Université de Montréal

Un modèle rétroactif de réconciliation utilité-confidentialité sur les données d'assurance.

par
Jonathan Rioux

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en informatique

Avril 2016

RÉSUMÉ

Le partage des données de façon confidentielle préoccupe un bon nombre d'acteurs, peu importe le domaine. La recherche évolue rapidement, mais le manque de solutions adaptées à la réalité d'une entreprise freine l'adoption de bonnes pratiques d'affaires quant à la protection des renseignements sensibles.

Nous proposons dans ce mémoire une solution modulaire, évolutive et complète nommée *PEPS*, paramétrée pour une utilisation dans le domaine de l'assurance. Nous évaluons le cycle entier d'un partage confidentiel, de la gestion des données à la divulgation, en passant par la gestion des forces externes et l'anonymisation. *PEPS* se démarque du fait qu'il utilise la contextualisation du problème rencontré et l'information propre au domaine afin de s'ajuster et de maximiser l'utilisation de l'ensemble anonymisé. À cette fin, nous présentons un algorithme d'anonymat fortement contextualisé ainsi que des mesures de performances ajustées aux analyses d'expérience.

Mots clés: Partage confidentiel de données, Gestion de la confidentialité, Données d'assurance, Mesure de l'utilité d'un ensemble de données anonymisé.

ABSTRACT

Privacy-preserving data sharing is a challenge for almost any enterprise nowadays, no matter their field of expertise. Research is evolving at a rapid pace, but there is still a lack of adapted and adaptable solutions for best business practices regarding the management and sharing of privacy-aware datasets.

To this problem, we offer *PEPS*, a modular, upgradeable and end-to-end system tailored for the need of insurance companies and researchers. We take into account the entire cycle of sharing data : from data management to publication, while negotiating with external forces and policies. Our system distinguishes itself by taking advantage of the domain-specific and problem-specific knowledge to tailor itself to the situation and increase the utility of the resulting dataset. To this end, we also present a strongly-contextualised privacy algorithm and adapted utility measures to evaluate the performance of a successful disclosure of experience analysis.

Keywords: Privacy-preserving data sharing, Confidentiality management, Insurance data, Utility measures for anonymized datasets.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

LISTE DES SIGLES

ED	Ensemble de données
IMC	Indice de masse corporelle
QID	Quasi-identifiant ou Quasi-identificateur

À ceux qui croient encore
au droit à la vie privée

REMERCIEMENTS

La rédaction d'un mémoire est une expérience unique et je suis très content d'avoir pu la partager avec plusieurs amis et collègues. Je m'excuse dès le départ si j'oublie quelques personnes. Soyez sûrs que je vous suis extrêmement reconnaissant.

En premier lieu, je souhaite remercier ma famille proche et moins proche, plus spécialement Olivier, qui s'est dévoué corps et âme pour que je puisse me concentrer durant les mois de rédaction. On passe trop souvent sous silence le rôle que le rire et le café chaud peuvent jouer, spécialement après plusieurs heures devant un écran d'ordinateur. Merci encore !

Un autre gros merci à mon frère Simon, qui comprend trop bien les aléas des études supérieures. Malgré la distance, ce fut un plaisir d'échanger, de se relire mutuellement et de s'engueuler. Je suis convaincu que mon expérience aurait été fort différente si tu n'avais pas été là.

Je remercie également mes collègues chez RGA Canada et plus spécifiquement mon (désormais ex) gestionnaire Marc-André Belzil. J'ai eu un plaisir inouï à jouer dans les données de notre équipe, à tester des stratégies, à remettre en question nos processus et à échanger sur tout et sur rien. Merci de tes encouragements et de la flexibilité dont tu as fait preuve pendant mes études. Je suis convaincu que nous collaborerons de nouveau dans le futur !

Peu d'étudiants possèdent la chance d'avoir une équipe de directeurs comme celle dont j'ai pu profiter. Esma, Gilles, vous faites une équipe inusitée et votre expérience m'a fait découvrir la recherche comme je n'aurais pu imaginer. Merci de la confiance que vous m'avez accordée et des nombreuses réunions de fins de journées qui ne se terminent que lorsque nous réalisons l'heure.

Finalement, mes plus sincères remerciements à mes collègues du laboratoire, et plus spécialement à Mouna Selmi. Ce fut un plaisir d'échanger avec vous toutes et tous et je vous souhaite un succès à la hauteur de vos ambitions.

Jonathan Rioux

CHAPITRE 1

INTRODUCTION

N'importe quel internaute sait à quel point ses renseignement personnels sont convoités par de nombreux individus et entreprises. Cela n'est pas nouveau : la connaissance acquise par des tests de sélection, des scores divers et des bilans médicaux alimente les entreprises depuis que celle-ci est disponible au grand public. À l'ère de la *ruée vers l'or numérique*, plusieurs opportunités sont présentés au propriétaires d'ensembles de données de partager, collaborer, voire vendre celles-ci.

Quand il ne s'agit pas de vos renseignements, la perception de la valeur des données est différente [2]. Il suffit de naviguer un peu pour voir des partages de données plus ou moins facilement réidentifiables [51]. Bien que nous sachions que la confidentialité parfaite est un rêve pour le moment impossible à atteindre [18], nous pouvons toutefois mesurer le risque de divulgation en utilisant des méthodes développées durant les dernières années [25]. Comment toutefois réconcilier le risque de réidentification, l'utilité de l'ensemble anonymisé et surtout de s'assurer que nous puissions faire évoluer notre pratique à travers le temps ?

