le k-anonymat, la l-diversité, la t-proximité et la confidentialité ϵ-différentielle.

3.4.1 Le k-anonymat

Le modèle k-anonymat a été proposé par Sweeney. Il se réalise en utilisant deux techniques que nous avons vues précédemment : la généralisation et la suppression (Wang *et al.*, 2009). Sa principale exigence est que chaque sortie des données doit être telle que chaque combinaison de valeurs de QI puisse être indistinctement associée à au moins k individus (Samarati, 2001). Chaque groupe d'enregistrements d'un ensemble de données qu'on ne peut distinguer les uns des autres s'appelle une classe d'équivalence (« equivalence class »). Dans un ensemble qui est considéré comme 2-anonyme les classes d'équivalence possèdent au moins 2 éléments chacun. À titre d'exemple, on peut dire que le tableau 3.4 satisfait le 2-anonymat parce que chaque combinaison des QI «âge» et «sexe» contient au moins 2 occurrences. Tandis que le tableau 3.5 ne satisfait pas le 2-anonymat à cause de l'occurrence unique du dernier enregistrement.

Age	Sexe	Casier Judiciaire
[21, 25] [21, 25]	M M	Vol Possession illégale d'arme à feu
[36, 40] [36, 40]	M M	Meurtre au premier degré Meurtre au premier degré
[15, 20] [15, 20]	F F	Voies de fait grave Voies de fait grave

Tableau 3.4 Tableau qui satisfait le tableau 2-anonymat

Age	Sexe	Casier Judiciaire
[21, 25] [21, 25]	M M	Vol Possession illégale d'arme à feu
[36, 40] [36, 40]	M M	Meurtre au premier degré Meurtre au premier degré
[15, 20] [15, 20]	F F	Voies de fait grave Voies de fait grave
[36, 40]	F	Voies de fait grave

Tableau 3.5 Tableau qui ne satisfait pas le 2-anonymat

En k-anonymat, il est difficile pour un attaquant de déterminer l'identité des individus en collectant les données contenant des informations personnelles. L'un des principaux avantages de la méthode k-anonymat c'est qu'elle vise à protéger les ensembles de données contre les attaques par lien d'enregistrements (Prasser *et al.*, 2014). Cependant, même lorsqu'on prend le soin d'identifier les QI, une solution respectant le k-anonymat peut rester vulnérable à certaines attaques, plus particulièrement les attaques par homogénéité et les attaques par connaissances auxiliaires (Sweeney, 2002).

Prenons l'exemple du tableau 3.4, qui satisfait le 2-anonymat, supposons que l'adversaire connaisse le sexe de Pierre (masculin) et son âge (39 ans). Il peut alors conclure que Pierre a commis un meurtre au premier degré facilement vu que l'attribut sensible «casier judiciaire» est le même pour la classe d'équivalence à laquelle appartient Pierre. Donc il y a toujours un risque pour que l'anonymat d'une personne soit compromis par la collecte d'informations à partir de différentes sources, car le modèle k-anonymat se concentre uniquement sur les QI et non sur les AS. Pour remédier à cette limitation du k-anonymat, (Machanavajjhala *et al.*, 2006) a introduit le modèle l-diversité.

3.4.2 La l-diversité

Définition 3.1. Une classe d'équivalence est dite l-diverse s'il existe au moins l valeurs distinctes pour l'attribut sensible. Une table est considérée comme satisfaisant la l-diversité si chaque classe d'équivalence de la table respecte la l-diversité (Machanavajjhala *et al.*, 2006).

Le modèle l-diversité a été conçu pour traiter certaines faiblesses du modèle k-anonymat, car protéger les identités au niveau de chaque classe d'équivalence n'équivaut pas à protéger les valeurs sensibles correspondantes, en particulier

lorsque les valeurs sensibles sont homogènes au sein d'un groupe. Pour ce faire, le concept de diversité intragroupe de valeurs sensibles est mis en avant dans le cadre de l'anonymisation (Machanavajjhala *et al.*, 2006).

Dans ce modèle, on fait en sorte que, pour un attribut sensible, il existe au moins l valeurs de cet attribut sensible au sein de tout groupe d'individu partageant le même QI. Par exemple, le tableau 3.6 est 3-anonyme et 3-divers car chaque classe d'équivalence contient plusieurs valeurs distinctes de l'AS «casier judiciaire».

Age	Sexe	Casier Judiciaire
[21, 25]	M	Vol
[21, 25]	M	Possession illégale d'arme à feu
[21, 25]	M	Voies de fait grave
[36, 40]	F	Meurtre au premier degré
[36, 40]	F	Voies de fait grave
[36, 40]	F	Vol
[15, 20]	M	Voies de fait grave
[15, 20]	M	Meurtre au premier degré
[15, 20]	M	Extorsion

Tableau 3.6 Exemple de tableau 3-anonymes et 3-diverses

L'un des avantages de la l-diversité est qu'elle garantit la confidentialité même lorsque le détenteur de données ne sait pas quel type de connaissances possède l'adversaire. L'idée principale derrière la l-diversité est l'exigence que les valeurs des attributs sensibles soient distinctes dans chaque groupe (Machanavajjhala *et al.*, 2006). Par contre, selon (Li *et al.*, 2007), l'un des problèmes avec la l-diversité est qu'elle est limitée face aux attaques par similarité, et elle n'assure pas la protection contre les attaques par inférences probabilistes dont celles par dissymétrie (par exemple, imaginez que la proportion de répondants qui ont commis un meurtre dans le groupe est beaucoup plus élevée que dans l'ensemble de données). Un adversaire peut obtenir des informations sur un attribut sensible à condition de disposer d'informations sur la distribution globale de cet attribut.

Pour surmonter les inconvénients de ce modèle, (Li *et al.*, 2007) ont introduit la t-proximité.

3.4.3 La t-proximité

Le modèle t-proximité exige que la distribution d'un attribut sensible dans une classe d'équivalence soit proche de la distribution de cet attribut dans la table globale (c'est-à-dire la distance entre deux distributions ne doit pas dépasser un seuil t). Afin d'intégrer les distances entre les valeurs des attributs sensibles, la métrique Earth Mover's Distance (EMD) permet de mesurer la distance entre deux distributions (Li *et al.*, 2009). L'EMD repose sur la quantité minimale de travail nécessaire pour transformer une distribution en une autre en déplaçant les masses de probabilités entre les distributions. Cela limite efficacement la quantité d'informations spécifiques qu'un attaquant peut apprendre sur un individu.

Selon (Rebollo-Monedero *et al.*, 2009), le modèle t-proximité a tendance à être plus efficace que plusieurs autres méthodes d'exploration de données préservant la confidentialité pour le cas des attributs numériques. Il permet de surmonter les attaques par similarité et les attaques par dissymétrie. Cependant, selon (Kiran et Kavya, 2012), il n'y a aucune procédure informatique qui a été spécifiée pour atteindre cette proximité, chaque attribut étant généralisé indépendamment. De plus, différents niveaux de protection ne peuvent pas être spécifiés pour les attributs sensibles. Les attaques par lien d'attribut ne peuvent pas être empêchées sur les attributs sensibles numériques. L'inconvénient le plus important est que plus t est petit, plus les données se dégradent, car la distribution des valeurs sensibles doit être identique dans tous les groupes QI.

3.4.4 La confidentialité différentielle

La confidentialité ϵ-différentielle a été introduite en 2006 par la chercheuse Cynthia Dwork de Microsoft (Dwork *et al.*, 2006b). Son objectif principal est de cacher la contribution d'un individu particulier à un calcul effectué sur une base de données à laquelle son profil appartient en rajoutant du bruit de manière à s'assurer que la distribution des sorties possibles observées par l'adversaire soit indistinguable que son profil fasse ou non partie de la base de données. Elle rassemble des méthodes qui protègent les données à caractère personnel contre le risque de ré-identification tout en maintenant la pertinence des résultats des requêtes. Pour garantir la confidentialité de systèmes où plusieurs requêtes sont autorisées, la confidentialité différentielle définit une limite au nombre de requêtes. Cette limite, appelée budget de confidentialité, est représenté par le paramètre ϵ. Les budgets de confidentialité empêchent la recréation de données via plusieurs requêtes. Et une fois le budget de confidentialité dépensé ou épuisé, les utilisateurs ne peuvent plus accéder aux données. Ce modèle permet de contrecarrer les attaques par inférences d'attributs et les attaques par appartenance (Dwork, 2011). La confidentialité différentielle permet l'exploitation statistique de données individuelles sans compromettre la vie privée des individus concernés. Elle est obtenue en appliquant un procédé qui introduit de l'aléa dans les données tout en maintenant leur potentielle exploitation. Elle a beaucoup retenu l'attention ces dernières années en tant que modèle général de protection des informations personnelles et a également été proposée comme modèle approprié pour les données sur la santé (Bambauer *et al.*, 2013).

Donnons un exemple devenu classique d'algorithme satisfaisant la confidentialité différentielle. C'est une technique qui a été inventé à l'origine dans le contexte des sondages et qui utilise la notion de réponse randomisée. À titre d'exemple, supposons un groupe de personnes où on trouve à la fois des délinquants et des

innocents. Supposons que l'on cherche à estimer la proportion de délinquants. Ceux-ci ne se dénonceront jamais si cette information sensible peut leur être associée personnellement. L'approche classique consiste à demander à chaque personne de jouer à pile ou face. Si l'on obtient pile, l'individu répond sincèrement. Si l'on obtient face, on doit relancer la pièce pour répondre au hasard à la question : face donne la réponse « oui, je suis un délinquant » et pile donne « non, je suis innocent ». Dans ce cas les délinquants n'auront donc plus peur d'avouer, vu qu'ils savent que beaucoup d'innocents diront aussi « oui, je suis un délinquant ».

La confidentialité différentielle peut être obtenue aussi en ajoutant un «bruit» (une valeur aléatoire) aux résultats de toutes les requêtes agrégées pour protéger les entrées individuelles sans modifier de manière significative le résultat. L'un des algorithmes les plus simples est le mécanisme de Laplace, qui peut post-traiter les résultats de requêtes agrégées en garantissent que l'attaquant ne peut pratiquement rien apprendre de plus sur un individu qu'il n'apprendrait si le dossier de cette personne était absent de l'ensemble de données (Holohan *et al.*, 2018).

Apple et Google utilisent des techniques de confidentialité différentielles dans iOS et Chrome respectivement. Google a récemment publié une version open source de sa bibliothèque de confidentialité différentielle utilisée par certains de ses produits. La bibliothèque est conçue pour aider les développeurs à créer des produits qui utilisent des données agrégées anonymisées de manière à préserver la confidentialité (Garfinkel *et al.*, 2018). L'exemple de plus grande envergure couramment cité, est celui du prochain recensement (Census) américain. L'équipe du recensement (le Census Bureau américain), est la plus grande agence de statistiques aux États-Unis. Ils utilisent le modèle de confidentialité différentielle pour protéger la confidentialité des répondants lors du recensement 2020 qui sera rendu public sous peu. Selon cette nouvelle approche, le Bureau du recensement a veillé

à ce qu'aucune publication de recensement ne permette de relier des réponses de recensement à des personnes spécifiques. Le système de traitement des données du recensement de 2020 commence par tenter de collecter des données auprès de toutes les personnes vivant aux États-Unis par divers moyens, dont un instrument en ligne, un système de réponse vocale par téléphone, un formulaire qui peut être envoyé par la poste et des «recenseurs» qui se déplacent de maison en maison pour le suivi des non-réponses (NRFU). Ces données confidentielles sont collectées et traitées pour créer le fichier CUF («Census Unedited File»), qui contiendra une liste bloc par bloc de chaque personne aux États-Unis. Ce fichier est utilisé pour produire le fichier édité du recensement «Census Edited File»(CEF). Suite à la création du CEF, les données du répondant sont acheminées vers une application spécialement conçue, appelée «Disclosure Evitement System» (DAS). Le DAS utilisera un nouveau mécanisme de confidentialité différentielle pour ajouter du bruit au CEF, produisant le fichier de détail des micro-données (MDF) que le système de tabulation du Census Bureau utilisera pour créer les produits de données traditionnels (Garfinkel et Leclerc, 2020).

La garantie robuste amenée par la confidentialité différentielle se paye par l'ajout d'un bruit conséquent au résultat des requêtes et des analyses. Selon (Bun et Steinke, 2016), dans sa forme la plus simple, la confidentialité différentielle (pure) est paramétrée par un nombre réel $\epsilon > 0$, qui contrôle le niveau de «perte de confidentialité [1]» qu'un individu peut subir lorsqu'un calcul (c'est-à-dire une tâche d'analyse statistique de données) est effectué sur ses données. Une caractéristique particulière de la confidentialité différentielle est qu'elle se dégrade de manière régulière et prévisible sous la composition de calculs multiples. De ce fait (Dwork

1. La perte de confidentialité est une variable aléatoire qui quantifie la quantité d'informations révélées sur un individu par un calcul impliquant ses données ; elle dépend du résultat du calcul, de la façon dont le calcul a été effectué et des informations que l'individu veut cacher.

et al., 2006a) propose une relaxation largement utilisée de la confidentialité différentielle pure : la confidentialité approximative ou (ϵ, δ)-différentielle, qui garantit essentiellement que la probabilité qu'un individu subisse une perte de confidentialité dépassant ϵ est limitée par δ. Pour un δ suffisamment petit, la confidentialité approximative (ϵ, δ)-différentielle fournit une norme de protection de la vie privée comparable à la confidentialité ϵ-différentielle pure, tout en permettant souvent d'effectuer des analyses beaucoup plus utiles.

3.5 Synthèse

Dans ce chapitre, nous avons présenté premièrement les méthodes de protection de données perturbatrices et non perturbatrices. Le tableau 3.7, permet d'avoir un récapitulatif des méthodes de protection de données avec leurs formes d'anonymisation. Et le tableau 3.8, extrait (Fung *et al.*, 2010a), permet de catégoriser les modèles de protection de données présentés en fonction des types d'attaques auquel ils peuvent faire face.

Méthodes de protection	Perturbatrices	Non perturbatrices	Réversibles	Irréversibles
Généralisation		✔		✔
Suppression		✔		✔
Pseudonymisation		✔	✔	
Randomisation	✔			✔
Permutation	✔			✔
Microagrégation	✔			✔

Tableau 3.7 Méthodes de protection avec leurs formes d'anonymisation

Suite aux avantages et inconvénients de chaque modèle, nous pouvons conclure que le choix d'une seule technique ne suffit pas pour protéger l'anonymat et la confidentialité des utilisateurs. Cependant dans le tableau 3.8, nous pouvons constater que la confidentialité différentielle permet de protéger contre tout type d'attaque.

Modèles de protection	Modèles d'attaque			
	Lien d'enregistrements	Inférence d'attributs	Appartenance	Inférences probabilistes
k-Anonymat	✓			
l-Diversité	✓	✓		
t-Proximité	✓	✓		✓
ε-Différentielle	✓	✓	✓	✓

Tableau 3.8 Modèles de protection de données avec les modèles d'attaques (Fung *et al.*, 2010a)

En revanche, cela peut parfois se faire au détriment de l'utilité des données si le niveau de bruit à rajouter est trop important. Vu l'importance de protéger la confidentialité et l'anonymat des utilisateurs, il est crucial de proposer des solutions avantageuses afin de minimiser le risque de divulgations d'informations personnelles.

Dans le cadre de notre projet d'anonymisation de données, nous allons à partir d'une certaine démarche de conception, appliquer de concert les méthodes et modèles de protection de données. Le chapitre suivant présente notre démarche de conception.

CHAPITRE IV

ANALYSE ET CONCEPTION DE PROCOM

4.1 Introduction

Nous avons vu précédemment qu'il est crucial de proposer des solutions avantageuses afin de minimiser le risque de divulgations d'informations personnelles des utilisateurs. Notre objectif consiste à protéger l'anonymat et la confidentialité des utilisateurs d'un service livré par une plateforme mobile. Cela implique de mettre en oeuvre une démarche d'anonymisation que nous avons appelée PROCOM. Elle va nous permettre de combiner les techniques et les méthodes de protection de données que nous avions identifiées dans notre revue de littérature.

Ce chapitre présente la conception et l'analyse de PROCOM. Dans un premier temps, nous débuterons par la formulation du problème. Ensuite dans un deuxième temps, nous présenterons le modèle «Privacy by Design» qui permet de protéger la confidentialité des utilisateurs dès le début d'un projet et en dernier lieu nous présenterons l'ensemble des étapes à suivre pour pouvoir anonymiser les données.

4.2 Formulation du problème

Supposons qu'un groupe de chercheurs d'université développent une application mobile qui collecte des données confidentielles des utilisateurs dans le cadre d'un projet de recherche. Par la suite, ils décident de publier les données collectées à

des tiers à des fins de recherche ou d'étude. Si ces données ne sont pas anonymes ou mal anonymisées n'importe quel attaquant pourra en déduire des informations personnelles et sensibles sur les utilisateurs. Selon (Patil *et al.*, 2017), même si l'identité n'est pas publiée, sur la base de certains attributs informatifs et de données accessibles au public, les attaquants peuvent accéder aux informations qui sont censées être privées, et le défi majeur, tout en préservant la vie privée d'un individu, est de conserver les données publiées utiles pour la recherche et l'analyse. Car, en anonymisant les données on risque de perdre des informations potentiellement utiles au projet de recherche. De ce fait, il faut avoir un mécanisme ou une démarche de conception qui permettrait d'avoir une anonymisation au cas par cas et adaptée aux usages prévus.

4.2.1 Modèle de confidentialité dès la conception «Privacy by Design»

D'après la littérature consultée, il existe plusieurs modèles qui permettent de protéger la confidentialité des utilisateurs. Par exemple, il existe le modèle «Privacy by Design» (PbD), qui consiste en un ensemble de principes qui peuvent être appliqués dès le début du développement d'un système tout en permettant de réduire les risques d'atteinte à la vie privée. Ce modèle est au coeur du «Règlement Général pour la Protection des Données» (RGPD) (Levallois-Barth, 2018), qui impose que les règles établies par le RGPD soient intégrées dès la conception d'un projet, produit, service, outils de récolte utilisant des données personnelles. L'objectif est d'éviter un risque de violation des données en mettant en place des mesures de protection le plus en amont possible et ne nécessitant aucune action particulière de l'utilisateur. Ces protections sont un pré-requis pour se mettre en conformité avec le règlement.

Selon (Langheinrich, 2001), PbD est un remède à l'insuffisance des moyens actuels de protection de la vie privée, mais aussi une démarche indispensable, car il est très

difficile d'améliorer la protection de la vie privée dans des systèmes qui n'ont pas été conçus selon cette exigence. PbD repose sur sept (7) principes fondamentaux (Cavoukian *et al.*, 2009) :

1. Il faut prendre des mesures pro-actives et non réactives, des mesures préventives et non correctives. PbD anticipe et empêche les événements envahissant la vie privée. Il n'attend pas que les risques se matérialisent et n'offre pas de recours pour résoudre les infractions une fois qu'elles se sont produites, il vise à les empêcher de se produire. PbD commence par une reconnaissance explicite de la valeur et des avantages de l'adoption pro-active de solides pratiques de protection de la vie privée, à un stade précoce et de manière cohérente (par exemple, en prévenant les violations de données (internes) dès le départ). Cela implique :
   - Un engagement clair, au plus haut niveau, à fixer et à appliquer des normes élevées de confidentialité - généralement plus élevées que les normes établies par les lois et réglementations mondiales.
   - Un engagement de confidentialité qui est manifestement partagé par les communautés d'utilisateurs et les parties prenantes, dans une culture d'amélioration continue.
   - Des méthodes établies pour reconnaître les conceptions de confidentialité médiocres, anticiper les mauvaises pratiques et résultats en matière de confidentialité, et corriger les impacts négatifs, bien avant qu'ils ne se produisent de manière pro-active, systématique et innovante.
2. Il faut assurer la confidentialité par défaut. PbD offre le maximum de confidentialité en garantissant que les données personnelles sont automatiquement protégées dans tout système informatique ou pratique commerciale donnée. Si un individu ne fait rien, sa vie privée reste intacte. Aucune ac-

tion n'est requise de la part de l'individu pour protéger sa vie privée. Ce principe est implémenté par les pratiques suivantes :

— Spécification des finalités : les finalités pour lesquelles les informations personnelles sont collectées, utilisées, conservées et divulguées doivent être communiquées à l'individu (personne concernée) au moment ou avant le moment où les informations sont collectées. Les finalités spécifiées doivent être claires, limitées et adaptées aux circonstances.

— Limitation de la collecte : la collecte de renseignements personnels doit être juste, légale et limitée à ce qui est nécessaire aux fins spécifiées.

— Minimisation des données : la collecte d'informations personnellement identifiables doit être réduite au strict minimum. La conception des programmes, des technologies de l'information et des communications et des systèmes doit commencer par des interactions et des transactions non identifiables, par défaut. Dans la mesure du possible, l'identifiabilité, l'observabilité et la liaison des informations personnelles doivent être minimisées.

— Limitation de l'utilisation, de la conservation et de la divulgation : l'utilisation, la conservation et la divulgation des renseignements personnels doivent être limitées aux fins pertinentes identifiées à l'individu, pour lesquelles il a consenti, sauf disposition contraire de la loi. Les informations personnelles ne seront conservées que le temps nécessaire pour atteindre les objectifs énoncés, puis détruites en toute sécurité.

  Lorsque le besoin ou l'utilisation des informations personnelles n'est pas clair, il doit exister une présomption de confidentialité et le principe de précaution doit s'appliquer : les paramètres par défaut doivent être ceux qui protègent le plus la vie privée.

3. La confidentialité est intégrée dans la conception et l'architecture des systèmes informatiques. Elle n'est pas mise en place comme un complément, après coup. Le résultat est que la confidentialité devient un élément essentiel de la fonctionnalité de base fournie. La confidentialité fait partie intégrante du système, sans diminuer la fonctionnalité. Elle est intégrée au système, par défaut. Dans la mesure du possible, des évaluations détaillées de l'impact sur la vie privée et des risques devraient être effectuées et publiées, documentant clairement les risques pour la vie privée et toutes les mesures prises pour atténuer ces risques, y compris l'examen d'alternatives et la sélection de paramètres.