Voici *PEPS*, un outil gérant le cycle entier d'une divulgation anonymisée couronnée de succès. *PEPS* accompagne l'utilisateur avec un processus facile à comprendre et à implémenter et permet de divulguer des ensembles anonymisés selon les conditions prédéterminées. Le système est également évolutif, conservant les succès passés pour inspirer ceux futurs. Celui-ci vient accompagné d'un algorithme de confidentialité fortement contextualisé permettant de tirer avantage de l'information que nous possédons. Le résultat est un ensemble de données anonymisé et bâti sur-mesure pour donner des résultats satisfaisants. Nous utiliserons les données d'assurance comme fondation, mais nous pouvons le migrer vers d'autres domaines avec aisance.

1.1 Objectifs

La plupart des protocoles ou systèmes de partages confidentiels sont souvent tributaites de l'une ou de plusieurs des caractéristiques suivantes :

- pour mesurer la performance ou l'utilité de l'ensemble de données résultant, les types d'analyse assumés sont souvent limités à des primitives d'apprentissage machine ou des opérations statistiques de base ;
- les attaques réussies contre un ensemble de données anonymisé n'ont aucun impact sur les divulgations futures ;
- dans le cas où nous utilisons un algorithme de confidentialité utilisant une procédure de généralisation (où une donnée spécifique est remplacée par une plus générale), les heuristiques employées pour rediscrétiser les données — lorsque nécessaire — sont souvent primitives, nuisant à l'utilité résiduelle de l'ensemble anonymisé.

Ces caractéristiques nuisent à l'implantation d'une solution intégrée de divulgation confidentielle des données et limitent l'impact des solutions existantes pour plusieurs acteurs. En prenant comme cas spécifique l'industrie de l'assurance, nous tenterons de combler les lacunes mentionnées ci-dessus. Les blocs constitutifs de notre système seront toutefois conçus de sorte à être utilisables avec peu de modifications pour un autre type d'industrie.

Plus spécifiquement, en reprenant les caractérisques mentionnées-ci haut, nous

- utiliserons une caractérisation plus poussée des analyses les plus susceptibles d'être utilisées pour préciser l'utilité des données post-anonymisation ;
- proposerons un système basique de préservation des divulgations passées afin de pouvoir mesurer l'impact d'une attaque et de réagir convenablement ;
- explorerons différentes méthodes de sérialisation afin d'obtenir une meilleure utilité pour une technique d'anonymisation donnée.

Ces objectifs s'inscrivent dans la création d'un système gérant l'ensemble des opérations entourant une divulgation confidentielle réussie, de la sélection des données jusqu'à l'audit de la divulgation, en tenant compte des facteurs externes connus. Ces étapes s'inscrivent dans une structure de gestion du risque en entreprise [60] appliquée aux actifs numériques.

1.2 Scénario

Notre analyse se concentrera plus spécifiquement sur les données d'étude pour les compagnies d'assurances de personnes[1]. Pour un acteur hors du milieu, le flux des données ainsi que les partages-types peuvent apparaître nébuleux. À cette fin, nous définirons un scénario type sur lequel les prochaines sections s'appuieront. Plusieurs termes utilisés ici sont définis dans la Section 2.1.

Soit une compagnie d'assurance de personnes que nous appelerons *Moon Life*[2], propriétaire d'une base de données $BD(A)$. Nous supposerons par souci de simplicité que $BD(A)$ encapsule *l'ensemble des données collectées et disponibles pour une analyse spécifique*[3]. Moon Life souhaite partager avec un individu B (pouvant être une université, une autre compagnie, un ordre professionnel[4], un consultant ou le public en général) un ensemble de données $ED(A)$ permettant de faire un travail d'analyse. Moon Life souhaite que l'ensemble de données soit le plus précis possible sans toutefois courir un trop grand risque de bris de confidentialité.

Préalablement, Moon Life a fait un examen de son appétit face au risque de confidentialité et évalué sa tolérance face à la perte d'information découlant de l'anonymisation des microdonnées. Suite à ces exercices, Moon Life a pu bâtir des bornes de confiden-

[1] Par opposition aux compagnies d'assurance IARD (Incendie, Accident et Risques Divers. Nous verrons toutefois que plusieurs liens peuvent être tissés entre les deux.)

[2] À ne pas confondre avec la version diurne de cette entreprise...

[3] La distinction est importante puisque, par exemple, les lignes d'assurance vie et de dommages (ou les services bancaires) ne peuvent partager des renseignements sur leurs assurés. Advenant qu'une compagnie oeuvre dans plus d'un des secteurs mentionnés, $BD(A)$ sera alors mutable et représentera les données disponibles pour une analyse donnée.

[4] Prenons exemple de l'Institut Canadien des Actuaires au Canada qui collecte des extraits d'assurés sur une base régulière afin de produire des études d'industrie.

tialité et d'utilité minimales. Ces bornes feront partie des *politiques* qui gouverneront les futurs partages d'ensembles de données.

Suite à une demande pour un ensemble $ED(A)$, Moon Life tentera de trouver un compromis utilité-confidentialité adapté à la situation. Un partage réussi créera un protocole permettant d'avoir une base de paramètres pour des échanges similaires futurs.

1.3 Un exemple de partage confidentiellement insuffisant

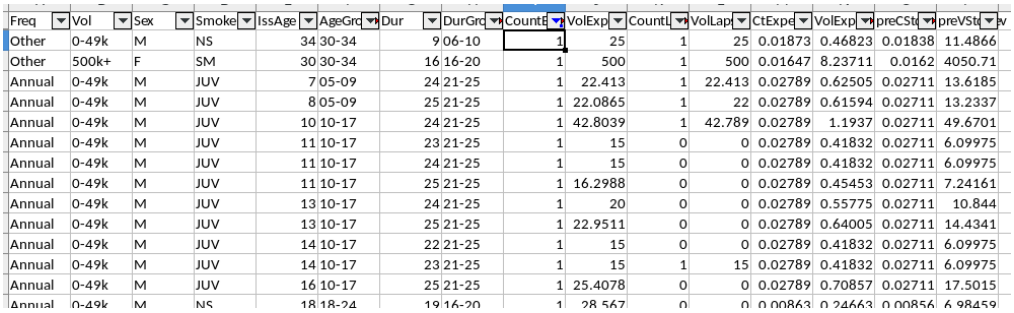
La motivation derrière ce mémoire est de fournir aux acteurs du milieu des moyens suffisants pour partager confidentiellement des données potentiellement sensibles, sans toutefois réduire l'utilité au point que le partage soit inutile.

En reprenant le modèle de scénario de la Section 1.2, considérons le cas suivant : le 23 septembre 2015, l'Institut Canadien des Actuaires publiait un papier de recherche nommé *Lapse Experience under Universal Life Level Cost of Insurance Policies* [9]. L'acquisition des données s'est fait de façon standard : les compagnies furent invitées à partager leurs données sérialisées via un portail que nous considérerons sécuritaire. Les données des compagnies participantes furent mises en commun et ont permis de faire une étude servant à l'ensemble de la communauté actuarielle canadienne. Le rapport est disponible à quiconque sur la page de l'Institut, ainsi qu'un ensemble de données dépersonnalisé.

En faisant une manipulation de tri (sur l'exposition de chaque groupe) et un filtre (afin d'isoler les groupes ayant une exposition de 1[5]), nous pouvons voir que 6598 tuples tombent dans cette catégorie. En considérant que nous possèdons l'ensemble de l'information permettant de recréer les 8 champs identifiant uniquement un tuple, nous pouvons *identifier uniquement ces individus* et inférer que ces polices n'ont pas terminé durant cette année.

Nous avons sélectionné un exemple où l'information inférée n'est pas particulièrement dangereuse pour les assurés présents dans cet ensemble de données. Toutefois,

[5] Étant donné que l'ensemble de données est précis à plusieurs décimales sur ce champ, nous supposons qu'un tuple ayant une exposition précisément de 1 constitue un seul individu exposé pendant une période d'un an.



Freq	Vol	Sex	Smoke	IssAge	AgeGrc	Dur	DurGrc	CountE	VolExp	CountL	VolLap	CtExpe	VolExp	preCSt	preVSt	v
Other	0-49k	M	NS	34	30-34	9	06-10	1	25	1	25	0.01873	0.46823	0.01838	11.4866	
Other	500k+	F	SM	30	30-34	16	16-20	1	500	1	500	0.01647	8.23711	0.0162	4050.71	
Annual	0-49k	M	JUV	7	05-09	24	21-25	1	22.413	1	22.413	0.02789	0.62505	0.02711	13.6185	
Annual	0-49k	M	JUV	8	05-09	25	21-25	1	22.0865	1	22	0.02789	0.61594	0.02711	13.2337	
Annual	0-49k	M	JUV	10	10-17	24	21-25	1	42.8039	1	42.789	0.02789	1.1937	0.02711	49.6701	
Annual	0-49k	M	JUV	11	10-17	23	21-25	1	15	0	0	0.02789	0.41832	0.02711	6.09975	
Annual	0-49k	M	JUV	11	10-17	24	21-25	1	15	0	0	0.02789	0.41832	0.02711	6.09975	
Annual	0-49k	M	JUV	11	10-17	25	21-25	1	16.2988	0	0	0.02789	0.45453	0.02711	7.24161	
Annual	0-49k	M	JUV	13	10-17	24	21-25	1	20	0	0	0.02789	0.55775	0.02711	10.844	
Annual	0-49k	M	JUV	13	10-17	25	21-25	1	22.9511	0	0	0.02789	0.64005	0.02711	14.4341	
Annual	0-49k	M	JUV	14	10-17	22	21-25	1	15	0	0	0.02789	0.41832	0.02711	6.09975	
Annual	0-49k	M	JUV	14	10-17	23	21-25	1	15	1	15	0.02789	0.41832	0.02711	6.09975	
Annual	0-49k	M	JUV	16	10-17	25	21-25	1	25.4078	0	0	0.02789	0.70857	0.02711	17.5015	
Annual	0-49k	M	NS	18	18-24	19	16-20	1	28.567	0	0	0.00863	0.24663	0.00856	6.98459	

Figure 1.1 – Données publiées, dépersonnalisation et 1-anoynymat

notre exemple illustre l'importance de considérer la confidentialité dans notre processus de décision de partage et de faire les ajustements nécessaires au besoin.

1.4 Contribution

Notre contribution principale est *PEPS*, défini en Section 3.2. PEPS est un sytème permettant de gérer le partage d'un ensemble de données anonymisé selon des critères d'intimite et d'utilité bien définis et modulables. Il généralise les systèmes existants en fournissant des fondations solides sur les modules externes à l'opération d'anonymisation, notamment la gestion des données et la réaction aux forces externes.