4. La prise en compte de la vie privée ne doit pas empêcher la mise en oeuvre d'autres fonctionnalités, mais doit être un avantage concurrentiel. La protection de la vie privée doit être considérée avec une approche «gagnant-gagnant». Par exemple, la prise en compte de la vie privée ne doit pas empêcher un haut niveau de sécurité. Il est possible de réaliser plusieurs objectifs à la fois sans les compromettre. Lors de l'intégration de la confidentialité dans une technologie, un processus ou un système donné, cela doit être fait de telle manière que toutes les fonctionnalités ne soient pas altérées et, dans la mesure du possible, que toutes les exigences soient optimisées.

5. Il faut assurer la sécurité de bout en bout, pendant toute la période de conservation des renseignements. La confidentialité, ayant été intégrée au système avant la collecte du premier élément d'information, s'étend tout au long du cycle de vie des données concernées, du début à la fin. Cela garantit qu'à la fin du processus, toutes les données sont détruites en toute sécurité, en temps opportun.

    — Les entités doivent assumer la responsabilité de la sécurité des informations personnelles (généralement proportionnelles au degré de sen-

sibilité) tout au long de leur cycle de vie, conformément aux normes élaborées par des organismes d'élaboration de normes reconnus.

— Les normes de sécurité appliquées doivent garantir la confidentialité, l'intégrité et la disponibilité des données personnelles tout au long de leur cycle de vie, y compris, entre autres, des méthodes de destruction sécurisée, un cryptage approprié et des méthodes de contrôle d'accès et d'enregistrement solides.

6. Il faut assurer la visibilité et la transparence. Grâce à la protection intégrée de la vie privée, tous les intervenants seront assurés que sans égard aux pratiques ou aux technologies employées, le système fonctionne conformément aux promesses et aux objectifs établis, sous réserve d'une vérification indépendante. Les éléments et le fonctionnement du système demeurent visibles et transparents, tant pour les utilisateurs que pour les fournisseurs.

7. Il faut respecter la vie privée des utilisateurs. Par-dessus tout, PbD exige que les architectes et les opérateurs gardent les intérêts de l'individu au premier plan en offrant des valeurs par défaut élevées en matière de respect de la vie privée. Permettre aux personnes concernées de jouer un rôle actif dans la gestion de leurs propres données peut être le moyen de contrôle le plus efficace contre les abus et les utilisations abusives de la vie privée et des données personnelles. Le respect de la vie privée des utilisateurs est pris en charge par les pratiques équitables en matière d'information :

— Consentement : Le consentement libre et spécifique de l'individu est requis pour la collecte, l'utilisation ou la divulgation de renseignements personnels, sauf autorisation contraire de la loi. Plus les données sont sensibles, plus la qualité du consentement requis est claire et précise. Le consentement peut être retiré à une date ultérieure.

— Exactitude : Les informations personnelles doivent être aussi exactes, complètes et mise à jour aussi souvent que nécessaire pour atteindre les objectifs spécifiés.

— Accès : Les individus doivent avoir accès à leurs informations personnelles et être informés de leurs utilisations et divulgations. Les individus doivent être en mesure de contester l'exactitude et l'exhaustivité des informations et les faire modifier le cas échéant.

— Conformité : Les organisations doivent établir des mécanismes de plainte et de recours, et communiquer des informations à leur sujet au public.

PbD s'applique aux nouvelles technologies, et notamment aux systèmes informatiques et aux infrastructures des réseaux. Ses principes peuvent s'appliquer à tous les types de renseignements personnels, mais ils devraient l'être avec une rigueur particulière pour les données sensibles telles que les renseignements médicaux et financiers. Plus les données sont sensibles, plus les mesures de protection de la vie privée tendent à être strictes. Quand on fait une anonymisation a priori nous devons collecter que les données adéquates, pertinentes et limitées à la réalisation de la finalité choisie. Cependant, si l'anonymisation a priori n'a pas été effectuée au début du développement du projet nous devons alors avoir une méthode ou un processus qui va permettre d'anonymiser la données a posteriori. Dans la section suivant nous allons présenter de manière détaillée les cinq étapes du processus d'anonymisation de PROCOM.

4.3 Présentation générale de PROCOM

Pour aider les chercheurs dans le choix d'une ou de plusieurs techniques dans le but de protéger la confidentialité et l'anonymat des utilisateurs d'un service livré par une plateforme mobile, PROCOM met à leur disposition un ensemble d'étapes

à suivre, chacune des étapes fournissant un ensemble d'informations qui peuvent éclairer leurs choix.

Comme le montre la figure 1.2, PROCOM compte cinq (5) étapes principales. À l'étape 1 on s'intéresse aux objectifs d'anonymisation des chercheurs afin de préserver les besoins de l'application. À l'étape 2, on fait une vérification ou une mise à l'épreuve de la protection accordée aux données de départ. À l'étape 3, on précise quelles données seront anonymisées. L'étape 4 présente les différents niveaux ou formes d'anonymisation et l'étape 5 permet de choisir les techniques ou méthodes de protection de données. Toutes ces étapes peuvent se répéter jusqu'à ce que le détenteur des données soit satisfait des résultats obtenus.

4.4 Première étape : évaluation de la préservation des besoins de l'application

La première étape avant d'anonymiser les données est de pouvoir identifier les objectifs d'anonymisation du projet de recherche. Les détenteurs de données veulent souvent collecter le plus de données possible tout en voulant un niveau d'anonymisation maximale. Il faut donc faire un choix d'anonymisation en fonction des besoins et du niveau de protection souhaité. Nous avons vu dans la section précédente qu'en anonymisant les données on risque de perdre des informations qui seront peut-être utiles dans le cadre de chaque projet de recherche spécifique. Pour pouvoir évaluer la préservation des besoins de l'application nous avons catégorisé les besoins en fonction de la qualité et de la confidentialités des données.

4.4.1 La qualité des données

Comme mentionné plus haut nous devons faire un compromis entre la qualité des données et l'anonymisation. Cependant le détenteur des données peut définir, quelles sont les données qui sont essentielles pour la recherche. Dans le but d'appliquer des techniques de protection de données qui n'affectent pas négativement

la qualité de ces données, il est important que le détenteur des données dise précisément quelle donnée est importante ou non et aussi la manière dont il compte utiliser cette donnée. Par exemple, pour une donnée d'âge, a t-il besoin du nombre de participants qui ont entre [18-25] ans ? Si oui, nous pourrons généraliser la donnée âge en plusieurs intervalles au lieu de conserver l'âge exact des participants. Le détenteur de données peut aussi réduire à cinq ans d'intervalle [18-22] pour ne pas perdre trop en qualité de donnée.

4.4.2 La confidentialité des données

Au niveau de la confidentialité des données, il est important de savoir si les données recueillies, considérées en elles-mêmes ou croisées avec d'autres données, pourraient permettre l'identification d'individus ? Si oui, qui aura accès aux données ? Les données seront-elles sécurisées afin d'éviter les fuites de données ? En effet, si les données seront croisées avec d'autres données à des fins d'utilisation secondaire, elles seront alors fortement exposées aux différents types d'attaques mentionnées dans le chapitre 3. Dans ce cas, pour préserver la confidentialité des données, il faudra que le détenteur des données ait accès aux autres données pour pouvoir évaluer le risque d'identification des participants.

4.4.3 La protection et l'utilité des données

Dans le but de fournir un niveau satisfaisant de protection des données, les techniques de protection de données altèrent les données de telle sorte qu'aucun individu ne puisse être identifié de manière unique. Par exemple, une table peut être trivialement généralisée à une seule classe d'équivalence en supprimant tous les QI. Cette approche assure un maximum de protection, cependant les données qui en résultent sont inutiles. Vu que les données anonymisées doivent permettre d'effectuer des tâches de recherche et d'analyse, il est important d'assurer un bon

compromis entre la protection de la vie privée et l'utilité des données. Dans le chapitre 5, nous présentons la métrique $ILoss$ qui permet de mesurer la perte d'utilité ou perte d'information de la généralisation d'une valeur spécifique à une valeur générale (par exemple quand on passe de l'âge exact à une tranche d'âge).

4.5 Deuxième étape : vérification de la protection accordée aux données

Avant d'examiner la possibilité d'appliquer une ou plusieurs méthodes ou techniques de protection de données qui ont le moins de risque de dégrader les données de départ, nous devons dans un premier temps pouvoir vérifier ou mettre à l'épreuve la protection accordée pour ces données. Nous pouvons le faire manuellement à l'aide des outils disponibles en ligne (Google, Facebook, etc.) afin de vérifier si on est dans l'impossibilité de retrouver l'identité d'une personne parmi ces données, ce qui requiert de se mettre dans la peau d'un attaquant. Cette vérification sera plus ou moins efficace selon que le vérificateur est habile à retrouver l'identité ou les informations sensibles.

Cette vérification peut se faire en trois (3) étapes. Premièrement, nous devons choisir un individu au hasard dans les données de départ. Deuxièmement, collecter toutes les informations disponibles dans ces données pour cet individu afin d'avoir un profil unique et troisièmement d'essayer de croiser ces données grâces à des outils disponible en ligne. Il se peut qu'on ne trouve aucun résultat soit parce que la table protège adéquatement les données de cet individu ou que la recherche ait été mal faite. Si des résultats sont obtenus, quelle que soit l'exactitude des résultats obtenus, la vie privée de cet individu est compromise.

Cette méthode est limitée car elle consiste à choisir un individu au hasard et faire une recherche manuelle. Cependant si on doit anonymiser une base de données de plusieurs milliers de profils, il existe d'autres travaux connexes qui pourraient utilement être exploités pour compléter l'automatisation de cette étape. Par exemple,

le logiciel «open source» ARX Data Anonymization Tool permet d'anonymiser les données sensibles. Comme le montre la figure 4.1, il comporte trois phases : la configuration, l'exploration et l'analyse.

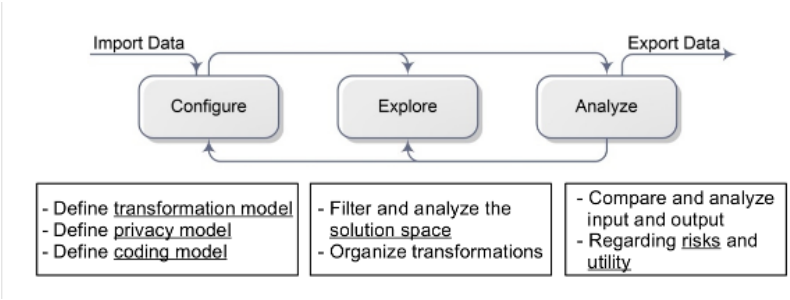


Figure 4.1 Processus d'anonymisation d'ARX.

Lors de la phase d'analyse, ARX permet de faire la vérification de la protection accordée aux données de départ en mesurant les risques de ré-identification ainsi que des estimations de l'unicité des enregistrements. Par exemple dans la figure 4.2, des seuils peuvent être fournis pour le risque le plus élevé de tout enregistrement, pour les enregistrements qui présentent un risque supérieur à ce seuil et pour la fraction moyenne des enregistrements qui peuvent être ré-identifiés avec succès.

Figure 4.2 Estimation de risques fournie par ARX.

Plus de détails peuvent être trouvés dans le manuel d'utilisation d'ARX sur son

site officiel [1]. Il décrit toutes les fonctionnalités de l'outil à l'aide de textes et de vidéos.

4.6 Troisième étape : comment repérer ou identifier les données à anonymiser ?

L'anonymisation des données consiste à modifier le contenu des enregistrements avant de les publier. Nous devons être en mesure d'identifier les données à généraliser, à supprimer ou à permuter. C'est-à-dire de pouvoir identifier les attributs QI, AS, ANS et de supprimer les IE.

4.6.1 Identifier les attributs QI, AS, ANS et IE

L'étape 3 de PROCOM est une étape importante dans le processus d'anonymisation des données, elle nous permet d'identifier tous les attributs QI, AS et ANS mais aussi d'identifier et de supprimer les attributs IE. D'après la littérature consultée, un QI fait référence à un sous-ensemble d'attributs qui peuvent identifier de manière unique la plupart des enregistrements d'une table.

Selon (Motwani et Xu, 2007), savoir comment classer ces attributs dans un tableau de données est un défi auquel les détenteurs de données sont confrontés. De plus, une fois les IE supprimés et les QI déterminés, les attributs restants sont regroupés en AS et ANS en fonction de leur sensibilité. Cette étape peut se faire aussi de manière automatique avec l'outil ARX. Il existe aussi d'autres travaux de recherche visant la détection automatique des attributs formant les QI. Citons (Agrawal *et al.*, 2014) qui combine une analyse statique et une analyse dynamique des programmes qui accèdent aux données sensibles. (Akoka *et al.*, 2014) fournissent un processus semi-automatique pour la détection des données sensibles en utilisant des techniques d'analyse syntaxique et sémantique.

---

1. http ://arx.deidentifier.org/overview/

4.7 Quatrième étape : identifier le type d'anonymisation

Après avoir supprimé les IE, et identifié les QI, AS et ANS, nous devons maintenant faire le choix d'un type d'anonymisation, c'est-à-dire assurer une transformation irréversible ou réversible des variables qui permettent l'identification d'un individu (nom, prénom, date de naissance, sexe).

4.7.1 Anonymisation irréversible ou réversible

L'anonymisation irréversible et réversible ont des objectifs distincts. L'anonymisation réversible se prête aux situations qui nécessitent ou permettent un retour en arrière afin d'être en mesure de retrouver la personne concernée par la donnée. Au contraire, l'anonymisation irréversible ne s'inscrit pas dans une démarche ultérieure de ré-identification et ne permet pas un retour en arrière. Ces deux types d'anonymisation répondent donc à deux objectifs distincts : conserver ou non le caractère personnel des informations.

L'anonymisation irréversible permet de conserver les informations anonymisées après la réalisation de la finalité du traitement, tandis que l'anoymisation réversible ne permet pas de conserver les informations nominatives au-delà de la durée de conservation prévue dans le traitement. Lors de l'anonymisation réversible, il faut être vigilant dans la mesure où une ré-identification peut intervenir à partir d'informations partielles (par exemple, la combinaison des données ville et date de naissance peut être suffisante). L'anonymisation irréversible est à privilégier chaque fois que la connaissance de la valeur d'une donnée personnelle n'est pas essentielle pour le bon fonctionnement des systèmes ou des traitements des intermédiaires qui l'exploitent (Arfaoui *et al.*, 2020).

L'anonymisation réversible trouve surtout son utilité dans le domaine de la santé, où par exemple les données personnelles des patients qui sont atteints d'une mala-

die incurable peuvent être anonymisées de façon réversible dans le but d'effectuer des recherches sur cette maladie, pour par la suite administrer le traitement trouvé aux patients atteints de la maladie et qui pourront être ré-identifiés. Nous devons donc faire le choix d'un type d'anonymisation en fonction des besoins (voir figure 4.3).

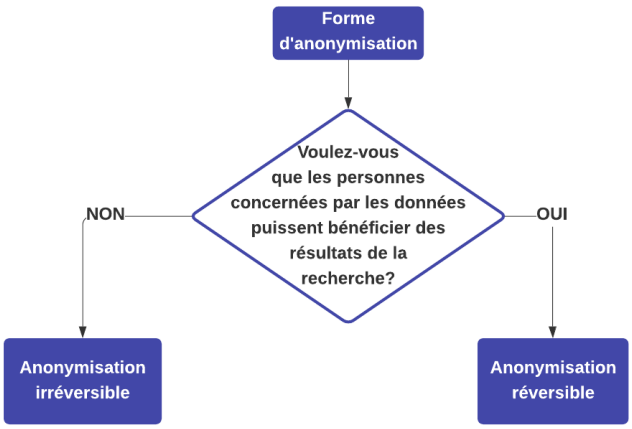


Figure 4.3 Arbre de décision pour pouvoir choisir une forme d'anonymisation.

4.8 Cinquième étape : quels méthodes et techniques de protection de données choisir ?

Dans le chapitre 3, nous avons présenté les méthodes et techniques de protection des données les plus utilisées de nos jours. Cependant dans cette section, à titre informatif, nous avons fait le choix de présenter quelques méthodes et techniques qui s'adaptent mieux entre elles dans le but de guider vers les meilleures combinaisons possibles selon les besoins spécifiques.

Comme le montre la figure 4.4 nous pouvons faire le choix entre les méthodes de type irréversible ou réversible. Nous pouvons aussi choisir de combiner ou pas les méthodes de protection perturbatrices et non perturbatrices représentées en fond

bleu et les différentes techniques en fond vert. Toutes ces méthodes et techniques peuvent être combinées ou utilisées séparément.

4.8.1 Combinaison 1 : généralisation et suppression

La généralisation permet de remplacer les QI par des valeurs moins spécifiques. L'avantage d'utiliser cette méthode est qu'elle préserve la véracité des informations. La suppression, quant à elle, permet de retirer des données qui présentent un risque élevé de ré-identification.

4.8.2 Combinaison 2 : randomisation et permutation

La randomisation permet d'ajouter du bruit aux données afin de masquer la valeur des attributs des enregistrements. Cette méthode peut être combinée avec la méthode de permutation car le bruit peut être introduit en interchangeant des valeurs de variables sensibles entre des enregistrements individuels.

4.8.3 Combinaison 3 : combinaison 2 et k-anonymat

La combinaison 1 (généralisation et suppression) va permettre de renforcer le k-anonymat. La combinaison 2 associée au k-anonymat va transformer les QI de telle sorte qu'il existe au moins k individus qui aient la même valeur de QI. Ainsi cette combinaison va protéger les données contre les attaques par liens d'enregistrements.

4.8.4 Combinaison 4 : combinaison 3 et micro-agrégation

La micro-agrégation contribue au renforcement du k-anonymat, elle fait en sorte que les enregistrements correspondent à des groupes d'au moins k individus appelés micro-agrégats. Pour pouvoir satisfaire la confidentialité, la technique de micro-agrégation va remplacer les valeurs originales par une moyenne.

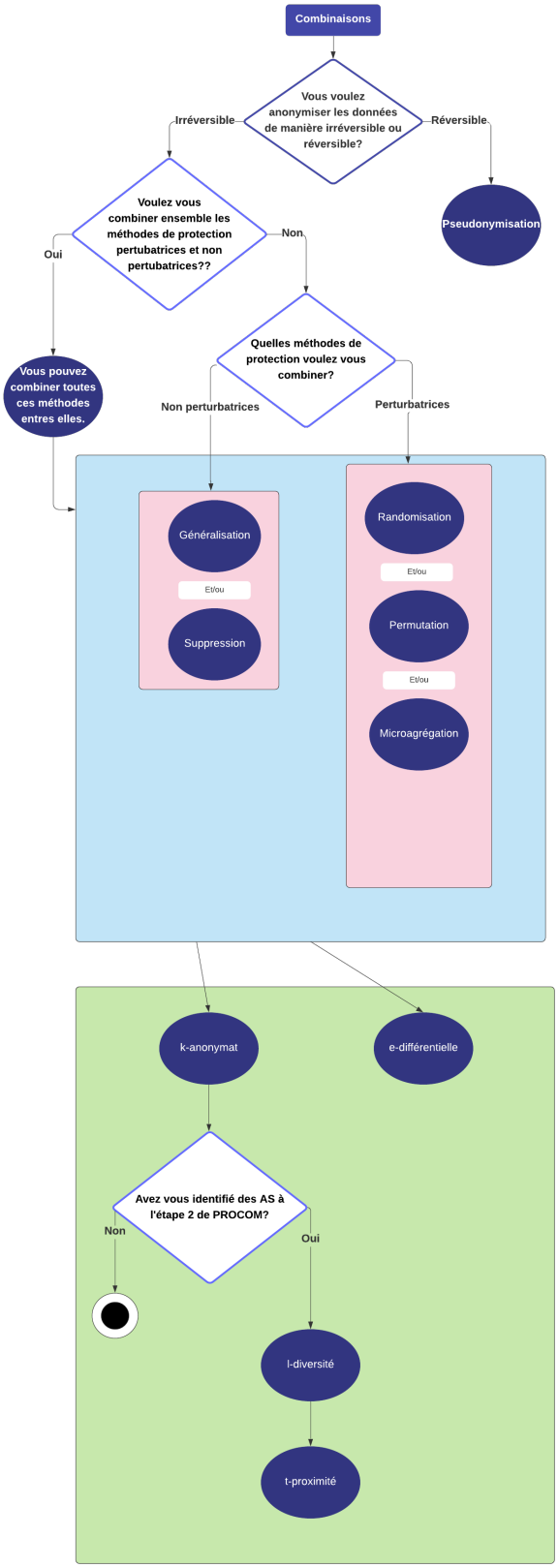


Figure 4.4 Arbre de décision montrant les différentes combinaisons possibles.

4.8.5 Combinaison 5 : combinaison 4 et confidentialité différentielle

Le k-anonymat et la confidentialité différentielle adoptent des approches de limitation de la divulgation qui sont fondamentalement différentes. Toutefois il existe dans la littérature, quelques travaux qui lient les deux modèles. Par exemple, (Soria-Comas *et al.*, 2014), montrent que la quantité de bruit nécessaire pour respecter la confidentialité différentielle peut être considérablement réduite si la requête est exécutée sur une version k-anonyme de l'ensemble de données, obtenue par micro-agrégation de tous les attributs (au lieu de l'exécuter sur les données brutes).

4.8.6 Combinaison 6 : combinaison 4 et l-diversité

Si à l'étape 3 de PROCOM nous avions identifié des AS nous pouvons renforcer la combinaison 3 en la combinant à la technique de l-diversité, car la technique de $k-$anonymat est sans effet sur les AS. Ainsi, cette technique, fait en sorte que, pour un AS, il existe au moins l valeurs distinctes de cet attribut sensible au sein de tout groupe d'individu partageant le même QI.

4.8.7 Combinaison 7 : combinaison 6 et t-proximité

Bien que les données puissent être protégées par la combinaison 5, il est cependant possible qu'un attaquant puisse obtenir des informations sur les AS dès qu'il dispose d'informations sur la distribution globale de ces attributs. Pour cela nous pouvons la combiner à la technique de t-proximité. Cette technique introduit un rapprochement entre deux distributions et propose que cette distance ne dépasse pas le seuil t, tel que détaillé dans le chapitre 3.