À travers le développement de *PEPS*, nous avons également défini le concept de *Protocole*, qui permet de s'assurer d'avoir des opérations de partage reproductibles et d'auditer ces opérations à postériori. Nous avons également développé un algorithme de *LKC*-anonymat pour une contextualisation plus poussée que ce qui était disponible auparavant : *CLiKC*. À travers le développement de *CLiKC*, nous avons également proposé plusieurs hiérarchies contextualisées : celles-ci pourront servir de pierres angulaires pour d'autres opérations d'anonymat, en assurance ou dans un autre domaine.

Bien que cet outil est externe à notre but premier, la nécéssité d'avoir des données suffisament détaillées et publiquement divulgables nous à conduit à développer un outil basique de simulation d'assurés. Cet outil, nommé *Synthure*, nous a libéré de la contrainte d'utiliser des données propriétaires, contrainte incontournable dans notre cas.

1.5 Organisation du mémoire

Le mémoire débute avec une présentation de l'état de l'art au Chapitre 2. Nous présentons d'abord un historique des deux principales familles d'anonymat, syntaxique et sémantique, et expliquons leur utilité à notre système. Suivant cela est une présentation de mesures d'utilité ayant servi dans le passé et aujourd'hui pour juger de la pertinence d'une opération d'anonymat pour un ou certains types d'analyse. Nous joignons ensuite ces deux réalités pour une courte présentation sur le compromis confidentialité-utilité, avant de présenter les systèmes existants.

Dans le Chapitre 3, nous présentons nos hypothèses ainsi que le scénario dans lequel nous conduirons nos expériences. *PEPS* est ensuite introduit et nous passons en revue les modules constitutifs ainsi que les intrants et extrants de chacun de ceux-ci. Suite à cela, nous décrivons l'ensemble de données simulées avec l'outil *Synthure* ainsi que les hypothèses gouvernant la distribution des tuples. L'algorithme *CLiKC* est ensuite illustré et accompagné de sa preuve de *LKC*-anonymat et de ses garanties. Nous terminons avec les hiérarchies suggérées pour l'ensemble susmentionné.

Le chapitre 4 explique en détail le processus suivi par *PEPS* en utilisant les données simulées en entrée. Nous discutons des choix de paramètres et des contraintes rencontrées, et proposons des outils pour reconstruire un ensemble de données discrets. Nous poursuivons le processus jusqu'à la sortie de l'ensemble anonymisé avant de reprendre un exemple sur des données publiques. Nous discutons des résultats et des limites de notre système en Chapitre 5 avant de conclure le mémoire.

Une présentation de l'outil *Synthure*, ainsi que la description des principaux fichiers disponibles sur le dépôt de Code accompagnant le mémoire sont en Annexe.

CHAPITRE 2

ÉTAT DE L'ART

2.1 Définitions

La littérature sur la protection de la vie privée comporte un vocabulaire bien à elle et est majoritairement anglophone. Afin d'éviter les ambiguités, une définition des principaux termes utilisés est présentée. Un bon nombre des définitions théoriques est une adaptation de celles présentes en [29]. Par souci de cohérence, nous avons repris autant que possible les termes des systèmes existants lorsque ceux-ci représentaient des concepts similaires au nôtres. Pour une présentation de quelques systèmes existants, voir la Section 2.6.

L'**anonymat** des informations sur un individu présent dans la base de données est la propriété d'être indistinguable d'autres individus selon un certain aspect. Formellement, pour un ensemble de requêtes $A, (A_i, i \in [1,n])$ et deux individus P_j, P_k, nous considérons que les deux individus sont anonymisés sous A si, pour toutes les requêtes en A, un attaquant ne peut différencier entre P_j et P_k après l'ensemble des requêtes posées. Dans le cas où plusieurs individus sont indistinguables, on appelle l'ensemble de ceux ci l'**ensemble d'anonymat**.

Les **microdonnées** sont les données avant traitement statistique. Nous considérerons ici les données sérialisées et les microdonnées comme deux concepts identiques et interchangeables. Nous utilisons le terme *sérialisé* lorsque chaque élément possède son propre tuple (voir la définition ci-dessous), par opposition aux données agglomérées, où elles sont en classes ou groupes.

Les microdonnées sont le plus souvent organisées en *tables*, ensemble ordonné ou non de *tuples* (ou *lignes*) comportant des valeurs d'un ensemble d'*attributs* ou *paramètres*. Les attributs peuvent être regroupés sous quatre catégories :

- un **identificateur** est un attribut qui à lui seul identifie de façon unique un individu ;
- un **quasi-identificateur** est un attribut ou un ensemble d'attributs qui, une fois liés à de l'information externe, permet de significativement réduire le domaine des individus possibles ou de réidentifier certains individus de la table ;
- un **attribut confidentiel** est un attribut sensible que nous ne souhaitons pas dévoiler ;
- un **attribut non-confidentiel** est un attribut dont la divulgation peut se faire sans danger. Nous ne considérerons pas ces attributs.

Une **divulgation**[1] consiste dans notre cas en l'identification réussie d'un individu par un adversaire. Formellement, en reprenant la définition mathématique de l'anonymat, une divulgation se produirait si l'adversaire était en mesure de distinguer P_j de P_k.

Une **inférence** est similaire à une divulgation : celle-ci se produit lorsque nous pouvons identifier un individu ou un groupe d'individus avec une probabilité assez forte.