4.9 Synthèse

Dans ce chapitre nous avons présenté de manière générale notre approche PROCOM. Nous avons vu que l'anonymisation pouvait être faite a priori (lors de la collecte de données), ou a posteriori (après la collecte de données). Nous avons vu qu'il est préférable d'effectuer une anonymisation a priori en collectant les données dont on a vraiment besoin. Nous avons présenté le modèle PbD qui est un ensemble de principes qui peuvent être appliqués dès le début du développement d'un système.

Nous avons vu qu'une publication imprudente des attributs QI, AS et IE entraînera une exposition de la vie privée, qu'il faut savoir comment classer ces attributs. Nous avons aussi vu que l'anonymisation irréversible est la méthode d'anonymisation par excellence, du point de vue du niveau de protection atteint, mais pas nécessairement du point de vue de l'utilité des données, cependant dans le domaine médical il est préférable d'effectuer une anonymisation réversible afin que les personnes concernées par les données puisse bénéficier des résultats des recherches scientifiques.

Et dans le but d'avoir une anonymisation efficace nous avons aussi suggéré un ensemble de combinaisons possibles des méthodes et techniques de protection de données en fonction des besoins spécifique à chaque application. Dans le prochain chapitre, nous allons présenter l'application de notre démarche PROCOM à un cas pratique.

CHAPITRE V

MISE EN OEUVRE DE PROCOM

5.1 Introduction

Après avoir défini notre démarche de conception PROCOM, il est pertinent de pouvoir appliquer notre démarche à un cas pratique. Dans cette optique, nous avons eu l'autorisation du chercheur Monsieur André Mondoux responsable du groupe de recherche sur l'information et la surveillance au quotidien (Grisq) de l'école des médias à l'Université du Québec À Montréal (UQAM), pour pouvoir analyser de manière anonyme les données d'un projet de recherche baptisé USAGES MOBILES. Ce projet vise à étudier les pratiques et les perceptions des utilisateurs de médias socionumériques dans un contexte de mobilité. Pour pouvoir collecter les données, une phase de pré-test a été lancée en février 2018 et s'est terminée en mars 2018. Il y a eu plus d'une 20 vingtaines de participants tous âgés d'au moins 18 ans.

Dans ce chapitre, nous allons présenter le projet de recherche USAGES MOBILES, par la suite nous allons appliquer notre démarche de conception aux données collectées par USAGES MOBILES et afin d'exploiter un autre aspect de PROCOM qu'USAGES MOBILES ne permettait pas de prendre en compte nous avons monté un scénario secondaire avec des données fictives.

5.2 Application mobile de traçabilité

USAGES MOBILES est un projet qui vise à mieux comprendre les dynamiques de surveillance via les fonctions de traçabilité des usagers en temps réel ainsi, que leurs rôles par rapport aux pratiques et aux perceptions des usagers. Le but principal est de pouvoir documenter et comprendre quelles applications sont utilisées en contexte de mobilité, le moment et la durée d'utilisation et le volume de données échangées, le tout en lien avec les données de géolocalisation.

Cette application est installée sur un appareil mobile Android, qui permet de compiler en temps réel l'utilisation des médias socionumériques des utilisateurs. Elle recueille les informations suivantes : applications utilisées, fonctions activées (sans leur contenu), temps d'utilisation, position géolocalisée, la fréquence du trafic échangé et volume (quantité de Mo) des données produites et reçues. Cependant, le contenu (texte saisi au clavier, images, vidéos, etc.) n'est pas recueilli. Et pour avoir accès à USAGES MOBILES, les participants doivent avoir en leur possession un appareil mobile Android, ils doivent se rendre vers un site web fourni par le Grisq, pour pouvoir signer un formulaire d'information et de consentement et remplir un questionnaire (voir annexe B).

Après avoir répondu aux questions du formulaire, le participant va recevoir un courriel avec un identifiant universel unique (UUID) à six (6) chiffres ainsi qu'un lien pour qu'il puisse télécharger l'application USAGES MOBILES à partir de son appareil mobile Android.

Le UUID est requis pour pouvoir accéder à l'application. Par la suite le participant doit autoriser USAGES MOBILES à accéder aux notifications, à la géolocalisation et aux informations d'accessibilité (sonnerie, vibration, ouverture ou fermeture d'écran, etc..) de l'appareil Android.

À partir de cet instant, USAGES MOBILES commence à collecter les données d'usage de l'utilisateur en arrière-plan. Un participant peut en tout temps arrêter de participer au projet de recherche soit en cliquant sur le bouton désinstaller USAGES MOBILES disponible dans le menu de l'application ou en le faisant dans les réglages de l'appareil (voir annexe C).

5.3 Rapport du comité éthique

Le groupe de recherche a reçu l'approbation au plan de l'éthique pour le projet de recherche USAGES MOBILES le 13 janvier 2017, car le Comité institutionnel a jugé que le rapport était conforme aux normes établies par la Politique no 54 sur l'éthique de la recherche avec des êtres humains (2015) valide jusqu'au 31 août 2019. L'objectif de ce comité est d'analyser les enjeux éthiques et de respect de la vie privée du projet. Dans cette demande, le groupe de recherche devait justifier les critères de sélection de participants (âge, groupe d'appartenance, affiliation à un organisme, sexe, etc.), mais aussi présenter les critères d'exclusion des participants (origine ethnique, culture, sexe, âge, langue, etc.) ainsi les motifs qui justifient ces exclusions.

5.4 La collecte des données

Pour pouvoir offrir une protection maximale, il faudrait effectuer une anonymisation a priori, car comme mentionné dans le chapitre 4, l'anonymisation dès la conception est l'approche la plus protectrice des données. Cependant, dans le cadre de ce projet vu que la collecte des données a déjà été effectuée nous sommes donc dans l'obligation d'effectuer une anonymisation a posteriori des données. Pour pouvoir récupérer les données déjà collectées, nous avons dû nous connecter au réseau privé de l'UQAM en utilisant un code d'accès reçu du Grisq. Nous avons

obtenu quatre (4) fichiers au format SQL [1] : participant.sql, usage.sql, localisation.sql et questionnaire.sql.

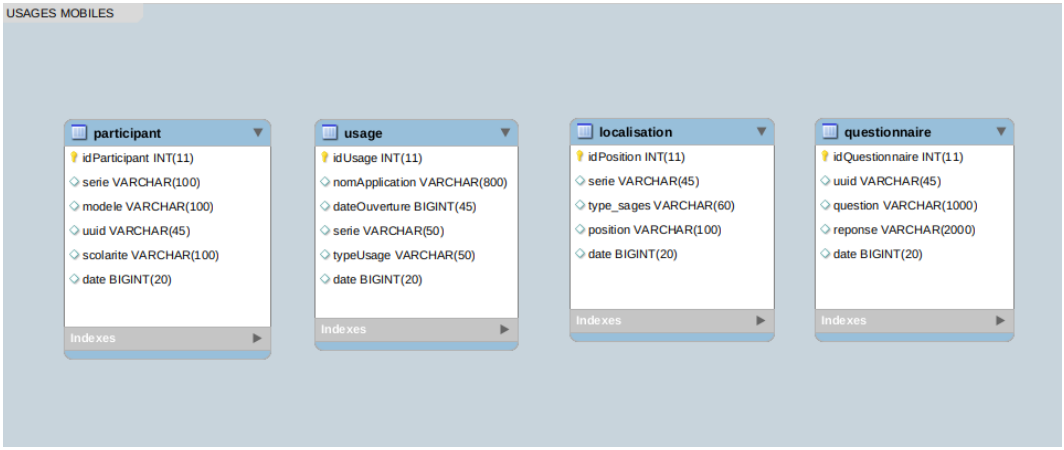


Figure 5.1 Structure de la base de données du Grisq

— Le fichier participant.sql contient des informations sur tous les participants ayant participé au projet de recherche. Par exemple, le numéro de série et le modèle de leur appareil Android.

— Le fichier usage.sql contient des informations sur les données d'usage des participants, par exemple, appel envoyé et reçu, fermeture et ouverture d'écran.

— Le fichier localisation.sql contient les différentes positions des participants représentées sous forme de code postal.

— Le fichier questionnaire.sql contient les réponses aux questions du formulaire de participation.

---

1. SQL est un langage informatique normalisé servant à exploiter des bases de données relationnelles.

Nous avons utilisé le logiciel MYSQL WORKBENCH [2] pour pouvoir visualiser les tables et manipuler les données reçues (voir figure 5.1). Dans le tableau 5.1, nous avons présenté la description détaillée de toutes les colonnes de chacune des tables reçues. Afin de mieux structurer les données brutes de départ, nous avons utilisé plusieurs requêtes SQL pour pouvoir enlever les données des participants qui ont rempli le formulaire, mais qui n'ont pas installé l'application USAGES MOBILES. Et nous avons remarqué que 10 des 20 participants n'ont pas installé l'application. Pour des raisons de confidentialité les membres du Grisq n'ont pas précisé la raison pour laquelle ces participants ont refusé de continuer à participer au projet, nous sommes donc obligés d'utiliser les données générées par les 10 participants actifs.

Nom d'attribut	Type	Description
série	Chaîne de caractères	Numéro série des appareils android des participants.
modèle	Chaîne de caractères	Le modèle des appareils mobiles androide des participants.
uuid	Chaîne de caractères	Un identifiant unique est attribué à chaque participant.
scolarité	Chaîne de caractères	Le diplôme académique le plus élevé des participants.
occupation	Chaîne de caractères	Information permettant de savoir si le participant a un emploi ou est aux études.
emploi	Chaîne de caractères	Le poste occupé par le participant s'il a un emploi.
nomApplication	Chaîne de caractères	Le nom des applications que les participants ouvrent.
dateOuverture	Timestamp	Date d'ouverture de chaque application..
typeUsage	Chaîne de caractères	Le type d'usage correspond à l'action que les participants font sur les applications qu'ils ouvrent. Par exemple s'ils cliquent, défilent ou écrivent du texte.
position	Chaîne de caractères	La position géographique de chaque participant lors de chaque usage.
question	Chaîne de caractères	Le libellé de chaque question du formulaire de consentement.
réponse	Chaîne de caractères	La réponse des participants pour chaque question du formulaire de consentement.
date	Timestamp	La date d'enregistrement de chaque action.

Tableau 5.1 Description des attributs des données du questionnaire

2. MySQL Workbench est un logiciel de gestion et d'administration de bases de données MySQL

Nous avons par la suite modifié la table participant, en ajoutant les colonnes âge, sexe, occupation et statut à partir des réponses trouvées dans la table questionnaire dans le but d'avoir uniquement la table participant (voir tableau 5.2).

L'objectif du groupe de recherche est d'obtenir le plus d'informations possibles sur les pratiques et les perceptions des utilisateurs de média socionumériques dans un contexte de mobilité. Ces informations sont stockées dans les tables application et localisation qui contiennent beaucoup d'enregistrements, alors pour faire nos tests, nous avons du fusionner ces deux tables. À titre d'exemple, nous avons sélectionné un enregistrement pour chaque participant (voir tableau 5.3).

numéro série	sexe	age	uuid	modèle	scolarité	occupation	statut
HT5B4BE10292	Homme	29	4fcdc8	HTC One A9	Université-baccalauréat	Étudiant(e)	Séparé(e)/Divorcé(e)
ZY2248JPLM	Homme	31	621fb8	Moto G Play	Université-certificat 1er cycle	En emploi	Célibataire
ZY223NMW4F	Homme	45	bec81e	XT1650	Collégial-technique	En emploi	Marié(e)/Union libre/Conjoint(e) de fait
LGM320d94ce438	Femme	67	3d49e2	LG-M320G	Université-baccalauréat	En emploi	Célibataire
ce12160c3112d31504	Homme	23	8137cb	SM-G935 W8	Secondaire général	Étudiant(e)	Célibataire
ZY2238P4VR	Homme	23	b7d441	Moto G (4)	Collégial-général	En emploi	Marié(e)/Union libre/Conjoint(e) de fait
LGM320f0f7d09a	Homme	28	296d0f	LG-M320G	Université-Maîtrise	En emploi	Célibataire
TA3970A8MJ	Femme	30	2883bd	MotoE2(4G-LTE)	Université-baccalauréat	En emploi	Célibataire
ZY222ZC9WV	Homme	23	19f922	MotoG3	Université-baccalauréat	Étudiant(e)	Célibataire
ce12160c658d6c2d01	Femme	65	e652fe	SM-G935 W8	Secondaire général	En emploi	Marié(e)/Union libre/Conjoint(e) de fait

Tableau 5.2 Table originale des participants

uuid	usage	position	date
4fcdc8	Appel envoyé	H4R 1L1	2018-02-01 20:43
621fb8	Ouverture WhatsApp	HT6116	2018-02-01 21:32
bec81e	Clique Google	H2X 3Y7	2018-02-19 20:51
3d49e2	Ouverture WhatsApp	J5L 2P7	2018-02-23 12:15
8137cb	Défile Facebook	J4P 2L5	2018-03-03 20:26
b7d441	Clique Google	H2X 1L2	2018-02-12 14:26
296d0f	Ouverture Gmail	H1W 2B3	2018-02-14 17:51
2883bd	Appel envoyé	H4G 1X5	2018-02-09 21:09
19f922	Ouverture Gmail	H2P 1T5	2018-02-06 16:03
e652fe	Ouverture Samsung Internet	H2L 2C4	2018-02-07 11:34

Tableau 5.3 Table d'usage des participants

5.4.1 Étape 1- évaluation de la préservation des besoins de l'application

Dans le but d'établir un compromis entre la qualité des données et leur anonymisation nous nous sommes entretenus avec les membres du groupe de recherche afin de bien comprendre les besoins de l'application. Nous avons analysé le questionnaire de participation et voici les points qui sont essentiels pour le groupe avant d'entamer notre processus d'anonymisation.

— Toutes les données de la table usage sont essentielles.

— La position géographique des participants peut être réduite à la ville.

— Le sexe des participants n'est pas important.

— L'âge peut être généralisé en sept ou dix ans d'intervalles, mais pas davantage.

— Le niveau de scolarité ainsi que le modèle de téléphone des participants ne sont pas essentiels.

Au niveau de la confidentialité des données, il est intéressant de noter que sur la première page du formulaire de consentement (voir annexe A), les chercheurs précisent que les données pourrait être protégées dès la cueillette (avant même leur transmission à la base de données de la recherche) par l'utilisation d'un pseudonyme qui sera associé au profil de chaque participant. Ainsi, toutes les manipulations qui seront effectuées sur les données le seront à partir des données pseudonymisées. Le seul document qui comportera un lien entre le pseudonyme et le nom des participants sera déposé dans un fichier unique verrouillé par mot de passe et conservé dans un local fermé à clé. De plus, les renseignements personnels seront détruits à la fin du projet de recherche (qui dure 5 ans). Et seules les données qui ne permettent pas d'identifier les participants seront conservées après cette date. Après avoir identifié les besoins essentiels du projet USAGE MOBILES nous pouvons maintenant passer à l'étape 2 de PROCOM afin de vérifier la protection accordée aux données collectées par l'application.

5.4.2 Étape 2- vérification de la protection accordée aux données reçues

Nous pouvons constater que l'équipe de recherche a effectué une anonymisation a priori afin de protéger l'anonymat des participants, car dans le formulaire on ne demande pas aux participants leurs noms, leurs adresses et leurs numéros de téléphone. Cependant, selon (Samarati, 2001) la désidentification des données ne fournit aucune garantie d'anonymat. Les informations publiées contiennent souvent d'autres données, telles que l'origine ethnique, la date de naissance, le sexe et le code postal, qui peuvent être liées à des informations accessibles au public pour ré-identifier les répondants et déduire des informations qui n'étaient pas destinées à être divulguées.

Nous devons donc appliquer l'étape 2 de PROCOM pour pouvoir mettre à l'épreuve la protection accordée par la technique de protection de données utilisée par le Grisq.

Nous avons choisi au hasard dans le tableau 5.2, le participant ayant le UUID «621fb8» et grâce aux données des quatre tables reçues, nous savons qu'il est un homme, célibataire, analyste programmeur qui détient un diplôme de premier cycle, né en 1989, qu'il prend surtout le transport en commun pour se rendre au travail qui est à 45 minutes de chez lui (voir annexe D). Vu que les données de localisation sont représentées sous forme de code postal, nous avons développé un outil en java pour afficher les coordonnées sur une carte Google. Et nous avons vu que ce participant se trouvait en Haïti plus précisément à Port-au-Prince (voir image 5.2).

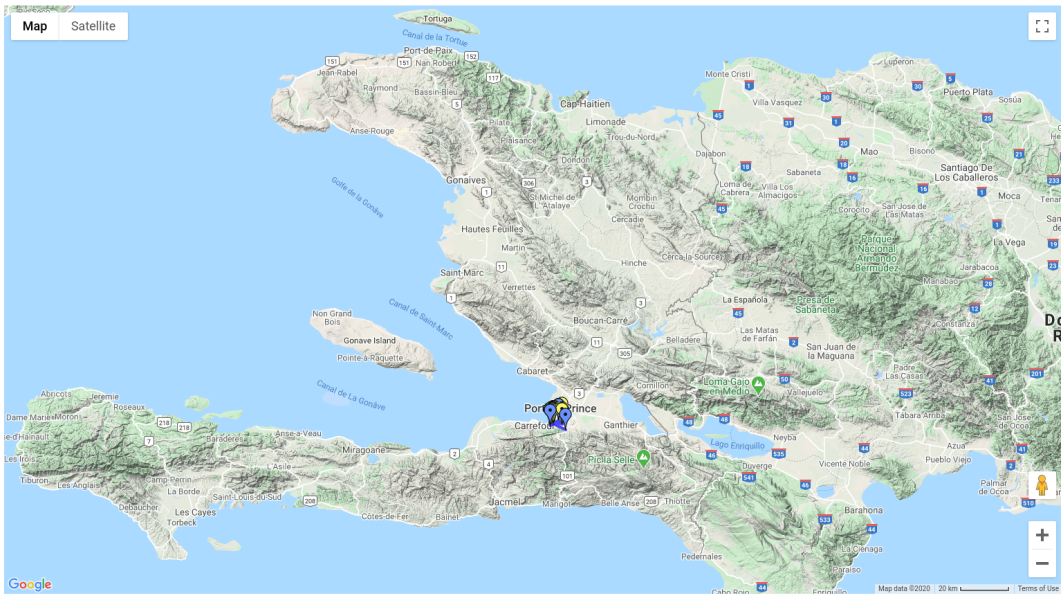


Figure 5.2 Carte d'Haïti avec les coordonnées géographiques du participant 621fb8

En analysant les points sur la carte nous avons vu que pendant toute la durée du pré-test, chaque matin vers 8h, ce participant se rendait toujours au 6 Delmas 48, Port-au-Prince, Haïti. En faisant une recherche sur Google nous avons vu que l'adresse correspondait à une entreprise appelée ATALOU MICROSYSTEM [3], et en allant sur le profil LinkedIn [4] de l'entreprise nous avons vu la liste des employés qui disent travailler chez ATALOU, parmi ces employés se trouve un analyste programmeur qui correspond au profil recherché. Nous pouvons donc constater qu'avec les réponses du questionnaire ainsi que les positions géographiques du participant nous avons pu déduire son identité en utilisant le moteur de recherche Google et le réseau social professionnel LinkedIn.

Nous avons aussi importé la table «participant» dans ARX et comme le montre la figure 5.3, des combinaisons d'attributs peuvent être analysées en ce qui concerne les risques associés à la ré-identification. La vue fournit des informations à quel point les combinaisons de variables séparent les enregistrements les uns des autres et à quel point les variables rendent les enregistrements distincts, et nous pouvons constater par exemple que la combinaison des attributs «sexe» et «age», pourraient permettre d'identifier les participants à 80%. Nous devons donc passer à la prochaine étape de PROCOM car la base de données du Grisq n'est pas anonymisée dans sa version brute.

---

3. ATALOU MICROSYSTEM est une entreprise qui offre ses services dans le domaine de l'informatique (http ://atalou.com/)

4. LinkedIn est un réseau social professionnel en ligne créé en 2002 à Mountain View.

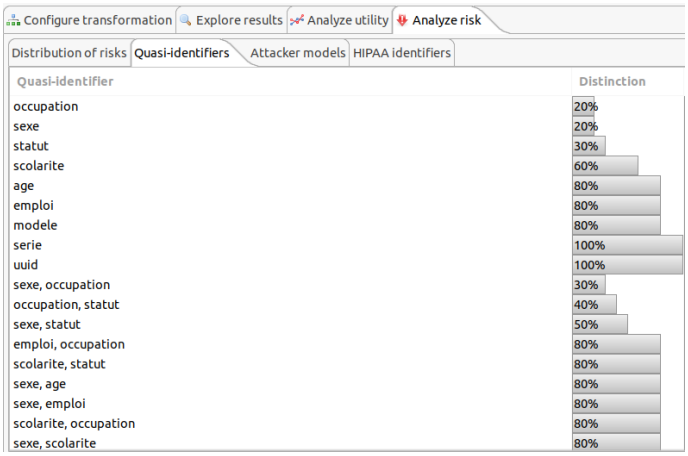


Figure 5.3 Analyse des risques de ré-identification de la table «participant» dans ARX

5.5 Étape 3 - identification des données à anonymiser

Nous avons vu dans la section précédente que les données reçues n'étaient pas tout à fait anonymisées, car nous avons pu en quelques heures identifier une personne avec une quasi-certitude, d'autant plus que nous avons utilisé pour le faire deux sources gratuites Google et LinkedIn, il faut savoir qu'un attaquant pourrait utiliser tous les moyens nécessaires pour parvenir à identifier l'identité d'une victime.

Pour appliquer l'étape 3 de PROCOM nous devons être en mesure d'identifier les QI, AS, ANS et supprimer les IE. Dans le tableau 5.2, les attributs «numéro série» et «uuid» sont considérés comme des IE car ces deux données peuvent explicitement identifier les participants du projet de recherche. Et grâce à la détection automatique des QI dans ARX, comme nous pouvons le constater dans la figure 5.4, tous les autres attributs sont considérés comme étant des QI. Effectivement, la combinaison des attributs «modèle», «âge», «occupation» et «statut» nous a permis dans la section 5.4.2 d'identifier au moins un participant du projet de

recherche.

Figure 5.4 Détection automatique des QI dans ARX

Dans le tableau 5.3, le «uuid» est considéré comme un IE, les attributs «usage», «position» et «date» sont considérés comme des QI (voir tableau 5.4 et 5.5).

QI					
sexe	**age**	**modèle**	**scolarité**	**occupation**	**statut**
Homme	29	HTC One A9	Université-baccalauréat	Étudiant(e)	Séparé(e)/Divorcé(e)
Homme	31	Moto G Play	Université-certificat 1er cycle	En emploi	Célibataire
Homme	45	XT1650	Collégial-technique	En emploi	Marié(e)/Union libre/Conjoint(e) de fait
Femme	67	LG-M320G	Université-baccalauréat	En emploi	Célibataire
Homme	23	SM-G935 W8	Secondaire général	Étudiant(e)	Célibataire
Homme	23	Moto G (4)	Collégial-général	En emploi	Marié(e)/Union libre/Conjoint(e) de fait
Homme	28	LG-M320G	Université-Maîtrise	En emploi	Célibataire
Femme	30	MotoE2(4G-LTE)	Université-baccalauréat	En emploi	Célibataire
Homme	23	MotoG3	Université-baccalauréat	Étudiant(e)	Célibataire
Femme	65	SM-G935 W8	Secondaire général	En emploi	Marié(e)/Union libre/Conjoint(e) de fait

Tableau 5.4 Résultat de la troisième étape de PROCOM appliquée aux données des participants

QI		
usage	**position**	**date**
Appel envoyé	H4R 1L1	2018-02-01 20:43
Ouverture WhatsApp	HT6116	2018-02-01 21:32
Clique Google	H2X 3Y7	2018-02-19 20:51
Ouverture WhatsApp	J5L 2P7	2018-02-23 12:15
Défile Facebook	J4P 2L5	2018-03-03 20:26
Clique Google	H2X 1L2	2018-02-12 14:26
Ouverture Gmail	H1W 2B3	2018-02-14 17:51
Appel envoyé	H4G 1X5	2018-02-09 21:09
Ouverture Gmail	H2P 1T5	2018-02-06 16:03
Ouverture Samsung Internet	H2L 2C4	2018-02-07 11:34

Tableau 5.5 Résultat de la troisième étape de PROCOM appliquée aux données d'usage

5.6 Étape 4 - identification du niveau d'anonymisation

Identifier le niveau d'anonymisation signifie de choisir une forme d'anonymisation c'est-à-dire de savoir si nous allons utiliser une anonymisation réversible ou irréversible. Vu que les chercheurs veulent publier les données à des fins de recherches et qu'ils n'ont pas besoin d'identifier individuellement chaque participant, nous avons donc répondu «non» à la question posée à l'étape 4 de PROCOM, à savoir si nous voulons que les personnes concernées par les données puissent bénéficier des résultats de la recherche (voir figure 4.3). De ce fait, nous leur proposons d'opter pour une anonymisation irréversible. Par contre les chercheurs doivent expliquer aux participants qu'ils qu'ils ne pourront pas bénéficier des résultats de la recherche parce que leurs données seront anonymisées de façon irréversible, ce qui ne permet pas de retourner aux données originelles.

Dans la prochaine section, nous allons montrer comment combiner les différentes techniques et méthodes de protection dites irréversibles.

5.7 Étape 5 - le choix des méthodes ou techniques de protection de données à combiner

Dans cette section nous allons appliquer l'étape 5 de PROCOM pour pouvoir rendre anonymes les tableaux 5.4 et 5.5. Pour cela, nous devrons répondre à un ensemble de questions fourni à l'étape 4 de PROCOM en fonction de nos besoins (voir figure 4.4).

Nous savons que l'équipe de recherche veut publier les données collectées à des tiers à des fins de recherche ou d'étude. De ce fait nous leur proposons de combiner les techniques de protection de données qui sont dites irréversibles. Parce que nous ne voulons pas que des attaquants puissent déduire des informations qui n'étaient pas destinées à être divulguées. Nous répondons «irréversible» à la première question. Nous savons aussi que l'objectif du groupe de recherche est de collecter le plus d'informations d'usage possible des participants, de ce fait nous leur suggérons de choisir les méthodes d'anonymisation qui ne perturbent pas les données c'est-à-dire de combiner les méthodes de protection qui sont dites non perturbatrices. À savoir, la technique de généralisation et la technique de suppression. Vu que nous n'avons pas identifié des AS à l'étape 3 il n'est pas nécessaire d'utiliser la technique de l-diversité. En fonction des réponses données, la «combinaison 2» de PROCOM parait la mieux adaptée au projet recherche USAGE MOBILES c'est-à-dire de combiner les méthodes de généralisation et de suppression avec la technique de k-anonymat pour pouvoir rendre anonymes les données des participants.

Nous allons dans un premier temps créer une hiérarchie de généralisation pour la plupart des attributs QI des deux tables en plusieurs niveaux. Plus on monte dans la hiérarchie moins les données seront précises. Puis dans un deuxième temps nous allons appliquer la technique de $k-$anonymat aux résultats obtenus par la généralisation et la suppression. ARX propose aussi différentes méthodes pour

créer des hiérarchies de généralisation pour différents types d'attribut. De plus, les spécifications de hiérarchie peuvent être importées et exportées pouvant ainsi être réutilisées pour anonymiser différents ensemble de données avec des attributs similaires.

5.7.1 Application de la généralisation et de la suppression à la table «participant»

Nous allons maintenant généraliser le QI «sexe» en un niveau S_1 qui peut être aussi considéré comme une suppression de l'attribut «sexe» :

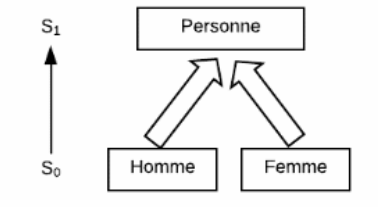


Figure 5.5 Généralisation du QI «sexe»

Pour le QI «âge», nous proposons plusieurs intervalles différents A_1 et A_2 :

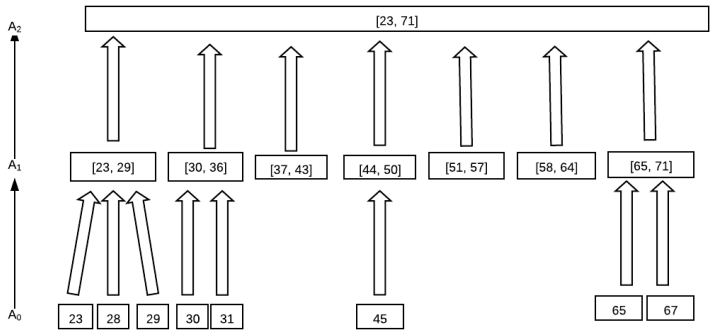


Figure 5.6 Généralisation du QI «âge»

Le QI «modèle» a été généralisé en deux niveaux M_1 et M_2 :

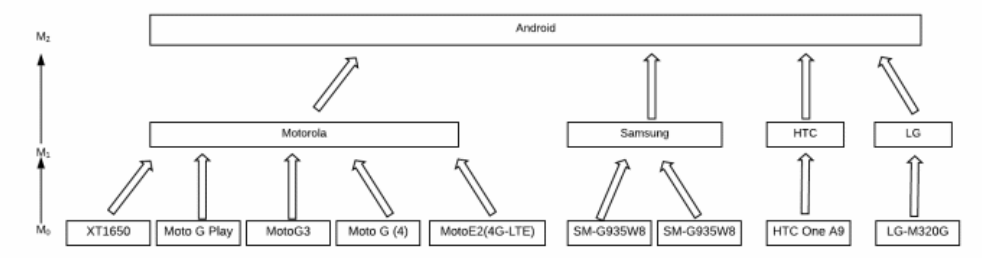


Figure 5.7 Généralisation du QI «modèle»

Et le QI «scolarité» a été généralisé en deux niveaux ; E_1 qui correspond au niveau secondaire et université, et E_2 qui correspond à tous les niveaux d'études :

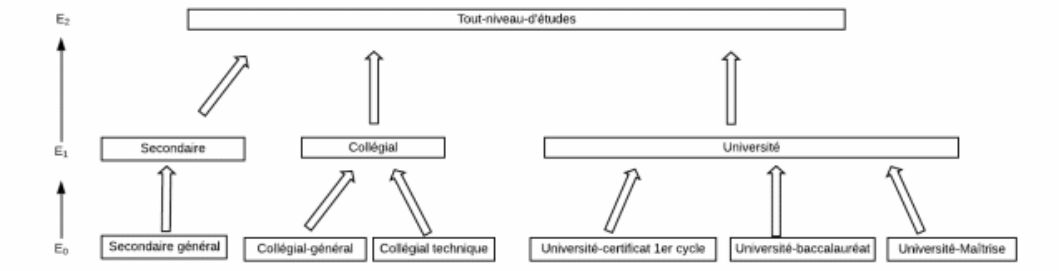


Figure 5.8 Généralisation du QI «scolarité»

La finalité du projet USAGE MOBILES est de collecter les informations d'usage de chaque participant dans un contexte de mobilité alors nous avons jugé non pertinente la colonne «occupation». De ce fait nous avons fait une suppression globale du QI «occupation». Le tableau 5.6, présente le résultat final de l'application de la combinaison 1 de PROCOM à la table «participant».

sexe	age	modèle	scolarité	statut
Personne	[23,29]	Android	Université	Séparé(e)/Divorcé(e)
Personne	[30,36]	Android	Université	Célibataire
Personne	[44,50]	Android	Collégial	Marié(e)/Union libre/Conjoint(e) de fait
Personne	[65,71]	Android	Université	Célibataire
Personne	[23,29]	Android	Secondaire	Célibataire
Personne	[23,29]	Android	Collégial	Marié(e)/Union libre/Conjoint(e) de fait
Personne	[23,29]	Android	Université	Célibataire
Personne	[30,36]	Android	Université	Célibataire
Personne	[23,29]	Android	Université	Célibataire
Personne	[65,71]	Android	Secondaire	Marié(e)/Union libre/Conjoint(e) de fait

Tableau 5.6 Table généralisée des participants

5.7.2 Application de la généralisation et de la suppression à la table «usage»

En généralisant le QI «position» nous avons obtenu trois niveaux ; P_1 qui correspond à la ville, P_2 à la province et P_3 au pays :

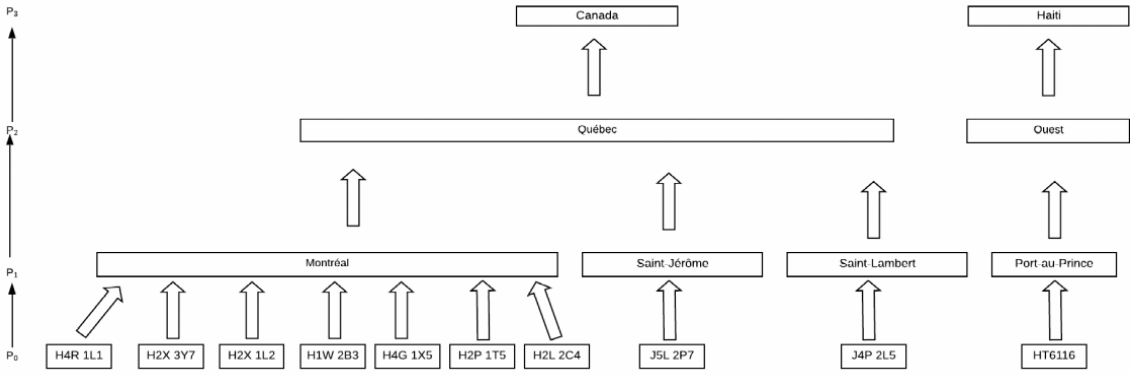


Figure 5.9 Généralisation du QI «position»

Pour le QI «date» nous avons obtenu deux niveaux ; D_1 qui correspond au couple mois-année et D_2 à l'année :

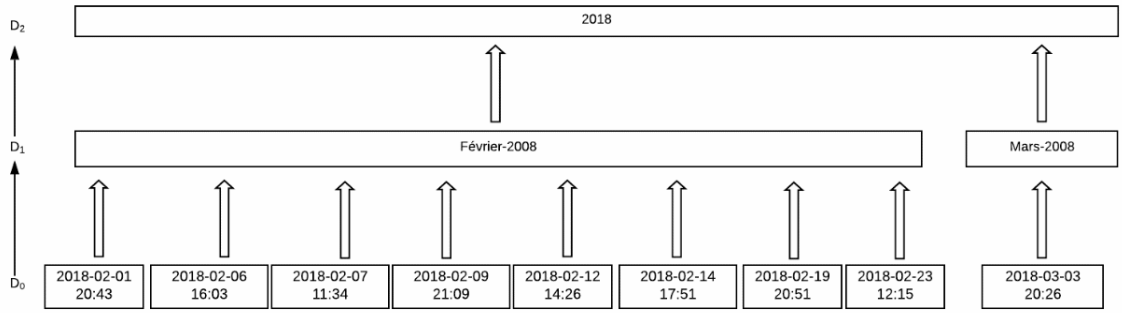


Figure 5.10 Généralisation du QI «date»

L'équipe de recherche était stricte concernant la pertinence des données d'usage, pour cela nous n'avons pas généralisé le QI «usage». Cependant grâce à la généralisation des QI «position» et «date», le QI «usage» à lui seul présente peu de danger à la vie privée des participants. Le tableau 5.7 présente le résultat final de la généralisation des données de la table usage.

usage	**position**	**date**
Appel envoyé	Montréal	Février-2008
Ouverture WhatsApp	Ouest	Février-2008
Clique Google	Montréal	Février-2008
Ouverture WhatsApp	Saint-Jérôme	Février-2008
Défile Facebook	Saint-Lambert	Mars-2008
Clique Google	Montréal	Février-2008
Ouverture Gmail	Montréal	Février-2008
Appel envoyé	Montréal	Février-2008
Ouverture Gmail	Montréal	Février-2008
Ouverture Samsung Internet	Montréal	Février-2008

Tableau 5.7 Table généralisée de la table «usage»

5.7.3 Application de la technique $k-$anonymat aux données des participants

La principale exigence du $k-$anonymat est que chaque sortie des données doit être telle que chaque combinaison de valeurs de QI puisse être indistinctement associée à au moins k individus. Nous pouvons donc voir que le tableau 5.6 ne satisfait pas le $k-$anonymat parce que la combinaison des QI [sexe, âge, modèle, scolarité, statut] nous permet d'identifier au moins un participant de manière unique. Par exemple, si on sait que le participant a un appareil Android, il a 60 ans et qu'il est marié on peut l'identifier facilement, car il est unique dans le tableau. Pour cela dans le tableau 5.8, nous avons mis en rouge les valeurs à supprimer et en bleu les valeurs à masquer pour pouvoir satisfaire le 2-anonymat.

sexe	age	modèle	scolarité	statut
Personne	[23,29]	Android	Université	Séparé(e)/Divorcé(e)
Personne	[23,29]	Android	Université	Célibataire
Personne	[23,29]	Android	Collégial	Célibataire
Personne	[23,29]	Android	Université	Marié(e)/Union libre/Conjoint(e) de fait
Personne	[23,29]	Android	Secondaire	Célibataire
Personne	[30,36]	Android	Université	Célibataire
Personne	[30,36]	Android	Université	Célibataire
Personne	[44,50]	Android	Secondaire	Marié(e)/Union libre/Conjoint(e) de fait
Personne	[67,71]	Android	Université	Célibataire
Personne	[67,71]	Android	Collégial	Marié(e)/Union libre/Conjoint(e) de fait

Tableau 5.8 Tableau des participants qui ne satisfait pas le 2-anonymat

Le tableau 5.9 présente le résultat des données de participants qui satisfait le 2-anonymat.

sexe	age	modèle	scolarité	statut
Personne	[23,29]	Android	Université	***
Personne	[23,29]	Android	Université	***
Personne	[23,29]	Android	Tout-niveau-d'études	Célibataire
Personne	[23,29]	Android	Tout-niveau-d'études	Célibataire
Personne	[23,29]	Android	Tout-niveau-d'études	Célibataire
Personne	[30,36]	Android	Université	Célibataire
Personne	[30,36]	Android	Université	Célibataire
Personne	[67,71]	Android	Tout-niveau-d'études	***
Personne	[67,71]	Android	Tout-niveau-d'études	***

Tableau 5.9 Table participant qui satisfait le 2−anonymat

Le tableau 5.10 représente le résultat de la deuxième itération aux données usages des participants, en bleu ce que nous allons masquer et en rouge ce que nous allons supprimer. Et le tableau 5.11 affiche le résultat final de la table d'usage qui respecte le 2-anonymat.

usage	position	date
Appel envoyé	Montréal	Février-2008
Appel envoyé	Montréal	Février-2008
Ouverture WhatsApp	Ouest	Février-2008
Ouverture WhatsApp	Saint-Jérôme	Février-2008
Clique Google	Montréal	Février-2008
Clique Google	Montréal	Février-2008
Défile Facebook	Saint-Lambert	Mars-2008
Ouverture Gmail	Montréal	Février-2008
Ouverture Gmail	Montréal	Février-2008
Ouverture Samsung Internet	Montréal	Février-2008

Tableau 5.10 Deuxième itération aux données d'usage

usage	position	date
Appel envoyé	Montréal	Février-2008
Appel envoyé	Montréal	Février-2008
Ouverture WhatsApp	***	Février-2008
Ouverture WhatsApp	***	Février-2008
Clique Google	Montréal	Février-2008
Clique Google	Montréal	Février-2008
Ouverture Gmail	Montréal	Février-2008
Ouverture Gmail	Montréal	Février-2008

Tableau 5.11 Table usage qui satisfait le 2−anonymat

5.8 L'anonymisation est-elle satisfaisante ?

Pour pouvoir répondre à cette question, nous avons présenté les résultats des deux tables anonymisées «participant» et «usage» au Grisq. Ils nous ont demandé d'ajuster la table «participant». À des fins de statistiques, ils voudraient si possible avoir le sexe des participants ainsi que le modèle de leurs téléphones. Selon le Grisq, les usages réels des appareils mobiles ont pour objectif d'analyser plusieurs facteurs, dont le stress professionnel causé par l'accessibilité et la présence constante des appareils mobiles, d'observer l'accélération sociale technologique, de connaître les habitudes et les effets des utilisations d'appareils mobiles au quotidien.

Les données : «sexe», «âge» et «position» permettent de faire le regroupement démographique des usages et problématiques qui émergent des données. La position géographique sert aussi à faire des regroupements par territoire (par exemple ville vs campagne), mais aussi à calculer la vitesse de déplacement pour connaître les usages durant les déplacements.

Nous avions envisagé dans notre modèle de conception une phase de retour en arrière pour chaque étape au cas où on aurait trop enlevé d'éléments. Nous sommes donc retournés à l'étape 5 et nous avons choisi le niveau S_0 pour l'attribut «sexe» (voir figure 5.5) et le niveau M_1 pour l'attribut «téléphone» (voir figure 5.7). Le tableau 5.12 montre le résultat final de la table «participant» qui a été accepté par le Grisq.

sexe	age	modèle	statut
Homme	[23,29]	HTC One A9	***
Homme	[23,29]	Moto G (4)	***
Homme	[23,29]	SM-G935W8	Célibataire
Homme	[23,29]	MotoG3	Célibataire
Homme	[23,29]	SM-G935W8	Célibataire
Personne	[30,36]	Android	Célibataire
Personne	[30,36]	Android	Célibataire
Femme	[67,71]	LG-M320G	***
Femme	[67,71]	LG-M320G	***

Tableau 5.12 Résultat final de la table participant qui satisfait le 2−anonymat et accepté par le Grisq

Le modèle k-anonymat permet la publication de données à usage général avec une utilité raisonnable quelles que soient les utilisations des données, au prix de certaines faiblesses en matière de confidentialité. Par contraste, a confidentialité différentielle offre une garantie de confidentialité très robuste au prix d'une limitation substantielle de l'utilité des sorties anonymisées (Soria-Comas *et al.*, 2014). Nous avons dû faire un compromis entre la sécurisation des données et le maintien de la précision en fonction de l'usage des données fait par le Grisq. Et à la demande du Grisq l'application de notre démarche PROCOM à USAGES MOBILES

est, pour le moment limitée à la combinaison 2 (Généralisation + suppression + k-anonymat).

5.9 Autre aspect de PROCOM

Comme nous pouvons le constater, les données de l'application USAGES MOBILES ne contiennent pas d'attributs sensibles, dans ce cas afin d'exploiter un autre aspect de PROCOM nous avons monté un scénario secondaire avec des données fictives. Prenons l'exemple d'une table dans une base de données médicales sur des étudiants et des enseignants d'une université (voir tableau 5.13).

L'attribut «nom» est considéré comme un IE, les attributs [«activité», «age»] sont considérés comme des QI et l'attribut «maladie» est considéré comme un AS.

Uuid	**Activité**	**Age**	**Maladie**
Suze	Maîtrise	25	Cirrhose
Paul	Chargé de cours	30	Cancer
Dany	Professeur	32	Cancer
Boby	Baccalaureat	22	VIH
Billy	DESS	21	Cirrhose
Samy	Professeur	29	Cancer
John	Maîtrise	25	Covid 19
Jymmy	Maîtrise	26	Covid 19
Tomy	DESS	22	Insuffisance cardiaque

Tableau 5.13 Données médicales brutes

5.9.1 Anonymisation irréversible et réversible

Pour nos tests nous allons considérer deux scénarios. Dans le premier scénario nous allons opter pour une anonymisation réversible, et dans le deuxième scénario une anonymisation irréversible.