Une **classe d'équivalence** [45] est un ensemble de tuples t possédant un ou plusieurs attributs identiques, regroupés sur ces attributs identiques. Nous pouvons faire le parallèle avec la directive `group by` du language relationel SQL [35]. Par exemple, si nous avons des fumeurs et des non-fumeurs dans notre ensemble de données, grouper selon ce paramètre va nous donner 2 classes d'équivalence, une contenant les fumeurs et une autre les non-fumeurs.

[1] *disclosure* en anglais

2.2 La gestion du risque en entreprise et les actifs numériques

Les médias sociaux et la culture des *start-ups* ont créé un engouement pour la collecte et l'analyse des données : le terme *data science* fait désormais partie du vocabulaire et cette discipline fait partie des professions les plus prometteuses [23]. Ce phénomène n'est toutefois pas nouveau et s'étend à ce que nous appelerons les consommateurs de données d'ancienne génération. Nous pouvons penser aux banques, compagnies d'assurances, agences de crédit (Equifax et TransUnion, principalement) dont la collecte et l'utilisation des données est un élément-clé de leur fonctionnement. La rapide montée en popularité de cet *or numérique* ont crée une triple pression sur ces acteurs que nous illustrons sous forme de *besoins* :

- Le besoin de *sophistication* : les techniques et processus ont gagné en complexité et en pouvoir prédictif. Ces acteurs doivent désormais compétitionner contre des joueurs non-traditionnels. Prenons par exemple l'arrivée en Janvier 2016 de Besurance [5], qui tente de faire renaître l'assurance communautaire en misant sur des outils technologiques. Les gros joueurs sont conscients de la pression et multiplient les créations d'incubateurs et laboratoires d'innovation [11] [55] ;

- parallèlement, l'accès à l'information est facilité par les aggrégateurs de prix (pensons Kanetix en assurance de dommages [37] à Term4Sale pour les produits d'assurance vie temporaires [64]) et un accès au marché de plus en plus transparent. Les acteurs traditionnels sont présentement coincés avec une course au plus bas prix ;

- Le besoin de *confidentialité* : la survie d'une entreprise financière réside notamment sur la confiance des clients. Les brêches d'information, de plus en plus fréquentes, doivent être prises au sérieux puisqu'elle peuvent nuire à leur réputation. Dans un domaine où la compétition est féroce et où les produits sont relativement homogènes, une compagnie considérée moins sécure aura de la difficulté à conserver ses parts de marché.

Considérer les données acquises et la capacité d'analyse comme un actif facilite l'intégration des éléments numériques au sein d'une stratégie d'entreprise. Puisque celles-ci contribuent à la création de valeur, nous pouvons les considérer comme un actif [60]. Nous pouvons tracer quelques parallèles :

- les données peuvent avoir une valeur différente selon le *propriétaire* : une compagnie d'assurance accordera une valeur plus grande à des renseignements médicaux sur ses assurés qu'une entreprise de taxi-covoiturage (e.g. Uber) ;
- nous devons imputer un coût d'*usure* aux actifs numériques : tout comme un grand nombre d'actifs tangibles, l'âge des données est souvent négativement corrélé avec les valeurs ;
- nous devons finalement imputer un *coût d'entreposage et d'utilisation* : outre le matériel et l'énergie requises pour le stockage, il faut également considérer le temps et les ressources nécessaires au traitement.

Un traitement financier rigoureux des microdonnées dépasse le cadre de ce mémoire mais est un sujet fréquemment traité [72] [73]. Nous ferons l'hypothèse que les données partagées font partie intégrante de la stratégie de l'entreprise concernée et que leur valeur intrinsèque est convenablement calculée à tout moment. Cette hypothèse nous permet de nous libérer des contraintes économiques

2.3 La protection de la confidentialité

2.3.1 De l'inefficacité de la dépersonnalisation naïve

La dépersonnalisation est probablement la façon la plus intuitive, bien que fort limitée [21], de procéder à l'anonymisation d'une base de données, ce qui explique sa popularité.

Le procédé est assez simple : il suffit de déterminer les identificateurs parmi les différents attributs de la table, puis de les supprimer. Dans le cas d'informations individuelles,

nous pouvons penser au numéro d'assurance sociale, de permis de conduire, au code permanent ou au prénom/nom de famille. Dès le départ, nous pouvons voir que l'absence de traitement des quasi-identificateurs (QID) est une faille fatale [50]. Une table comportant de nombreux champs et combinaisons de champs permettant d'identifier de façon aisée un individu (comme les tables utilisés pour la recherche médicale ou en assurance de personnes) s'exposent particulièrement à ce genre d'attaque.

Plus spécifiquement, posons un adversaire Adv souhaitant identifier le tuple où les valeurs sensibles d'un assuré V se retrouvent dans la table T de $BD(A)$ (nous délaisserons le (A) pour cet exemple et ceux subséquents). L'adversaire a accès à l'ensemble des champs non-sensibles de T et posséde la connaissance que nous dénoterons qid. qid comporte au minimum le vecteur QID_v des quasi-identificateurs de l'assuré V mais peut contenir plus d'éléments. Armé de cette information, l'adversaire peut alors identifier un groupe de tuples dénommé $T[QID_v]$ qui comportent les mêmes quasi-identificateurs que QID_v. Soit $|T[QID_v]|$ le nombre de tuples dans $T[QID_v]$, l'adversaire peut alors attaquer de deux façons différentes, selon le résultat obtenu :

si $|T[QID_v]| = 1$, alors Adv utilise ce que nous appelerons une *attaque par lien identitaire*[2]. Nous voyons alors que $|TQID_v|$ retourne alors un unique tuple, révélant ainsi les attributs sensibles de v.
Si toutefois $|T[QID_v]| > 1$, alors Adv utilise ce que nous appelerons une *attaque par lien attributif*[3]. À partir des tuples obtenus, Adv peut tenter d'inférer une partie des attributs sensibles de v. Ce type d'attaque a donné naissance à une mesure d'anonymat appelée la *probabilité globale de réidentification* [21]. Pour v, la probabilité de réidentification sur l'ensemble de données précédemment mentionné est $1/|T[QID_v]|$. Nous pouvons faire cet exercice pour l'ensemble de ceux-ci présents dans la table. La probabilité globale de

[2]de l'anglais *identity linkage* [50]
[3]de l'anglais *attribute linkage*

réidentification sera alors de

$$\max_{v} \frac{1}{|T[QID_j]|} \tag{2.1}$$

Sans traitement, il est facile de voir que cette technique ne peut donner de résultats prometteurs à moins d'avoir un ensemble de données colossal et un nombre de quasi-identificateurs très limité. Le site internet *How Unique are You ?* [63] sert notamment à sensibiliser les utilisateurs sur la facilité de la réidentification basée sur trois éléments facilement accessibles : la date de naissance, le sexe et le *ZIP code*[4].

Nous concluons cette section par un exemple des attaques sur un ensemble de données fictif. Prenons une base de données d'assurabilité minimale[5]. Notons que les champs `NAS` (Numéro d'assurance sociale) et `nom` sont supprimés puisque nous ne pouvons pas les généraliser efficacement.

Tableau 2.I – Ensemble de données type

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

Supposons que le paramètre *Cause de sélection* est celui que nous souhaitons inférer. Un adversaire possédant le vecteur dans le Tableau 2.II peut alors identifier sans équivoque le tuple avec l'astérisque et constitue une attaque par lien identitaire réussie.

Dans ce cas, le couple *(Date de naissance, Sexe)* fut suffisant pour identifier sans équivoque le tuple souhaité. Il est important de noter que ce n'est pas la seule attaque

[4] Analogue américain du code postal, sans toutefois être aussi précis.

[5] Cet exemple est inspiré de celui présenté en [58].

Tableau 2.II – Identification d'un tuple via deux champs

Date de naissance	Sexe	Lieu de résidence	Statut marital	Cause de sélection
1964-03-25	M			

possible. Dans notre cas, le couple *(Date de naissance, Lieu de résidence)* est également suffisant, ainsi que le couple *(Lieu de résidence, Statut marital)*.

Une attaque par lien attributif est possible si l'adversaire possède le vecteur suivant :

Tableau 2.III – Vecteur suffisant pour une attaque par lien attributif

Date de naissance	Sexe	Lieu de résidence	Statut marital	Cause de sélection
	M	Montréal		

Dans notre cas, il est facile de voir que $|T[qid]| = 2$, mais les deux tuples possèdent la même valeur pour l'attribut *Cause de sélection*. La probabilité d'inférer le champ sensible étant donné *qid*, $P(\texttt{Hypertension}|qid) = 1(100\%)$.

Les exemples ci-dessus sont très simplistes mais illustrent bien le genre d'attaque auquel une entité peut faire face lors d'un partage de données. Les mécanismes de déidentifications naïfs sont clairement insuffisants et il nous faut des techniques plus sophistiquées. Avant de se lancer, il convient cependant de mettre en place des mesures d'anonymat plus exhaustives et robustes. Passons maintenant en revue les deux familles dominantes, soit le *k*-anonymat et la confidentialité différentielle.

2.3.2 Le k-anonymat

Le concept de *k*-anonymat émerge d'une généralisation de la dépersonnalisation par quasi-identificateurs. Puisque la majorité des ensembles de données n'ont pas la combinaison grande-taille/peu de quasi-identificateurs requise, une méthode plus robuste fut développée.

Le *k*-anonymat fait partie de ce que nous appelons les méthodes *syntaxiques* (par oppositions aux méthodes *sémantiques*, rencontrées en Section 2.3.6). Ces méthodes reposent sur une hiérarchisation des données, du plus général au plus spécifique, cou-

ramment appelée **taxonomie**[6]. Par exemple, une hiérarchisation très simple peut être représentée par la figure 2.1 : nous voyons que le Québec (élément général) peut être vu comme un ensemble de villes (Montréal, Québec, Rimouski), qui sont constituées de quartiers ou arrondissements. À la différence de cet exemple, les hiérarchies dans les méthodes d'anonymisation doivent être *totales*, c'est à dire que toutes les valeurs individuelles doivent s'y retrouver.

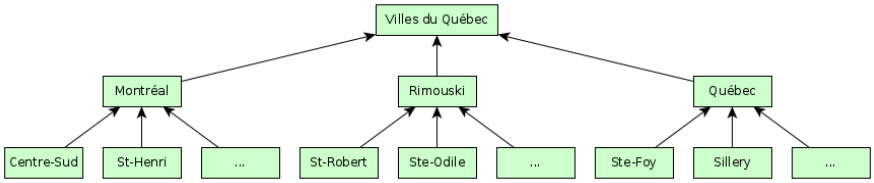


Figure 2.1 – Exemple de taxonomie : certaines villes et quartiers du Québec.