5.9.1.1 Scénario réversible

Selon l'étape 4 de PROCOM si nous voulions que la personne concernée par la donnée puisse bénéficier des résultats de la recherche nous devons appliquer la technique de pseudonymisation. Cette technique consiste à ajouter à chaque enregistrement un nouveau champ, appelé pseudonyme. Ce pseudonyme doit rendre impossible tout lien entre cette nouvelle valeur et la personne concernée par la donnée. Pour cela nous avons utilisé une fonction de génération aléatoire en langage java (voir figure 5.14), pour générer des identifiants uniques à six caractères.

Le tableau 5.15, montre le résultat de la pseudonymnisation appliquée à la table originale 5.13. Toutefois comme mentionné dans le chapitre 3, la pseudonymnisation ne donne pas un niveau de protection suffisamment élevé, car la combinaison d'autres champs peut permettre d'identifier les individus concernés.

```
public static String generateUUID() {
    String uuid = UUID.randomUUID().toString();
    return uuid.substring(0, 6);
}
```

Tableau 5.14 Fonction de généralisation aléatoire à six chiffres

5.9.1.2 Scénario irréversible

Pour illustrer le scénario irréversible nous avons choisi la combinaison 4 de PROCOM, car nous avons identifié un AS qui est l'attribut «maladie». Nous devons donc combiner k-anonymat et l-diversité. Dans la figure 5.11, nous avons généralisé l'attribut âge à quatre ans d'intervalle. Nous avons réduit le niveau de détail des données de telle sorte qu'il y a au moins k n-uplet différents qui ont la même valeur de QI. Nous avons aussi généralisé l'attribut «activité» des étudiants du

Uuid	Activité	Age	Maladie
8137cb	Maîtrise	25	Cirrhose
8862ac	Chargé de cours	30	Cancer
4fcdc8	Professeur	32	Cancer
2883bd	Baccalaureat	22	VIH
b7d441	DESS	21	Cirrhose
db90e4	Professeur	29	Cancer
1f4f15	Maîtrise	25	Covid 19
e652fe	Maîtrise	26	Covid 19
628c99	DESS	22	Insuffisance cardiaque

Tableau 5.15 Résultat de la pseudonymisation

(DESS et Baccalauréat) en «étudiant» et «professeur» et «chargé de cours» en «enseignant».

La technique de k-anonymat est sans effet sur les AS. À titre d'exemple, si on considère la figure 5.11, on peut facilement déduire qu'un enseignant ayant entre 29 et 32 ans a forcément le cancer. Si l'attaquant sait que Dany est un professeur de 32 ans, alors il peut déduire qu'il a le cancer. Pour cela nous devons utiliser la technique de l-diversité pour résoudre ce problème.

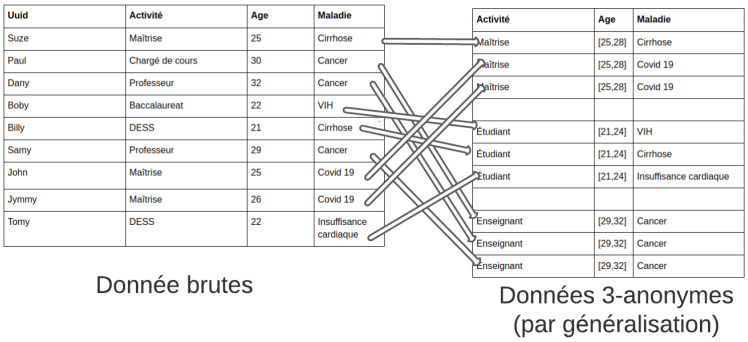


Figure 5.11 Données médicales brutes

Pour pouvoir satisfaire la confidentialité nous avons dans un premier temps utilisé la technique de micro-agrégation sur l'attribut «âge» pour pouvoir renforcer le k-anonymat. Cette technique va remplacer les valeurs originales par une moyenne. Et dans un deuxième temps nous avons rendu encore plus général l'attribut «activité». La figure 5.12 montre le résultat final de l'application de la combinaison 4 de PROCOM aux données médicales des étudiant et enseignants d'une université.

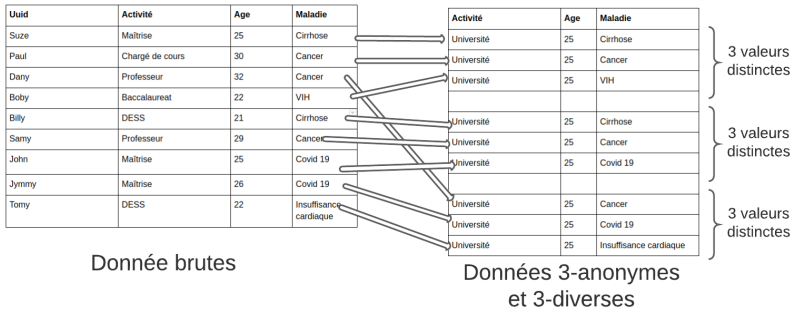


Figure 5.12 Résultat final de l'anonymisation des données médicales, contexte d'une université

5.9.1.3 Présentation des résultats

Après avoir appliqué notre démarche de conception PROCOM au projet USAGES MOBILES, nous devons maintenant avoir une mesure pour pouvoir évaluer la qualité des données produites par notre démarche en les comparant avec les données originales. Effectivement, d'après la littérature consultée, pour éviter que le processus d'anonymisation ne dégrade pas trop la précision des données et de leurs utilités, il existe des métriques de qualité ou d'évaluation appelée « data metrics » permettant de mesurer la qualité des données d'une table anonyme en la comparant à la qualité des données de la table originale. De ce fait, dans cette section nous allons appliquer aux données produites par notre démarche la métrique de complétudes des données et la métrique de perte d'information $ILoss$.

5.9.2 Métrique de complétude des données

Selon (Fung *et al.*, 2010a), la métrique de complétude des données évalue le degré de données manquantes dans une table anonymisée par rapport à la table d'origine. Elle est utile dans le cas où on utilise des techniques ou des méthodes de suppression globales. La complétude peut être mesurée comme le nombre d'enregistrements supprimés par rapport au nombre total d'enregistrements dans la table d'origine. En l'occurrence, le tableau 5.2 des participants contenait 10 enregistrements et après l'avoir anonymisé on a obtenu le tableau 5.9 avec 9 enregistrements. En appliquant la métrique de complétudes des données à ce tableau on peut dire que l'anonymisation nous a permis de conserver 90% des données originales. De même, l'anonymisation du tableau 5.3 en tableau 5.11 nous a permis de conserver 80% des données originales. Dans le cas notre scénario fictif comme le montre la figure 5.12, nous avons conservé 100% des données originales mais au détriment de la qualité des données car nous avons remplacé les valeurs des certains attributs par des valeurs moins spécifiques.

5.9.3 Métrique de perte d'information $ILoss$

$ILoss$ est proposé par (Xiao et Tao, 2006), pour capturer la perte d'information de la généralisation d'une valeur spécifique à une valeur générale vg ; la perte d'information pour un attribut spécifique est données par : $ILoss(v_g) = \frac{|v_g|-1}{|D_A|}$ où $|v_g|$ est le nombre de valeurs de domaine qui sont des descendants de v_g et $|D_A|$ est le nombre total de nœuds feuilles dans l'arbre de généralisation. $ILoss\ (v_g) = 0$ si v_g est une valeur de données d'origine dans le tableau. Et plus $ILoss\ (v_g)$ tend vers 1, plus la perte d'information est grande. En d'autres termes, $ILoss(v_g)$ mesure la fraction des valeurs de domaine généralisées par v_g. Par exemple, la généralisation d'une instance de Femme à Personne dans la figure 5.5 est : $ILoss(Personne) = \frac{2-1}{2} = 0,50$. La perte globale d'information d'une table anonyme T' de taille N

peut être calculée comme suit :

$$\frac{1}{N}\sum_{i=1}^{N} ILoss(v_{g_i})$$

Dans la figure 5.13, nous présentons la perte d'information découlant de la généralisation des attributs de la table 5.9, que nous avons proposée au Grisq. Par exemple, la perte d'information découlant de la généralisation de l'attribut «âge», pour l'intervalle [23-29] est $ILoss(Age[23-29]) = 0,25$. La figure 5.14, présente de son côté la perte d'information de la table modifiée 5.12, qui a été acceptée par le Grisq. Nous constatons que dans la figure 5.14, plusieurs attributs ont conservé leurs valeurs d'origine. Par exemple $ILoss(sexe) = 0$, $ILoss(HTC) = 0$ pour ne citer que ceux-là. En effet, dans le but de fournir un niveau satisfaisant de protection des données, à travers notre méthode PROCOM nous avons tenté d'altérer les données de telle sorte qu'aucun participant ne puisse être identifié de manière unique. Par exemple, dans la table proposée nous avons généralisé l'attribut «modèle» au niveau M_2. Cette généralisation à M_2 assure une meilleure protection plutôt que le niveau M_1, cependant les données qui en résultent sont inutiles par rapport aux objectifs fixés par le Grisq.

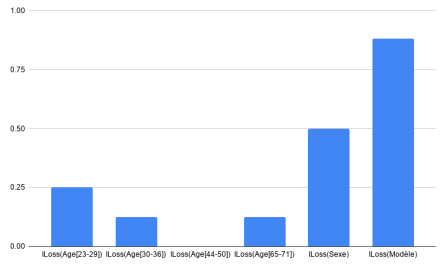


Figure 5.13 Perte d'information découlant de la généralisation des données proposé par le Grisq

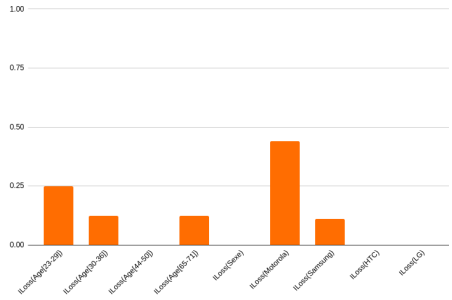


Figure 5.14 Perte d'information découlant de la généralisation des données accepté par le Grisq

En appliquant la formule de perte globale d'informations mentionnée dans la section 5.9.2, nous avons obtenu $ILoss(table5.9) = 0,31$ et $ILoss(table5.12) = 0,11$. La figure 5.15, montre la perte globale d'informations engendrée par la généralisation des tables 5.9 et 5.12. Nous pouvons constater que la perte globale d'information a diminuée de plus de moité à cause de l'assouplissement demandé par le Grisq. Cela entraîne automatiquement une perte de confidentialité, risque qui a été accepté par le Grisq. Vu que les données anonymisées doivent permettre d'effectuer des tâches de recherche et d'analyse, il est important d'assurer un bon compromis entre la protection de la vie privée et l'utilité des données.

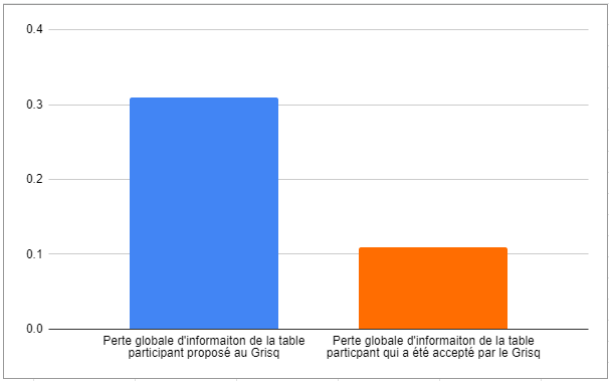


Figure 5.15 Perte totale.

5.10 Synthèse

Dans ce chapitre nous avons dans un premier temps présenté le projet USAGES MOBILES qui est une application mobile Android développée par le Grisq. Ce projet permet d'étudier les pratiques et les perceptions des utilisateurs de médias socionumériques dans un contexte de mobilité.

Dans un deuxième temps, grâce à la première étape de PROCOM nous avons eu des discussion avec le Grisq pour pouvoir préserver les besoins de l'application ensuite à l'étape 2 nous avons prouvé que les données collectées par USAGE

MOBILES n'étaient pas anonymisées correctement, pour cela nous avons sélectionné au hasard un participant de la recherche et en fonction des informations publiées par le groupe de recherche nous avons pu l'identifier. Par la suite nous avons appliqué les autres étapes de notre démarche de conception PROCOM aux données reçu du Grisq. Nous avons aussi monté un scénario secondaire fictif pour pouvoir exploiter d'autres cas de figures qu'USAGE MOBILES ne permettait pas de prendre en compte.

Dans le but de valider notre démarche de conception PROCOM, nous avons appliquer la métrique de complétude des données ainsi que la métrique $ILoss$ permettant de mesurer la qualité des données d'une table anonyme en la comparant à la qualité des données de la table originale. Dans le prochain chapitre, nous allons faire un rappel de la problématique de la recherche et ensuite répondre aux questions de recherche posées au chapitre 1.

CHAPITRE VI

RÉSULTATS, DISCUSSIONS ET CONCLUSION

6.1 Introduction

En regardant plus précisément les vols de données, nous voyons que, depuis ces dernières années, nous avons affaire à des chiffres astronomiques. Par exemple en 2015, il y avait à peu près 3 000 brèches de données c'est-à-dire des pertes de données ou des instruisions. Prenons comme exemple le scandale Strava qui est une société de «fitness» qui collecte la position géographique de ses utilisateurs. En novembre 2017, Strava a tracé des graphiques assez intéressants des parcours de tous ses utilisateurs. Il a été démontré que dans certains lieux assez reculés du globe, par exemple la province d'Helmand en Afghanistan, les seuls utilisateurs de cette application étaient des militaires américains se trouvant sur le sol afghan. Ainsi, en regardant les traces envoyées par les GPS de l'application «fitness» de ces personnes, il fut assez simple de retrouver très précisément la localisation de certaines bases militaires américaines ou d'autres pays (Véliz et Grunewald, 2018). Nous recommandons au Grisq de se prévenir des fuites données, de se protéger contre les attaques externes en contrôlant l'accès aux données, de se demander réellement quelles données ils veulent véritablement publier, et s'interroger sur la précision des données.

L'approche courante consistant à collecter autant de données que possible et à les

conserver le plus longtemps possible est le résultat de la perception que les données sont un bien inconditionnel - plus on en a, mieux c'est. Même si les utilisations des données ne sont pas encore claires, elles pourraient être utiles à l'avenir, selon ce point de vue. Cependant, conserver des données comporte des risques. Plus les données sont conservées longtemps, plus le risque d'utilisation abusive, accidentelle ou malveillante, est grand. Nous recommandons aussi au Grisq d'appliquer à l'avenir les principes de PbD. C'est-à-dire de prendre en compte la sécurité et la confidentialité au moment où ils conçoivent leurs systèmes. En effet, une bonne sécurité et une bonne confidentialité des données doivent se prévoir à l'avance. Il est en réalité beaucoup plus compliqué de prouver que la sécurité du système est atteinte lorsqu'on rajoute des couches de sécurité après, plutôt qu'au moment où on conçoit le système. C'est pourquoi, dès lors qu'on développe des applications autour d'objets connectés qui utilisent et exploitent des données personnelles, il est très important qu'au moment où l'on conçoit le système, on prenne en compte tous les éléments liés à la sécurité et à la confidentialité des informations.

6.2 Réponses aux questions de recherche

Dans cette section, nous allons présenter un bref rappel de problématique de la recherche. Par la suite, on présente la réponse à chacune des questions de recherche soulevées au chapitre 1 de cette étude.

La problématique soulevée au chapitre 1 était que les données collectées par les applications mobiles présentaient de graves problèmes de confidentialités, car les téléphones intelligents actuels ne sont pas compétents pour gérer les données sensibles des utilisateurs et font face à des fuites de données sensibles causées par une collecte excessive de données. Et d'après la littérature consultée, les techniques de protection de données à elles seules ne permettent pas de résoudre le problème de protection des données des utilisateurs.

6.2.1 Réponses aux questions de recherche secondaires

L'objectif principal de notre projet de recherche était de mettre en oeuvre un modèle de conception appelé PROCOM qui permet de combiner plusieurs techniques et méthodes de protection de données dans le but de protéger la confidentialité et protection des utilisateurs d'un service livré par une plateforme mobile.

De ce fait, pour atteindre cet objectif, on avait soulevé trois (3) questions de recherche secondaires.

La première question de recherche secondaire était la suivante :Comment peut-on préserver la confidentialité et l'anonymat des utilisateurs tout en maximisant l'utilité des données ?

Pour répondre à cette question, on s'est basé sur notre revue de littérature. Nous avons vu qu'il existe plusieurs types d'attaques à la vie privée à savoir les attaques par lien d'enregistrement, par lien de tables et par inférences probabilistes. Et qu'il existe aussi plusieurs modèles de protection des données qui permettent de contrer ces différents types d'attaques. Par exemple, le modèle k-anonymat permet de contrer les attaques par lien d'enregistrements et non les attaques par lien d'attributs tandis que le modèle $l-$diversité permet de protéger contre les attaques par lien d'attributs et non par les attaques par lien de tables.

En somme, pour pouvoir préserver la confidentialité des utilisateurs il faudrait avoir un guide d'anonymisation qui permettrait de combiner plusieurs méthodes ou techniques on fonction des besoins de chaque application tout en permettant d'offrir le maximum de confidentialité.

La deuxième question de recherche secondaire était la suivante : est-il possible d'éliminer ou d'assouplir certaines données que les applications collectent, et de conserver seulement le strict nécessaire, sans nuire aux besoins de l'application ?

La collecte des données est énorme, et démesurée par rapport à l'utilité qu'en fait le détenteur des données. En tant que concepteur d'applications il est possible d'appliquer le principe de collecte limitée des données, c'est-à-dire de ne collecter que les informations dont on a besoin pour remplir une tâche précise, et d'effacer ou de dégrader la précision de ces données une fois la tâche accomplie.

À travers la première étape de PROCOM, nous avons pu avoir des discussions avec le Grisq dans le but de faire l'évaluation de la préservation des besoins de l'application USAGES MOBILES. Nous avons remarqué que la qualité des données est généralement considérée comme un concept multidimensionnel qui implique dans certains contextes des paramètres objectifs et subjectifs. Parmi les différents paramètres possibles, la précision des données, la complétude des données et la cohérence des données sont considérées comme les paramètres les plus pertinents. Alors c'est au détenteur des données de faire un compromis entre la perte d'information et le gain en anonymisation.

6.2.2 Réponses à la question de recherche principale

La question de recherche principale a été formulée de la façon suivante : comment peut-on combiner plusieurs techniques de protection de données afin de concevoir un modèle de conception qui permettrait de protéger la confidentialité et l'anonymat des utilisateurs d'un service livré par une plateforme mobile ?

Pour répondre à cette question de recherche, on s'est basé sur les étapes de PROCOM. On va présenter sous quatre (4) volets la réponse à la question de recherche principale :

1. Il faut commencer à anonymiser les données dès la collecte des données, c'est-à-dire de collecter les données essentielles, de collecter le strict nécessaire en fonction des besoins de l'application.

2. Il faut identifier le niveau d'anonymisation c'est-à-dire de choisir une forme d'anonymisation réversible ou irréversible et évaluer les risques en matière d'anonymisation.
3. Par la suite, il faut combiner les différentes méthodes d'anonymisation et techniques présentés dans le chapitre 3 en fonction de chaque type d'attaques.
4. Enfin, il faut évaluer la qualité des données générées à partir des métriques de qualité de données.

6.3 Conclusion

Dans ce rapport, nous nous sommes intéressés au problème de la publication de données respectueuses de la vie privée et plus particulièrement à la conception d'un modèle de confidentialité et de l'anonymisation d'un service livré par une plateforme mobile. Comme nous pouvons le constater, les différentes méthodes d'anonymisation ont des limites importantes car la majorité des techniques et méthodes de protection de données ne pouvaient à elles seules protéger la confidentialité et l'anonymat des utilisateurs à 100%. En effet, il n'existe pas une technique meilleure ou moins bonne qu'une autre, mais qu'en fonction du type d'attaques, chaque technique pouvait trouver son utilité. Et aucun organisme ne peut prétendre avoir des données pleinement anonymes. Afin de renforcer l'anonymat nous avons mis en place un modèle de conception qui en cinq étapes (5) nous a permis de combiner les techniques et méthodes de protection de données les mieux considérées à ce jour. Nous avons pris acte que l'anonymisation n'est pas la seule voie à considérer, et qu'il faudrait peut-être mieux encadrer l'utilisation des données. Mais en même temps, il ne faut pas trop restreindre leur utilisation secondaire afin de pouvoir tirer des bénéfices des données pour l'intérêt collectif.

6.3.1 Perspectives de recherche

En termes de recherches future notre travail peut être poursuivi selon au moins quatre axes :

— Concevoir un algorithme qui pourrait combiner de façon optimale les différentes combinaisons de PROCOM en fonction des résultats des différentes métriques de qualité de données afin de générer automatiquement des tables anonymes.
— Étendre à d'autres techniques pour pouvoir mieux exploiter la confidentialité différentielle car au cours de la dernière décennie elle a pris une place prépondérante dans la protection des données sensibles.
— Concevoir un outil qui pourrait automatiser la plupart des étapes de PROCOM afin que notre démarche puisse être utilisée et déployée facilement par les chercheurs.
— Conduire une expérimentation à plus grande échelle incluant des utilisateurs pour mesurer l'utilité et l'utilisabilité de notre modèle PROCOM

Ceci complète la rédaction de ce mémoire. Il décrit une solution fonctionnelle aux problèmes liés à l'anonymisation des données personnelles dans le but d'en faire usage dans le domaine de la recherche. La section suivante fait état des recommandations pour la suite du projet.