Plus spécifiquement, cette méthodologie [58] repose sur deux concepts permettant de fournir des garanties prouvables sur la confidentialité de l'ensemble de données résultant :

- la *généralisation* des paramètres, soit de remplacer une valeur spécifique par une valeur générale sans la falsifier ;
- la *suppression* pure et dure de ceux ne pouvant pas survivre à une généralisation satisfaisante.

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

Définition 2.3.1. k-anonymat : Posons un ensemble de donnés $T(A_1,\ldots,A_n)$ et les quasi-identificateurs associés QI_T. L'ensemble T est réputé k-anonyme si pour chaque quasi-identificateur $QI \in QI_T$ chaque séquence de tuples apparaît au moins k fois dans $T[QI]$.

[6]Le terme est emprunté aux sciences de la vie, plus spécialement l'étude de l'évolution.

2.3.3 Le k-anonymat : un exemple d'application

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

En reprenant notre exemple précédent en Section 2.3.1 et en supposant que toutes les combinaisons des champs `Date de naissance`, `Sexe`, `Lieu de résidence` et `Statut marital` soient à notre disposition, posons ici $k = 2$. Nous pourrons alors obtenir l'ensemble de données 2-anonymisé suivant.

Tableau 2.IV – Ensemble de données-type 2-anonymisé

Date de naissance	Sexe	Lieu de résidence	Statut marital	Cause de sélection
$\leq$ 1970-12-31	F	*	*	IMC hors normes
$\leq$ 1970-12-31	M	*	*	Hypertension
$\leq$ 1970-12-31	F	*	*	IMC hors normes
$\leq$ 1970-12-31	M	*	*	Hypertension
$\leq$ 1970-12-31	F	*	*	Autre
$\leq$ 1970-12-31	F	*	*	Autre
$\leq$ 1970-12-31	F	*	*	IMC hors normes
$\leq$ 1970-12-31	M	*	*	Autre
$\leq$ 1970-12-31	M	*	*	Autre

Les champs dénotés par un astérisque représentent une suppression du champ.

Étant donné le faible nombre de tuples et la diversité des champs, les résultats de l'exemple ne sont pas très intéressants. Le but était ici d'illustrer succintement le processus de k-anonymisation. Avec un nombre plus élevé d'observations, nous pouvons augmenter le nombre de paramètres et leur granularite et possiblement augmenter le paramètre k.

2.3.4 Faiblesses et critiques du k-anonymat

Le k-anonymat est la formalisation d'un processus utilisé par certains systèmes de contrôle sur divulgation existants, notamment Datafly et Mu-Argus [53]. Le travail accompli a permis de corriger plusieurs lacunes dans ces systèmes et a orienté leurs efforts

futurs.

Afin que notre divulgation anonymisée soit couronné de succès, le *k*-anonymat suppose que nous connaissons l'ensemble des combinaisons possibles, présentes ou *futures*. Cette restriction nous laisse alors devant deux choix :

- considérer toutes les combinaisons de $1, 2, \ldots, n$ quasi-identificateurs dans notre ensemble de données ;
- faire l'hypothèse que les sources externes connues du propriétaire des données sont les seules qui seront disponibles pendant la période utile de la divulgation.

Dans certains cas, notamment si nous connaissons parfaitement la capacite d'inférence du récipiendaire, le second scénario peut être envisageable. Dans le cas d'une divulgation publique, le principe de précaution nous oblige à considérer l'ensemble des combinaisons, ce qui peut réduire dramatiquement la granularité de notre ensemble de données anonymisé et potentiellement son utilité.

Le choix du paramètre *k* est également d'une importance capitale. Posons *k* trop élevé et nous aurons un ensemble de données avec une suppression trop agressive, rendant caduc notre but premier. Un *k* trop faible va allouer une probabilité globale de réidentification trop élevée. Nous pouvons voir en section 1.3 un partage où le paramètre *k*, s'il fut utilisé, était clairement insuffisant.

Le modèle suppose également que chaque tuple représente un individu distinct. Les auteurs en [25] démontrent que si plusieurs tuples représentent un même individu — un scénario facilement envisageable pour une compagnie d'assurance, où un individu peut prendre plusieurs couvertures différentes à plusieurs moments — un groupe de *k* tuples peut représenter moins de *k* individus, diminuant nos garanties de confidentialité. L'application du *k*-anonymat sur des données transactionnelles a alors un interêt limité.

Finalement, le *k*-anonymat est sensible à ce qui a été appelé *la malédiction de la dimensionalité* [3][7]. L'application du *k*-anonymat sur un ensemble de données possédant un grand nombre de dimensions mène à une dégradation notable de la qualité de celles-ci, donc de son utilité résiduelle. Cette limitation nous pose spécialement problème étant

[7] traduction de *the curse of high-dimensionality [on k-anonymity].*

donné le nombre arbitrairement élevé de QID que nos ensembles de données peuvent posséder : il suffit d'imaginer les résultats d'un bilan médical servant à la sélection d'un assuré pour comprendre la problématique.

2.3.5 Le LKC-anonymat

Le concept de k-anonymat a, malgré ses faiblesses, pavé le terrain pour plusieurs extensions et améliorations. Celles-ci ont été surtout proposées afin d'esquiver le problème de dimensionalité ou d'aider à l'utilité éventuelle de l'ensemble de données. Parmi celles-ci, notons la l-diversité [47], l'(α,k)-anonymat [70], la t-proximité [46] et la (c,k)-sécurité [48]. Nous pouvons toutefois rassembler la plupart de ces techniques sous l'ombrelle du *LKC*-anonymat, présenté ci-bas.