6.3.2 Recommandations

L'application de notre modèle PROCOM aux données du Grisq nous a permis de faire les recommandations suivantes :

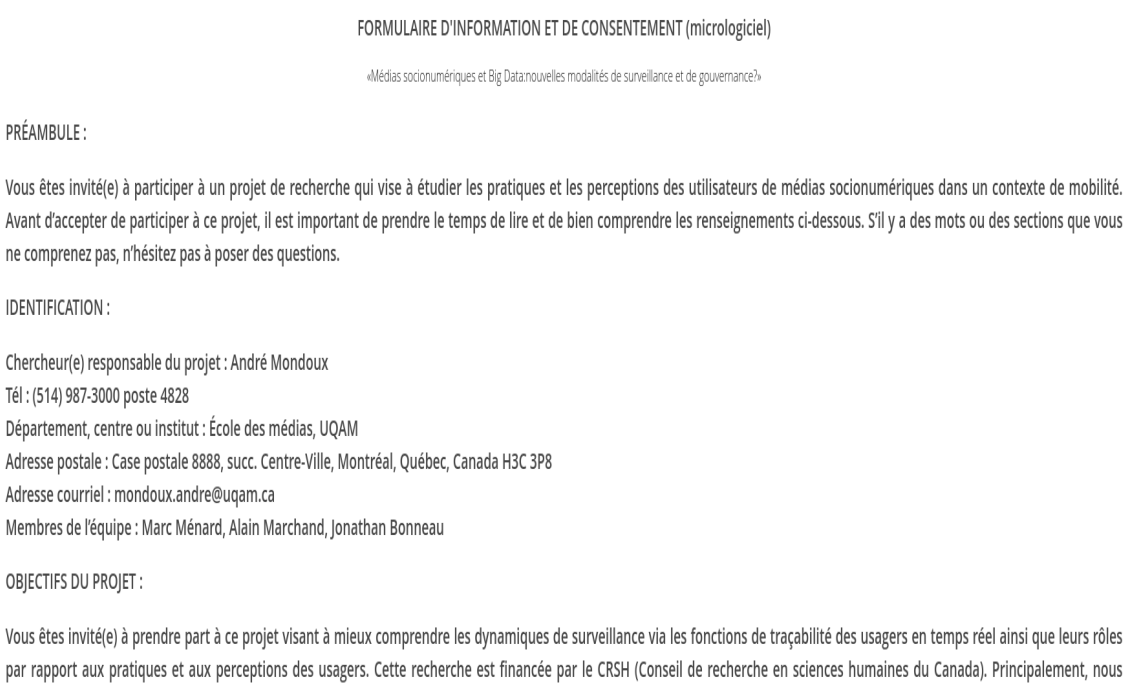
— Une seule solution assurée pour limiter au départ l'impact d'une fuite de données est de limiter la collecte des données, soit en collectant les données

dont on a vraiment besoin, soit en effaçant les données ou en dégradant la précision des données une fois la tâche accomplie.

— On devra toujours faire un compromis entre la protection de la vie privée et la qualité des données. L'anonymisation a un coût : moins une donnée est précise ou complète, moindre sera la qualité des données mais plus on gagnera en anonymisation, il faut toujours viser un juste équilibre.

ANNEXE A

FORMULAIRE DE CONSENTEMENT



FORMULAIRE D'INFORMATION ET DE CONSENTEMENT (micrologiciel)

«Médias socionumériques et Big Data:nouvelles modalités de surveillance et de gouvernance?»

PRÉAMBULE :

Vous êtes invité(e) à participer à un projet de recherche qui vise à étudier les pratiques et les perceptions des utilisateurs de médias socionumériques dans un contexte de mobilité. Avant d'accepter de participer à ce projet, il est important de prendre le temps de lire et de bien comprendre les renseignements ci-dessous. S'il y a des mots ou des sections que vous ne comprenez pas, n'hésitez pas à poser des questions.

IDENTIFICATION :

Chercheur(e) responsable du projet : André Mondoux
Tél : (514) 987-3000 poste 4828
Département, centre ou institut : École des médias, UQAM
Adresse postale : Case postale 8888, succ. Centre-Ville, Montréal, Québec, Canada H3C 3P8
Adresse courriel : mondoux.andre@uqam.ca
Membres de l'équipe : Marc Ménard, Alain Marchand, Jonathan Bonneau

OBJECTIFS DU PROJET :

Vous êtes invité(e) à prendre part à ce projet visant à mieux comprendre les dynamiques de surveillance via les fonctions de traçabilité des usagers en temps réel ainsi que leurs rôles par rapport aux pratiques et aux perceptions des usagers. Cette recherche est financée par le CRSH (Conseil de recherche en sciences humaines du Canada). Principalement, nous désirons documenter et comprendre quelles applications sont utilisées en contexte de mobilité, le moment et la durée d'utilisation et le volume de données échangées, le tout en lien avec les données de géolocalisation.

Figure A.1 Formulaire de consentement

PROCÉDURE :

Votre participation est requise pour l'intégration d'un micrologiciel de traçabilité à votre téléphone intelligent pour une période de deux mois. Développé de concert avec le Département d'informatique de l'UQAM, ce micrologiciel installé sur votre appareil mobile nous permettra de compiler en temps réel votre utilisation des médias socionumériques. Nous recueillerons les informations suivantes : applications utilisées, fonctions activées (sans leur contenu), temps d'utilisation, position géolocalisée, types (en provenance de et vers quelles applications ?), vélocité (fréquence du trafic échangé) et volume (quantité de Mo) des données produites et reçues. En aucun cas, le contenu (texte saisi au clavier, images, vidéos,sites Web visités, etc.) ne sera recueilli. Notez que les données recueillies sur support informatique ne permettront pas de vous identifier. Les données seront anonymisées dès la cueillette et un pseudonyme vous sera attribué.

AVANTAGES et RISQUES POTENTIELS :

En participant à cette recherche, vous aiderez à l'avancement des connaissances sur les pratiques des utilisateurs de médias socionumériques sur plateforme mobile, leurs effets sur les représentations sociales, la construction identitaire des individus, leur socialisation et sur une possible dynamique de banalisation de la surveillance liée à ce phénomène. Dans une certaine mesure, les entretiens réalisés pourraient vous permettre de mieux comprendre votre propre pratique en tant qu'usager.

En participant à cette recherche, vous courez des risques que nous considérons minimes. Les principaux inconvénients que nous sommes en mesure d'identifier sont reliés au monitorage des activités effectuées sur votre téléphone intelligent, lequel pourra peut-être créer un certain inconfort. Tel que spécifié ci-dessous, vous pouvez, à n'importe quel moment et sans avoir à donner de raisons spécifiques, désactiver le micrologiciel de traçabilité et ainsi mettre fin à votre participation à ce projet de recherche.

ANONYMAT ET CONFIDENTIALITÉ :

Plusieurs données non nominatives recueillies au cours de la recherche ne sont pas considérées confidentielles (applications utilisées, temps d'utilisation, volume de données intrantes et sortantes), c'est-à-dire que des extraits, correspondants aux métadonnées (types de données et non leur contenu), pourront être publiés pour documenter notre analyse et les résultats des chercheurs dans le contexte de la diffusion des résultats de la recherche. Toutes les informations susceptibles d'identifier les participants à l'étude seront ou bien retirées, ou bien modifiées de manière à assurer la confidentialité des données nominatives. Ces renseignements confidentiels ne seront accessibles que par les membres de l'équipe restreinte de recherche. Tout le matériel de recherche ainsi que votre formulaire de consentement seront conservés séparément en lieu sûr (sous clé) au Laboratoire de recherche en médias socionumériques et ludification de l'UQAM pour la durée totale du projet.

Ainsi, les noms des participants seront remplacés par des pseudonymes suivant un code associé à votre appareil mobile et ce, dès la cueillette des données. Ce code ne sera connu que du chercheur responsable du projet. Les noms resteront strictement confidentiels et ne seront transmis à aucun individu ou organisme. Les renseignements personnels seront détruits à la fin du projet de recherche (qui dure 5 ans). Seules les données qui ne permettent pas de vous identifier seront conservées après cette date.

Figure A.2 Formulaire de consentement

PARTICIPATION VOLONTAIRE et DROIT DE RETRAIT :

Votre participation à ce projet est volontaire. Cela signifie que vous acceptez de participer au projet sans aucune contrainte ou pression extérieure et que, par ailleurs, vous être libre d'y mettre fin, sans préjudice de quelque nature que ce soit et sans avoir à vous justifier. Dans ce cas, et à moins d'une directive contraire de votre part, les documents vous concernant seront détruits.

Votre accord à participer implique également que vous acceptez que l'équipe de recherche puisse utiliser aux fins de la présente recherche (articles, mémoires et thèses des étudiants membres de l'équipe, conférences et communications scientifiques) les renseignements recueillis à la condition qu'aucune information permettant de vous identifier ne soit divulguée publiquement à moins d'un consentement explicite de votre part.

Si, au cours de l'étude, de nouvelles informations ou des changements aux procédures de recherche susceptibles de vous faire reconsidérer votre décision de participer à l'étude surviennent, vous en serez avisé.

COMPENSATION FINANCIÈRE ou AUTRE :

Il est entendu que vous ne recevrez aucune compensation financière pour contribution au projet.

CLAUSE DE RESPONSABILITÉ :

En acceptant de participer à ce projet, vous ne renoncez à aucun de vos droits ni ne libérez les chercheurs, le commanditaire ou les institutions impliquées de leurs obligations légales et professionnelles.

DES QUESTIONS SUR LE PROJET OU SUR VOS DROITS ?

Pour des questions additionnelles sur le projet, sur votre participation et sur vos droits en tant que participant de recherche, ou pour vous retirer du projet, vous pouvez communiquer avec :

André Mondoux
Téléphone : (514) 987-3000, poste 4828
Courriel : mondoux.andre@uqam.ca

Le Comité institutionnel d'éthique de la recherche avec des êtres humains de l'UQAM a approuvé le projet de recherche auquel vous allez participer. Pour des informations concernant les responsabilités de l'équipe de recherche au plan de l'éthique de la recherche avec des êtres humains ou pour formuler une plainte, vous pouvez contacter la présidence du Comité, par l'intermédiaire de son secrétariat au numéro (514) 987-3000 # 7753 ou par courriel à CIEREH@UQAM.CA

Figure A.3 Formulaire de consentement

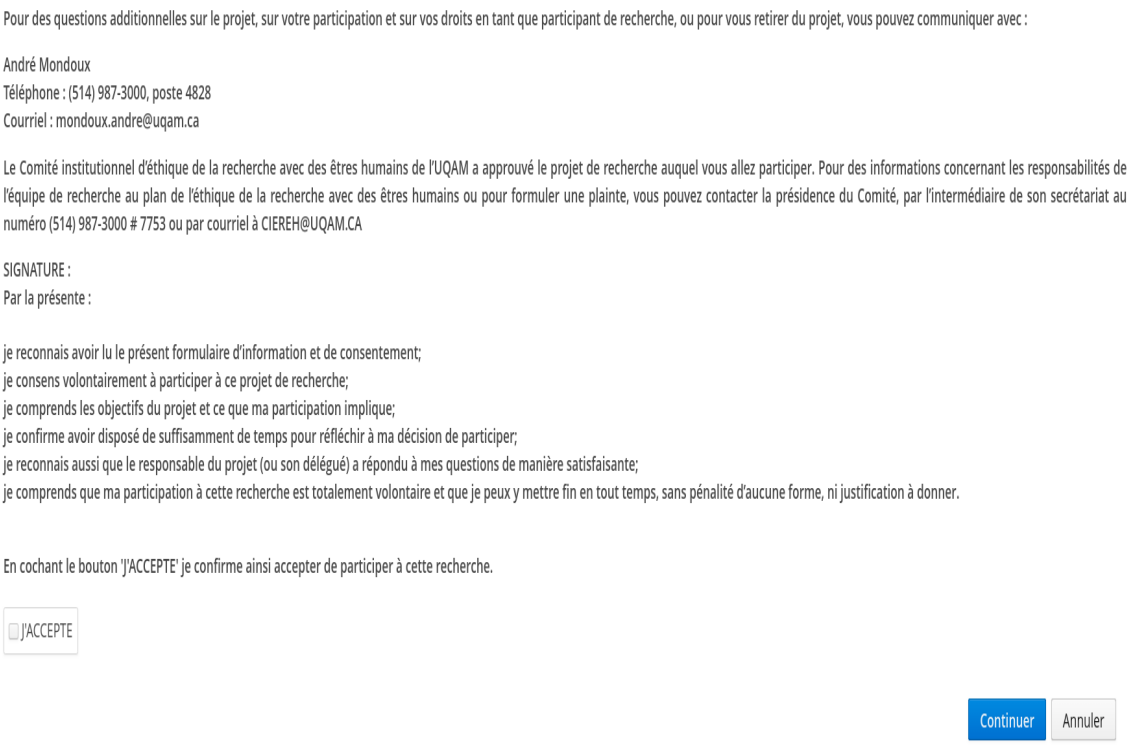


Figure A.4 Formulaire de consentement

ANNEXE B

QUESTIONNAIRE SUR LA MOBILITÉ

Figure B.1 Page d'accueil

Figure B.2 Volet Technique

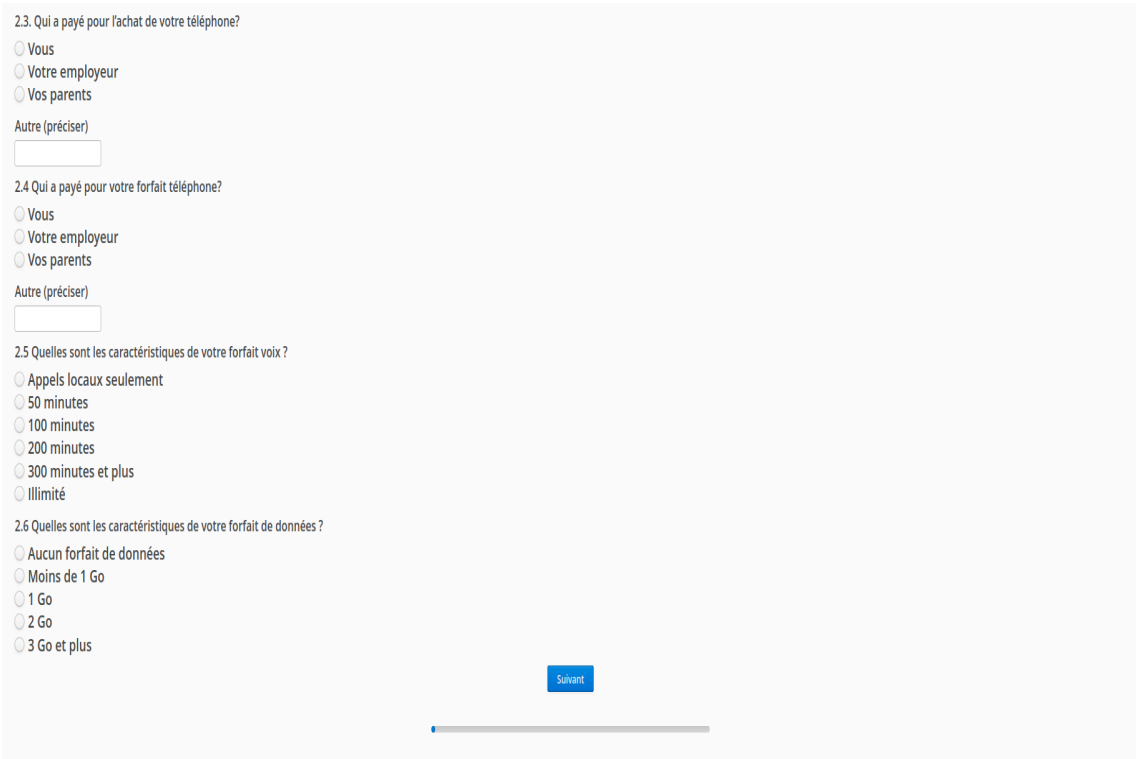
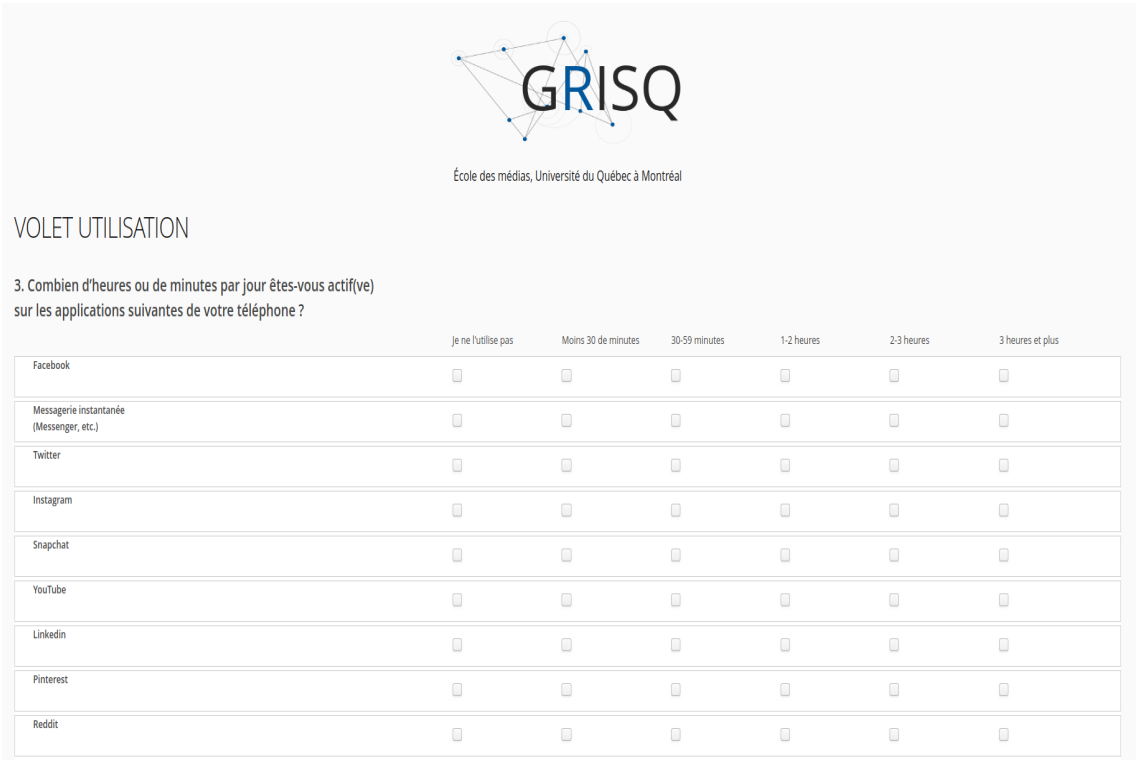


Figure B.3 Volet Technique



GRISQ

École des médias, Université du Québec à Montréal

VOLET UTILISATION

3. Combien d'heures ou de minutes par jour êtes-vous actif(ve) sur les applications suivantes de votre téléphone ?

	Je ne l'utilise pas	Moins 30 de minutes	30-59 minutes	1-2 heures	2-3 heures	3 heures et plus
Facebook	☐	☐	☐	☐	☐	☐
Messagerie instantanée (Messenger, etc.)	☐	☐	☐	☐	☐	☐
Twitter	☐	☐	☐	☐	☐	☐
Instagram	☐	☐	☐	☐	☐	☐
Snapchat	☐	☐	☐	☐	☐	☐
YouTube	☐	☐	☐	☐	☐	☐
Linkedin	☐	☐	☐	☐	☐	☐
Pinterest	☐	☐	☐	☐	☐	☐
Reddit	☐	☐	☐	☐	☐	☐

Figure B.4 Volet Utilisation

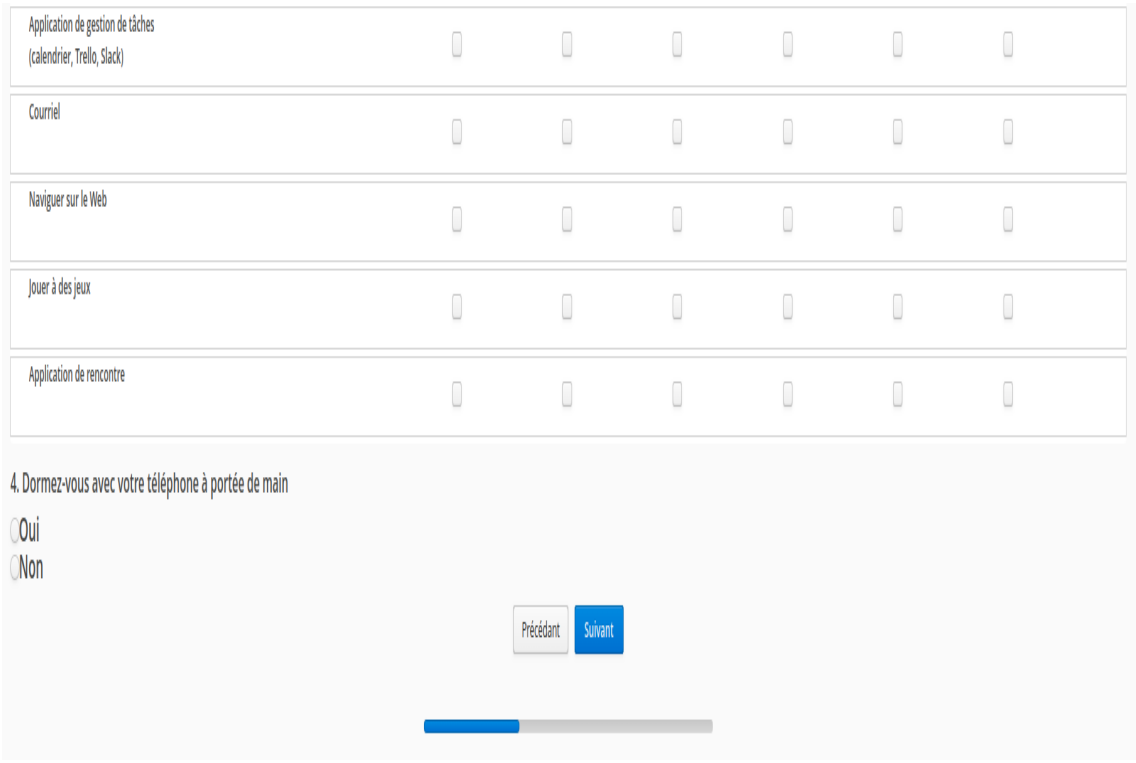


Figure B.5 Volet Utilisation

Figure B.6 Volet Travail

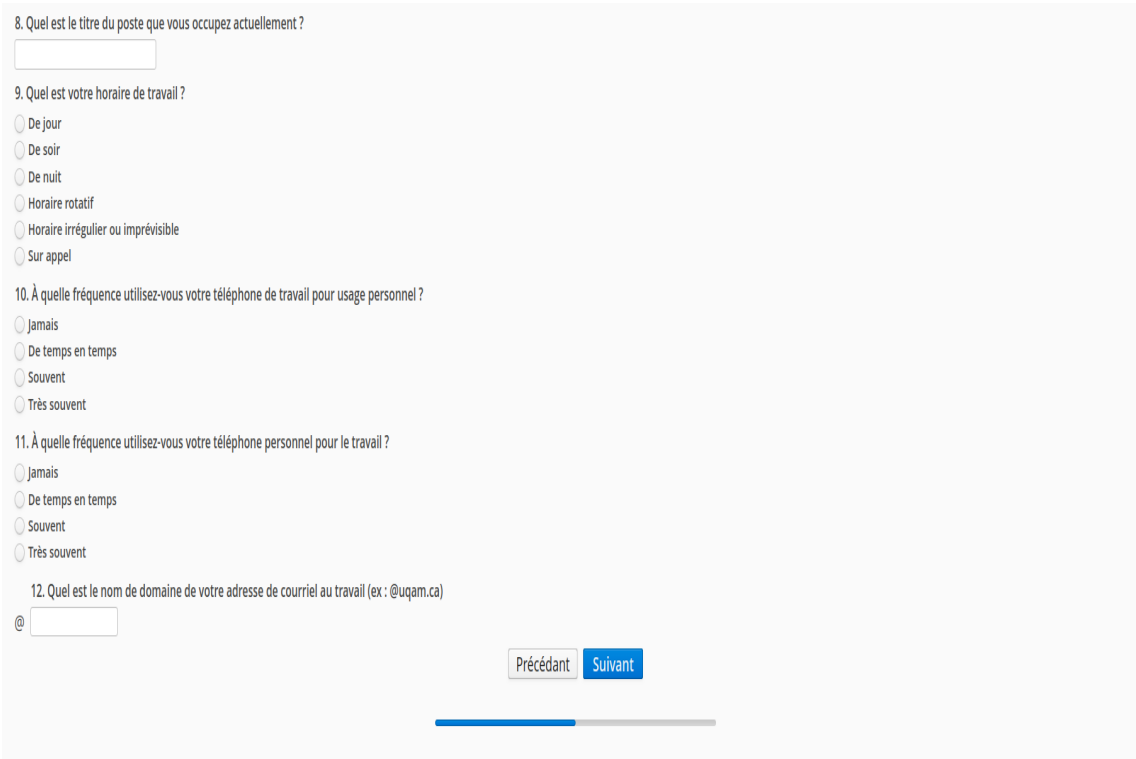


Figure B.7 Volet Travail

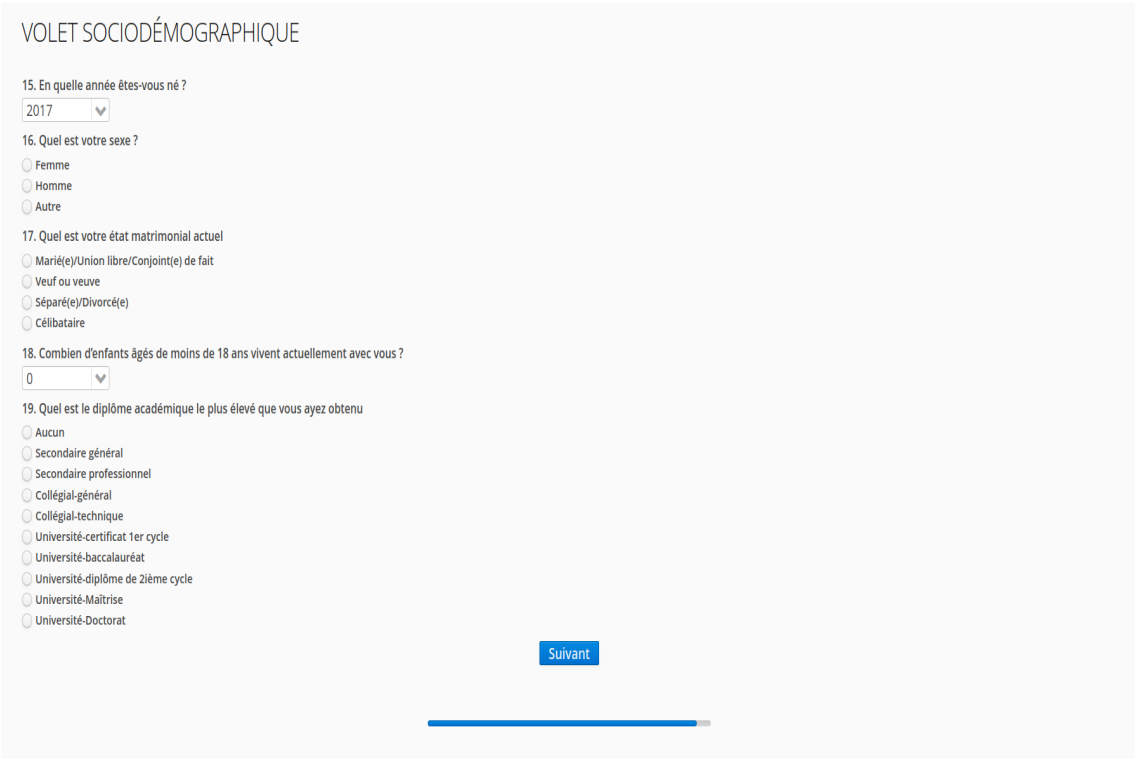


Figure B.8 Volet Sociodémographique

Figure B.9 Remerciements

ANNEXE C

USAGES MOBILES

Figure C.1 Vue principale Usage Mobile

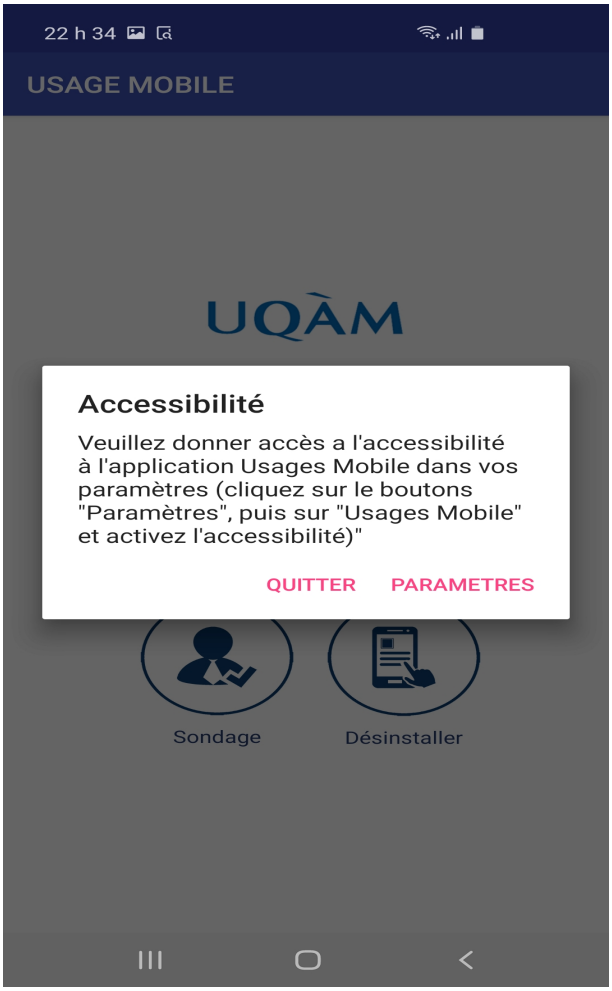


Figure C.2 Autorisation à l'accessibilité

Figure C.3 Autorisation à l'accessibilité

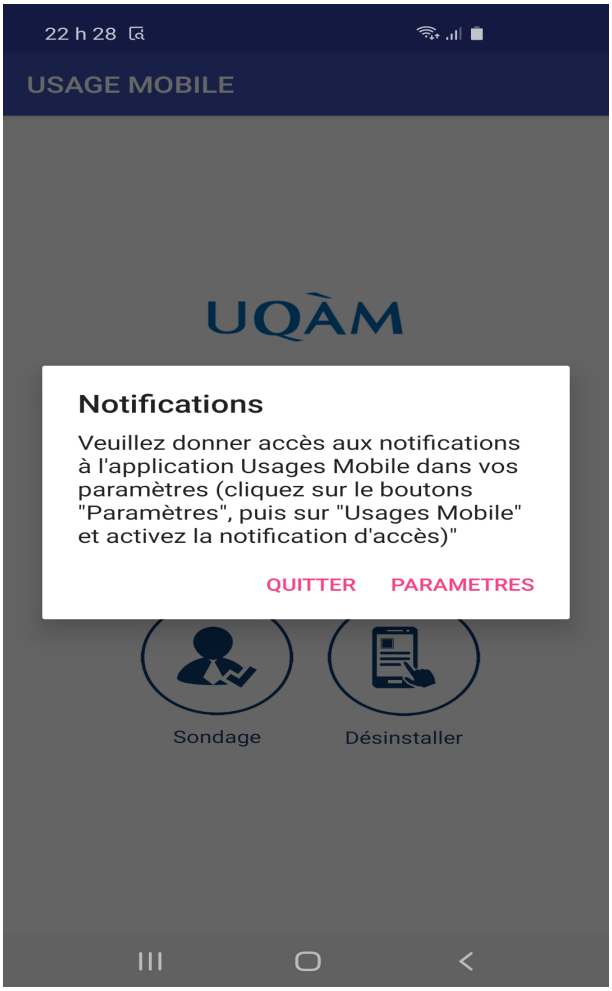


Figure C.4 Autorisation aux notifications

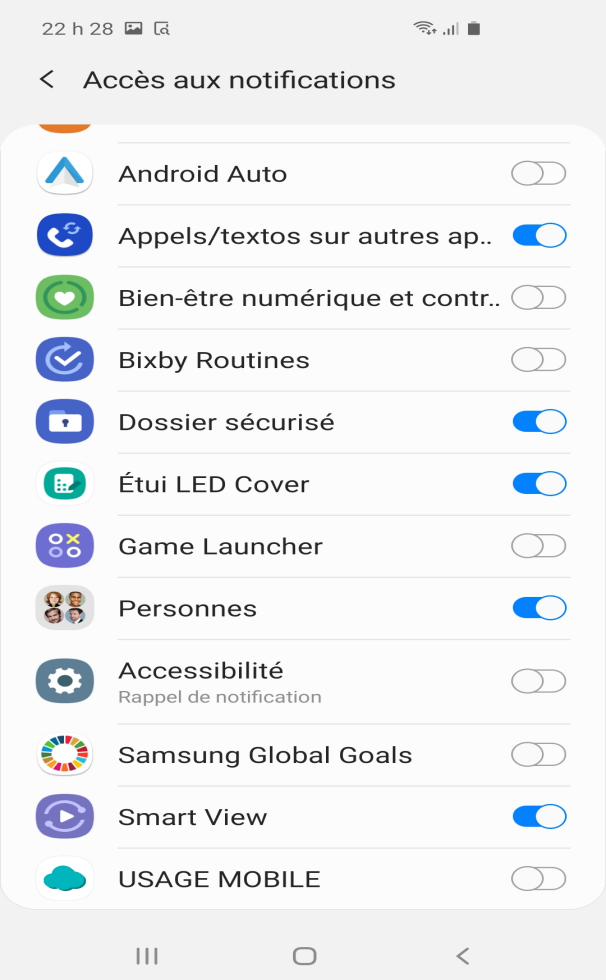


Figure C.5 Autorisation aux notifications

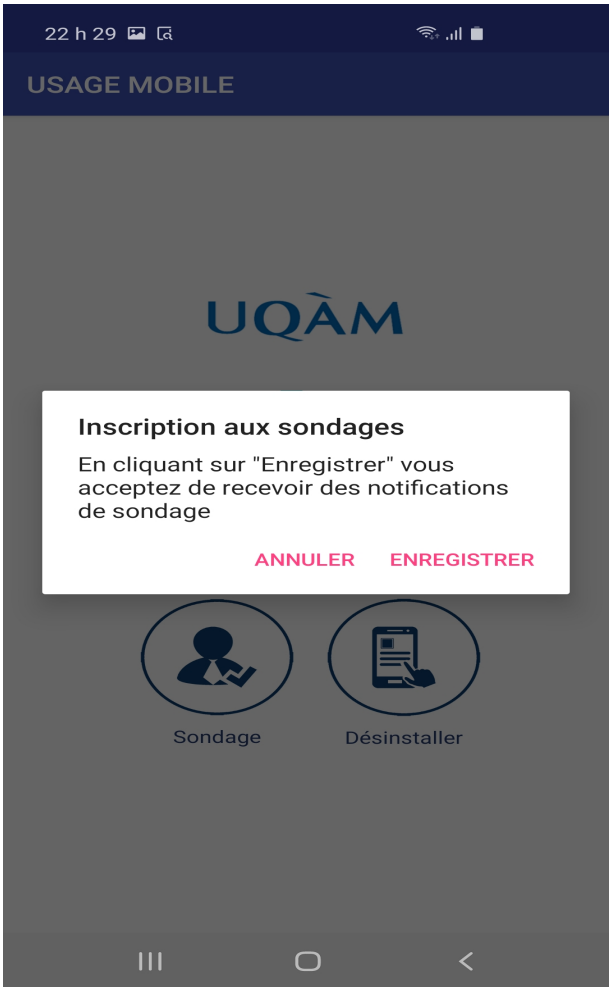


Figure C.6 Inscription aux sondages

Figure C.7 Suppression de Usage Mobile

ANNEXE D

LES DONNÉES BRUTES DU PARTICIPANT CHOISIT AU HASARD AU CHAPITRE 5

```
mysql> SELECT  distinct (p.serial), p.modele, p.uuid FROM participant AS p  RIGHT JOIN application AS a ON p.serial = a.Serial where p.serial is NOT NULL;
+------------------+--------------+--------+
| serial           | modele       | uuid   |
+------------------+--------------+--------+
| ZY2248JPLM       | Moto G Play  | 621fb8 |
| ZY223NMW4F       | XT1650       | bec81e |
| LGM320d94ce438   | LG-M320G     | db90e4 |
| ce12160c3112d31504 | SM-G935W8  | 8137cb |
| LGM320f0f7d09a   | LG-M320G     | 296d0f |
| ZY2238P4VR       | Moto G (4)   | b7d441 |
| TA3970A8MJ       | MotoE2(4G-LTE) | 2883bd |
| ZY222ZC9WV       | MotoG3       | 19f922 |
| ce12160c658d6c2d01 | SM-G935W8  | e652fe |
| RF8MC0L1CSL      | SM-G973F     | 4fcdc8 |
+------------------+--------------+--------+
10 rows in set (4.18 sec)
```

Figure D.1 Requêtes affichant tous les participants qui ont installés USAGES MOBILES

```
mysql> select * from form where uuid= '621fb8';
+--------+------------------------------------------------------------------------------------------------------+------------------
---------------+--------------------+--------+---------------+---------------+----------+
| idForm | question                                                                                             | answer
             | type               | uuid   | date          | createdate    | chercheur |
+--------+------------------------------------------------------------------------------------------------------+------------------
---------------+--------------------+--------+---------------+---------------+----------+
|    454 | 2.1. Combien de téléphones avez-eu en votre possession au cours de 5 dernières années ?              | Quatre
             | TECHNIQUE          | 621fb8 | 1517537444344 | 1517537444387 | Mesolu    |
|    455 | 2.2. Depuis combien de temps possédez-vous votre téléphone actuel ?                                  | Moins d'un an
             | TECHNIQUE          | 621fb8 | 1517537444412 | 1517537444443 | Mesolu    |
|    456 | 2.5 Quelles sont les caractéristiques de votre forfait voix ?                                        | Appels locaux seul
ement        | TECHNIQUE          | 621fb8 | 1517537444466 | 1517537444499 | Mesolu    |
|    457 | 2.6 Quelles sont les caractéristiques de votre forfait de données ?                                  | Aucun forfait de d
onnées       | TECHNIQUE          | 621fb8 | 1517537444527 | 1517537444558 | Mesolu    |
|    458 | 1. Veuillez indiquer les appareils que vous utilisez actuellement.                                   | [Téléphone cellula
ire]         | TECHNIQUE          | 621fb8 | 1517537444587 | 1517537444618 | Mesolu    |
|    459 | 2.3. Qui a payé pour l’achat de votre téléphone?                                                     | Vous
             | TECHNIQUE          | 621fb8 | 1517537444641 | 1517537444672 | Mesolu    |
|    460 | 2.4 Qui a payé pour votre forfait téléphone?                                                         | Vous
             | TECHNIQUE          | 621fb8 | 1517537444694 | 1517537444725 | Mesolu    |
|    461 | <b>Instagram</b>                                                                                     | Moins 30 de minute
s            | UTILISATION        | 621fb8 | 1517537514024 | 1517537514064 | Mesolu    |
|    462 | <b>Application de gestion de tâches <br>(calendrier, Trello, Slack)</b>                              | Moins 30 de minute
s            | UTILISATION        | 621fb8 | 1517537514092 | 1517537514124 | Mesolu    |
|    463 | <b>Pinterest</b>                                                                                     | Je ne l'utilise pa
s            | UTILISATION        | 621fb8 | 1517537514149 | 1517537514184 | Mesolu    |
|    464 | <b>Twitter</b>                                                                                       | Moins 30 de minute
s            | UTILISATION        | 621fb8 | 1517537514212 | 1517537514244 | Mesolu    |
|    465 | <b>Facebook</b>                                                                                      | 30-59 minutes
             | UTILISATION        | 621fb8 | 1517537514268 | 1517537514299 | Mesolu    |
|    466 | <b>Jouer à des jeux</b>                                                                              | 30-59 minutes
             | UTILISATION        | 621fb8 | 1517537514323 | 1517537514354 | Mesolu    |
|    467 | <b>Courriel</b>                                                                                      | 2-3 heures
             | UTILISATION        | 621fb8 | 1517537514375 | 1517537514406 | Mesolu    |
|    468 | <b>Naviguer sur le Web</b>                                                                           | 2-3 heures
             | UTILISATION        | 621fb8 | 1517537514430 | 1517537514465 | Mesolu    |
|    469 | 4. Dormez-vous avec votre téléphone à portée de main                                                 | Oui
             | UTILISATION        | 621fb8 | 1517537514487 | 1517537514518 | Mesolu    |
|    470 | <b>Messagerie instantanée <br>(Messenger, etc.)</b>                                                  | 1-2 heures
             | UTILISATION        | 621fb8 | 1517537514547 | 1517537514578 | Mesolu    |
|    471 | <b>Reddit</b>                                                                                        | Je ne l'utilise pa
s            | UTILISATION        | 621fb8 | 1517537514601 | 1517537514635 | Mesolu    |
|    472 | <b>Application de rencontre</b>                                                                      | Je ne l'utilise pa
s            | UTILISATION        | 621fb8 | 1517537514659 | 1517537514721 | Mesolu    |
|    473 | <b>YouTube</b>                                                                                       | 1-2 heures
```

Figure D.2 Requêtes affichant les réponses au questionnaire par le participant choisi au hasard

```
            | UTILISATION        | 621fb8 | 1517537514744 | 1517537514778 | Mesolu   |
|   474 | <b>Linkedin</b>                                                                                                                   | Je ne l'utilise pa
s           | UTILISATION        | 621fb8 | 1517537514801 | 1517537514833 | Mesolu   |
|   475 | <b>Snapchat</b>                                                                                                                   | Je ne l'utilise pa
s           | UTILISATION        | 621fb8 | 1517537514855 | 1517537514885 | Mesolu   |
|   476 | 6. Si vous n'occupez pas d'emploi, quelle est, actuellement, votre occupation principale ?                                        |
            | TRAVAIL            | 621fb8 | 1517537637994 | 1517537638036 | Mesolu   |
|   477 | 5. Occupez-vous actuellement un emploi ?                                                                                          | Oui
            | TRAVAIL            | 621fb8 | 1517537638063 | 1517537638098 | Mesolu   |
|   478 | 10. À quelle fréquence utilisez-vous votre téléphone de travail pour usage personnel ?                                            | De temps en temps
            | TRAVAIL            | 621fb8 | 1517537638120 | 1517537638154 | Mesolu   |
|   479 | 12. Quel est le nom de domaine de votre adresse de courriel au travail (ex : @uqam.ca)                                            | atalou.com
            | TRAVAIL            | 621fb8 | 1517537638177 | 1517537638212 | Mesolu   |
|   480 | 7. En moyenne, combien d'heures par semaine travaillez-vous à votre emploi ? <br>(Par exemple, si vous travaillez 17 heures 30, inscrivez 17.5) | 8 heures 0 minutes
            | TRAVAIL            | 621fb8 | 1517537638234 | 1517537638269 | Mesolu   |
|   481 | 9. Quel est votre horaire de travail ?                                                                                            | De jour
            | TRAVAIL            | 621fb8 | 1517537638291 | 1517537638326 | Mesolu   |
|   482 | 8. Quel est le titre du poste que vous occupez actuellement ?                                                                     | Analyste Programme
ur          | TRAVAIL            | 621fb8 | 1517537638349 | 1517537638383 | Mesolu   |
|   483 | 11. À quelle fréquence utilisez-vous votre téléphone personnel pour le travail ?                                                  | Souvent
            | TRAVAIL            | 621fb8 | 1517537638404 | 1517537638439 | Mesolu   |
|   484 | <b>Avez-vous l'impression que chaque<br> heure de travail vous fatigue ?</b>                                                      | Souvent
            | ÉPUISEMENT         | 621fb8 | 1517537709780 | 1517537709821 | Mesolu   |
|   485 | <b>Votre travail est-il épuisant <br> émotionnellement ?</b>                                                                      | Quelque peu
            | ÉPUISEMENT         | 621fb8 | 1517537709848 | 1517537709882 | Mesolu   |
|   486 | <b>Vous sentez-vous épuisé(e) à cause de <br> votre travail?</b>                                                                  | À un niveau élevé
            | ÉPUISEMENT         | 621fb8 | 1517537709906 | 1517537709940 | Mesolu   |
|   487 | <b>Votre travail suscite-t-il chez vous un <br> sentiment de frustration ?</b>                                                    | À un niveau élevé
            | ÉPUISEMENT         | 621fb8 | 1517537709966 | 1517537710001 | Mesolu   |
|   488 | <b>Avez-vous suffisamment d'énergie à <br> consacrer à votre  famille et à <br> vos amis pendant vos moments de loisir ?</b>      | Rarement
            | ÉPUISEMENT         | 621fb8 | 1517537710025 | 1517537710061 | Mesolu   |
|   489 | <b>Êtes-vous épuisée le matin <br> (ou le soir ou la nuit,si vous  travaillez <br>  sur un autre horaire) à l'idée <br> de faire  une autre journée de travail ?</b> | Souvent
            | ÉPUISEMENT         | 621fb8 | 1517537710084 | 1517537710119 | Mesolu   |
|   490 | <b>Vous sentez-vous exténué(e) après une <br> journée de travail ?</b>                                                            | Souvent
            | ÉPUISEMENT         | 621fb8 | 1517537710142 | 1517537710177 | Mesolu   |
|   491 | 14. Quel mode de transport utilisez-vous, habituellement, pour vous <br> rendre à votre lieu de travail ? (n'indiquez qu'une seule possibilité) | Surtout le transpo
rt en commun | DÉPLACEMENT        | 621fb8 | 1517537797274 | 1517537797334 | Mesolu   |
|   492 | 13. Combien de temps, en moyenne, mettez-vous pour arriver à votre <br>lieu de travail ?                                          | 0 heures 45 minute
s           | DÉPLACEMENT        | 621fb8 | 1517537797360 | 1517537797394 | Mesolu   |
|   493 | 15. En quelle année êtes-vous né ?                                                                                                | 1989
            | SOCIODÉMOGRAPHIQUE | 621fb8 | 1517537824294 | 1517537824334 | Mesolu   |
|   494 | 16. Quel est votre sexe ?                                                                                                         | Homme
            | SOCIODÉMOGRAPHIQUE | 621fb8 | 1517537824359 | 1517537824393 | Mesolu   |
|   495 | 18. Combien d'enfants âgés de moins de 18 ans vivent actuellement avec vous ?                                                     | 0
            | SOCIODÉMOGRAPHIQUE | 621fb8 | 1517537824417 | 1517537824450 | Mesolu   |
|   496 | 19. Quel est le diplôme académique le plus élevé que vous ayez obtenu                                                             | Université-certifi
cat 1er cycle | SOCIODÉMOGRAPHIQUE | 621fb8 | 1517537824475 | 1517537824507 | Mesolu   |
|   497 | 17. Quel est votre état matrimonial actuel                                                                                        | Célibataire
            | SOCIODÉMOGRAPHIQUE | 621fb8 | 1517537824530 | 1517537824564 | Mesolu   |
+-------+-----------------------------------------------------------------------------------------------------------------------------------+-------------------
------------+--------------------+--------+---------------+---------------+----------+
44 rows in set (0.01 sec)
```

Figure D.3 Requêtes affichant les réponses au questionnaire par le participant choisit au hasard

```
mysql> select idWifi, serial, type, from_unixtime(date/1000, "%Y-%m-%d %H:%i:%s")  from wifi limit 1000;
+--------+-------------+-----------------------------------------+-----------------------------------------+
| idWifi | serial      | type                                    | from_unixtime(date/1000, "%Y-%m-%d %H:%i:%s") |
+--------+-------------+-----------------------------------------+-----------------------------------------+
|   1584 | HT5B4BE10292 | Fermeture Ecran                         | 2018-02-01 20:34:02                     |
|   1585 | HT5B4BE10292 | Fermeture Ecran                         | 2018-02-01 20:36:12                     |
|   1586 | HT5B4BE10292 | Fermeture Ecran                         | 2018-02-01 20:36:25                     |
|   1587 | HT5B4BE10292 | Fermeture Ecran                         | 2018-02-01 20:41:23                     |
|   1588 | HT5B4BE10292 | Ouverture Settings                      | 2018-02-01 20:41:52                     |
|   1591 | HT5B4BE10292 | Ouverture Settings                      | 2018-02-01 20:41:54                     |
|   1592 | HT5B4BE10292 | Clique Settings                         | 2018-02-01 20:41:55                     |
|   1593 | HT5B4BE10292 | Ouverture Settings                      | 2018-02-01 20:41:55                     |
|   1595 | HT5B4BE10292 | Clique Settings                         | 2018-02-01 20:41:56                     |
|   1596 | HT5B4BE10292 | Ouverture Settings                      | 2018-02-01 20:41:56                     |
|   1597 | HT5B4BE10292 | Clique Settings                         | 2018-02-01 20:41:57                     |
|   1598 | HT5B4BE10292 | Ouverture Settings                      | 2018-02-01 20:41:57                     |
|   1599 | HT5B4BE10292 | Ouverture USAGE MOBILE                  | 2018-02-01 20:41:58                     |
|   1600 | HT5B4BE10292 | Ouverture Sense Home                    | 2018-02-01 20:42:00                     |
|   1601 | HT5B4BE10292 | Ouverture Sense Home                    | 2018-02-01 20:42:00                     |
|   1603 | HT5B4BE10292 | Clique Sense Home                       | 2018-02-01 20:43:26                     |
|   1604 | HT5B4BE10292 | Ouverture People                        | 2018-02-01 20:43:26                     |
|   1605 | HT5B4BE10292 | Clique People                           | 2018-02-01 20:43:27                     |
|   1606 | HT5B4BE10292 | Ouverture Phone                         | 2018-02-01 20:43:27                     |
|   1607 | HT5B4BE10292 | Appel envoye                            | 2018-02-01 20:43:28                     |
|   1608 | HT5B4BE10292 | Clique Phone                            | 2018-02-01 20:43:28                     |
|   1609 | HT5B4BE10292 | Ouverture People                        | 2018-02-01 20:43:29                     |
|   1610 | HT5B4BE10292 | Appel envoye                            | 2018-02-01 20:43:28                     |
|   1611 | HT5B4BE10292 | Ajout Contact                           | 2018-02-01 20:43:31                     |
|   1612 | HT5B4BE10292 | Ouverture Sense Home                    | 2018-02-01 20:43:33                     |
|   1613 | HT5B4BE10292 | Ouverture Sense Home                    | 2018-02-01 20:43:33                     |
|   1614 | HT5B4BE10292 | Clique Sense Home                       | 2018-02-01 20:43:35                     |
|   1615 | HT5B4BE10292 | Ouverture Messages                      | 2018-02-01 20:43:35                     |
|   1616 | HT5B4BE10292 | Clique Messages                         | 2018-02-01 20:43:36                     |
|   1618 | HT5B4BE10292 | Clique Messages                         | 2018-02-01 20:43:42                     |
|   1620 | HT5B4BE10292 | Clique Messages                         | 2018-02-01 20:43:42                     |
|   1622 | HT5B4BE10292 | Ecrit du Texte sur  Messages            | 2018-02-01 20:43:43                     |
|   1623 | HT5B4BE10292 | Ecrit du Texte sur  Messages            | 2018-02-01 20:43:44                     |
|   1624 | HT5B4BE10292 | Clique Messages                         | 2018-02-01 20:43:44                     |
|   1625 | HT5B4BE10292 | Texto envoye                            | 2018-02-01 20:43:44                     |
|   1627 | HT5B4BE10292 | Ecrit du Texte sur  Messages            | 2018-02-01 20:43:45                     |
|   1629 | HT5B4BE10292 | Ajout Contact                           | 2018-02-01 20:43:45                     |
|   1630 | HT5B4BE10292 | Ouverture Sense Home                    | 2018-02-01 20:43:45                     |
|   1631 | HT5B4BE10292 | Ouverture Sense Home                    | 2018-02-01 20:43:46                     |
|   1632 | HT5B4BE10292 | Ouverture Sense Home                    | 2018-02-01 20:44:11                     |
|   1633 | HT5B4BE10292 | Clique Sense Home                       | 2018-02-01 20:44:11                     |
|   1649 | HT5B4BE10292 | Clique Sense Home                       | 2018-02-01 20:44:14                     |
|   1651 | HT5B4BE10292 | Fermeture Ecran                         | 2018-02-01 20:44:14                     |
|   1652 | HT5B4BE10292 | Ouverture USAGE MOBILE                  | 2018-02-01 20:44:14                     |
|   1653 | HT5B4BE10292 | Clique USAGE MOBILE                     | 2018-02-01 20:44:36                     |
|   1654 | HT5B4BE10292 | Ouverture USAGE MOBILE                  | 2018-02-01 20:44:37                     |
|   1655 | HT5B4BE10292 | Clique USAGE MOBILE                     | 2018-02-01 20:44:37                     |
|   1656 | HT5B4BE10292 | Ouverture Sense Home                    | 2018-02-01 20:44:59                     |
```

Figure D.4 Requêtes affichant les milles premiers usages fait par le participant choisit au hasard

2154	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:35:21
2155	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:35:21
2156	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:35:21
2157	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:35:22
2158	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:35:22
2159	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:35:23
2160	ZY2248JPLM	Ajout Contact	2018-02-01 21:35:23
2162	ZY2248JPLM	Clique WhatsApp	2018-02-01 21:35:24
2163	ZY2248JPLM	Ouverture WhatsApp	2018-02-01 21:35:24
2170	ZY2248JPLM	Clique WhatsApp	2018-02-01 21:35:26
2171	ZY2248JPLM	Ouverture WhatsApp	2018-02-01 21:35:26
2173	ZY2248JPLM	Clique WhatsApp	2018-02-01 21:35:32
2178	ZY2248JPLM	Ouverture USAGE MOBILE	2018-02-01 21:35:33
2179	ZY2248JPLM	Clique USAGE MOBILE	2018-02-01 21:35:34
2180	ZY2248JPLM	Ouverture USAGE MOBILE	2018-02-01 21:35:34
2181	ZY2248JPLM	Clique USAGE MOBILE	2018-02-01 21:35:35
2182	ZY2248JPLM	Ouverture USAGE MOBILE	2018-02-01 21:35:35
2185	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:13
2186	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:13
2187	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:13
2188	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:13
2189	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:14
2190	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:16
2191	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:17
2192	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:17
2193	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:17
2194	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:18
2195	ZY2248JPLM	Ajout Contact	2018-02-01 21:36:21
2197	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:21
2198	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:21
2199	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:22
2200	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:22
2201	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:22
2202	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:22
2203	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:22
2204	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:23
2205	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:23
2206	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:24
2207	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:24
2208	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:25
2209	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:25
2210	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:25
2211	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:25
2212	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:25
2213	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:26
2214	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:26
2215	ZY2248JPLM	Ajout Contact	2018-02-01 21:36:27
2217	ZY2248JPLM	Clique WhatsApp	2018-02-01 21:36:27
2218	ZY2248JPLM	Ouverture WhatsApp	2018-02-01 21:36:28
2225	ZY2248JPLM	Clique WhatsApp	2018-02-01 21:36:29
2226	ZY2248JPLM	Ouverture WhatsApp	2018-02-01 21:36:29
2228	ZY2248JPLM	Clique WhatsApp	2018-02-01 21:36:31
2230	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:32
2231	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:32
2232	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:32
2233	ZY2248JPLM	Ecrit du Texte sur WhatsApp	2018-02-01 21:36:33

Figure D.5 Requêtes affichant les milles premiers usages fait par le participant choisit au hasard

BIBLIOGRAPHIE

Aggarwal, C. C. et Philip, S. Y. (2008a). A general survey of privacy-preserving data mining models and algorithms. In *Privacy-preserving data mining* 11–52. Springer.

Aggarwal, C. C. et Philip, S. Y. (2008b). *Privacy-preserving data mining : models and algorithms*. Springer Science & Business Media.

Agrawal, H., Cochinwala, M. et Horgan, J. R. (2014). Automated determination of quasi-identifiers using program analysis. US Patent 8,661,423.

Agrawal, R. et Srikant, R. (2000). Privacy-preserving data mining. Dans *ACM Sigmod Record*, volume 29, 439–450. ACM.

Akoka, J., Comyn-Wattiau, I., Du Mouza, C., Fadili, H., Lammari, N., Metais, E. et Cherfi, S. S.-S. (2014). A semantic approach for semi-automatic detection of sensitive data. *Information Resources Management Journal (IRMJ)*, *27*(4), 23–44.

Arfaoui, S., Belmekki, A. et Mezrioui, A. (2020). A qualitative-driven study of irreversible data anonymizing techniques in databases. Dans *Proceedings of the 13th International Conference on Intelligent Systems : Theories and Applications*, 1–6.

Bambauer, J., Muralidhar, K. et Sarathy, R. (2013). Fool's gold : an illustrated critique of differential privacy. *Vand. J. Ent. & Tech. L.*, *16*, 701.

Bonchi, F., Lakshmanan, L. V. et Wang, H. W. (2011). Trajectory anonymity in publishing personal mobility data. *ACM Sigkdd Explorations Newsletter*, *13*(1), 30–42.

Bun, M. et Steinke, T. (2016). Concentrated differential privacy : Simplifications, extensions, and lower bounds. Dans *Theory of Cryptography Conference*, 635–658. Springer.

Cavoukian, A. *et al.* (2009). Privacy by design : The 7 foundational principles. *Information and privacy commissioner of Ontario, Canada*, *5*.

Claerhout, B. et DeMoor, G. (2005). Privacy protection for clinical and genomic data : The use of privacy-enhancing techniques in medicine. *International Journal of Medical Informatics*, *74*(2-4), 257–265.

Dalenius, T. (1977). Towards a methodology for statistical disclosure control. *statistik Tidskrift*, *15*(429-444), 2–1.

Domingo-Ferrer, J. (2008). A survey of inference control methods for privacy-preserving data mining. In *Privacy-preserving data mining* 53–80. Springer.

Dwork, C. (2011). Differential privacy. *Encyclopedia of Cryptography and Security*, 338–340.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I. et Naor, M. (2006a). Our data, ourselves : Privacy via distributed noise generation. Dans *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, 486–503. Springer.

Dwork, C., McSherry, F., Nissim, K. et Smith, A. (2006b). Calibrating noise to sensitivity in private data analysis. Dans *Theory of cryptography conference*, 265–284. Springer.

Dwork, C., Roth, A. *et al.* (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, *9*(3-4), 211–407.

Enck, W., Gilbert, P., Han, S., Tendulkar, V., Chun, B.-G., Cox, L. P., Jung, J., McDaniel, P. et Sheth, A. N. (2014). Taintdroid : an information-flow tracking system for realtime privacy monitoring on smartphones. *ACM Transactions on Computer Systems (TOCS)*, *32*(2), 5.

Fienberg, S. E. et McIntyre, J. (2004). Data swapping : Variations on a theme by dalenius and reiss. Dans *International Workshop on Privacy in Statistical Databases*, 14–29. Springer.

Fung, B. C., Wang, K., Chen, R. et Yu, P. S. (2010a). Privacy-preserving data publishing : A survey of recent developments. *ACM Computing Surveys (Csur)*, *42*(4), 1–53.

Fung, B. C., Wang, K., Fu, A. W.-C. et Philip, S. Y. (2010b). *Introduction to privacy-preserving data publishing : Concepts and techniques*. Chapman and Hall/CRC.

Garfinkel, S. L., Abowd, J. M. et Powazek, S. (2018). Issues encountered deploying differential privacy. Dans *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, 133–137.

Garfinkel, S. L. et Leclerc, P. (2020). Randomness concerns when deploying differential privacy. Dans *Proceedings of the 19th Workshop on Privacy in the Electronic Society*, 73–86.

Ghinita, G., Kalnis, P. et Tao, Y. (2010). Anonymous publication of sensitive transactional data. *IEEE Transactions on Knowledge and Data Engineering*, *23*(2), 161–174.

Gunawan, D. et Mambo, M. (2019). Data anonymization for hiding personal tendency in set-valued database publication. *Future Internet*, *11*(6), 138.

Holohan, N., Antonatos, S., Braghin, S. et Mac Aonghusa, P. (2018). The bounded laplace mechanism in differential privacy. *arXiv preprint arXiv :1808.10410*.

Isaac, H. (2016). Données, valeur et business models.

Jia, X., Pan, C., Xu, X., Zhu, K. Q. et Lo, E. (2014). ρ-uncertainty anonymization by partial suppression. Dans *International Conference on Database Systems for Advanced Applications*, 188–202. Springer.

Kataoka, H., Ogawa, Y., Echizen, I., Kuboyama, T. et Yoshiura, H. (2014). Effects of external information on anonymity and role of transparency with example of social network de-anonymisation. Dans *2014 Ninth International Conference on Availability, Reliability and Security*, 461–467. IEEE.

Ke, H., Fu, A., Yu, S. et Chen, S. (2018). Aq-dp : A new differential privacy scheme based on quasi-identifier classifying in big data. Dans *2018 IEEE Global Communications Conference (GLOBECOM)*, 1–6. `http://dx.doi.org/10.1109/GLOCOM.2018.8647941`

Kiran, P. et Kavya, N. (2012). A survey on methods, attacks and metric for privacy preserving data publishing. *International Journal of Computer Applications*, *53*(18).

Kuhn, E., van der Meer, C., Owen, J. E., Hoffman, J. E., Cash, R., Carrese, P., Olff, M., Bakker, A., Schellong, J., Lorenz, P. *et al.* (2018). Ptsd coach around the world. *Mhealth*, *4*.

Kushida, C. A., Nichols, D. A., Jadrnicek, R., Miller, R., Walsh, J. K. et Griffin, K. (2012). Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical care*, *50*(Suppl), S82.

Langheinrich, M. (2001). Privacy by design—principles of privacy-aware ubiquitous systems. Dans *International conference on Ubiquitous Computing*, 273–

291. Springer.

Lee, H., Kim, S., Kim, J. W. et Chung, Y. D. (2017). Utility-preserving anonymization for health data publishing. *BMC medical informatics and decision making*, *17*(1), 1–12.

Levallois-Barth, C. (2018). La notion de data protection by design. *Revue de la Gendarmerie Nationale*, (263), https–www.

Li, N., Li, T. et Venkatasubramanian, S. (2007). t-closeness : Privacy beyond k-anonymity and l-diversity. Dans *2007 IEEE 23rd International Conference on Data Engineering*, 106–115. IEEE.

Li, N., Li, T. et Venkatasubramanian, S. (2009). Closeness : A new privacy measure for data publishing. *IEEE Transactions on Knowledge and Data Engineering*, *22*(7), 943–956.

Machanavajjhala, A., Gehrke, J., Kifer, D. et Venkitasubramaniam, M. (2006). l-diversity : Privacy beyond k-anonymity. Dans *22nd International Conference on Data Engineering (ICDE'06)*, 24–24. IEEE.

Motwani, R. et Xu, Y. (2007). Efficient algorithms for masking and finding quasi-identifiers. Dans *Proceedings of the Conference on Very Large Data Bases (VLDB)*, 83–93.

Neubauer, T. et Heurix, J. (2011). A methodology for the pseudonymization of medical data. *International journal of medical informatics*, *80*(3), 190–204.

Okuno, T., Ichino, M., Kuboyama, T. et Yoshiura, H. (2011). Content-based de-anonymisation of tweets. Dans *2011 Seventh International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 53–56. IEEE.

Patil, D., Mohapatra, R. K. et Babu, K. S. (2017). Evaluation of generalization based k-anonymization algorithms. Dans *2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS)*, 171–175. IEEE.

Peddapunnaiahp, G. et Kiran, Y. (2016). Anonymizing tree structure with privacy preserving data. *International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)*.

Prasser, F., Kohlmayer, F. et Kuhn, K. A. (2014). A benchmark of globally-optimal anonymization methods for biomedical data. Dans *2014 IEEE 27th International Symposium on Computer-Based Medical Systems*, 66–71. IEEE.

Rebollo-Monedero, D., Forne, J. et Domingo-Ferrer, J. (2009). From t-closeness-like privacy to postrandomization via information theory. *IEEE Transactions*

on Knowledge and Data Engineering, *22*(11), 1623–1636.

Rosati, P., Gogolin, F. et Lynn, T. (2020). Cyber-security incidents and audit quality. *European Accounting Review*, 1–28.

Samarati, P. (2001). Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering*, *13*(6), 1010–1027.

Samarati, P. et Sweeney, L. (1998). *Protecting privacy when disclosing information : k-anonymity and its enforcement through generalization and suppression*. Rapport technique, technical report, SRI International.

Singh, A. P. et Parihar, M. D. (2013). A review of privacy preserving data publishing technique. *International Journal of Emerging Research in Management &Technology*, *2*(6), 32–38.

Solé, M., Muntés-Mulero, V. et Nin, J. (2012). Efficient microaggregation techniques for large numerical data volumes. *International Journal of Information Security*, *11*(4), 253–267.

Soria-Comas, J. et Domingo-Ferrer, J. (2016). Big data privacy : challenges to privacy principles and models. *Data Science and Engineering*, *1*(1), 21–28.

Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D. et Martínez, S. (2014). Enhancing data utility in differential privacy via microaggregation-based k-anonymity. *The VLDB Journal*, *23*(5), 771–794.

Sweeney, L. (2002). k-anonymity : A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557–570.

Tinabo, R., Mtenzi, F. et O'Shea, B. (2009). Anonymisation vs. pseudonymisation : Which one is most useful for both privacy protection and usefulness of e-healthcare data. Dans *2009 International Conference for Internet Technology and Secured Transactions,(ICITST)*, 1–6. IEEE.

Truta, T. M., Campan, A. et Meyer, P. (2007). Generating microdata with p-sensitive k-anonymity property. Dans *Workshop on Secure Data Management*, 124–141. Springer.

Véliz, C. et Grunewald, P. (2018). Protecting data privacy is key to a smart energy future. *Nature Energy*, *3*(9), 702–704.

Wang, J., Luo, Y., Zhao, Y. et Le, J. (2009). A survey on privacy preserving data mining. Dans *2009 First International Workshop on Database Technology and Applications*, 111–114. IEEE.

Xiao, X. et Tao, Y. (2006). Anatomy : Simple and effective privacy preservation. Dans *VLDB*, volume 6, 139–150.

Zhang, L., Wang, X., Lu, J., Li, P. et Cai, Z. (2016). An efficient privacy preserving data aggregation approach for mobile sensing. *Security and Communication Networks*, *9*(16), 3844–3853.