Le *LKC*-anonymat [50] [61] [68] naît de l'hypothèse audacieuse qui veut que, dans une situation de divulgation courante, il est *très difficile* pour un adversaire d'acquérir *toute l'information* sur un tuple. Suivant cette hypothèse, nous pouvons, en bornant l'information obtenable par l'adversaire, réduire les contraintes du k-anonymat sans augmenter de façon innaceptable le risque de divulgation accidentelle.

Plus spécifiquement, nous supposons que l'information *a priori* de l'adversaire est contenue dans des vecteurs de taille au plus L. Par exemple, si nous avons 12 champs QID et que $L = 3$, alors l'adversaire ne peut lancer une attaque qu'en utilisant un vecteur compris d'un seul, 2 ou 3 champs parmi les 12.

Plus formellement, posons $T[qid]$ comme dans la Section 2.3.1 et appelons *Sens* l'ensemble des valeurs sensibles disponibles en T.

Définition 2.3.2. *LKC*-anonymat [50] : Soit L le nombre maximal de valeurs dans la connaissance *a priori* de *Adv* et $S \subseteq Sens$ un ensemble de données sensibles. Une table T satisfait le *LKC*-anonymat si et seulement si pour tout *qid* avec $|qid| \leq L$

1. $|T[qid]| \geq K$, où $K > 0$ est un entier[8], et
2. $P(s|qid) \leq C$ pour tout $s \in S$, où $0 < C \leq 1$ est le seuil de confiance.

[8] Notons la similitude avec le k-anonymat.

Les paramètres L, K, C sont définis par le propriétaire des données. Le pouvoir de l'adversaire réside dans le choix du paramètre L. La probabilite d'une attaque par lien identitaire est alors garantie d'être au plus $1/K$ et celle d'une attaque par lien attributif est alors au plus C.

Le *LKC*-anonymat possède deux propriétés intéressantes : en premier lieu, il généralise un bon nombre des extensions au k-anonymat, simplifiant la mise en place d'algorithmes traditionnels. Le Tableau 2.V résume cette généralisation. En second lieu, l'application à des ensembles de données à plusieurs dimensions est grandement simplifiée. Nous avons simplement besoin qu'un sous ensemble des QID soit partagé par au moins K tuples, contrairement à tous les QID pour le k-anonymat traditionnel.

Tableau 2.V – k-anonymat et dérivés en fonction du *LKC*-anonymat

Méthode	Paramètres utilises	Émulation via le *LKC*-anonymat
k-anonymat	K	$L = \|QID\|, K = k, C = 1$
Confidence bounding [69][9]	Aucun	$L = \|QID\|, K = 1, C = 0$
(α, k)-anonymat [70]	α, k	$L = \|QID\|, K = k, C = \alpha$

Le *LKC*-anonymat n'est cependant pas sans défaut [10]. Notons le travail supplémentaire non-négligeable que cette approche demande afin de s'assurer de la pertinence de considérer chaque attribut non-sensible comme un QID potentiel. Ce travail de contextualisation pour notre problème est fait en Section 3.8.

2.3.6 La confidentialité différentielle

La confidentialité différentielle [18] [14] [15] ou intimité différentielle se démarque des tentatives précédentes de formalisation de la confidentialité, notamment du k-anonymat, par une perspective différente de l'enjeu. Plutôt que de tenter de regrouper les individus dans des « grappes » où tous les tuples sont similaires, nous tentons ici de limiter le risque supplémentaire lié à l'utilisation d'un service où ses données se retrouvent dans l'ensemble à l'étude. Nous dirons qu'une opération sur un ensemble de données $BD(A)$ est différentiellement confidentielle[10] si nous pouvons inférer environ la même quantité

[9] Formalisation de la dépersonalisation par quasi-identificateurs.

[10] de l'anglais *differentially private*.

d'information sur un individu Y, *qu'il soit présent ou non dans* $BD(A)$.

Plus formellement, nous pouvons dire :

Définition 2.3.3. ε -confidentialité différentielle : Un algorithme randomisé Ad est ε-différentiellement confidentiel si, pour tous les ensembles de données BD_1 et BD_2 différents d'au plus les données d'un seul individu et $\hat{D} \subseteq Range(Ad)$:

$$\Pr[Ad(BD_1) \in \hat{D}] \leq e^{\varepsilon} \cdot \Pr[Ad(BD_2) \in \hat{D}] \tag{2.2}$$

où les probabilités sont sur celles de l'algorithme Ad.

Nous pouvons également relaxer la définition en ajoutant un paramètre additif δ qui nous permet d'ignorer les événements très rares [17].

Définition 2.3.4. (ε, δ) -confidentialité différentielle : Un algorithme randomisé Ad est (ε, δ)-différentiellement confidentiel si, pour tous les ensembles de données BD_1 et BD_2 différents d'au plus les données d'un seul individu et $\hat{D} \subseteq Range(Ad)$:

$$\Pr[Ad(BD_1) \in \hat{D}] \leq e^{\varepsilon} \cdot \Pr[Ad(BD_2) \in \hat{D}] + \delta \tag{2.3}$$

où les probabilités sont celles de l'algorithme Ad.

Les paramètres susmentionnés sont connus publiquement et leur valeur est inversement proportionnelle aux garanties de confidentialités suggérées par l'algorithme.

Cette définition peut paraître à prime abord comme une définition faible de la confidentialité. En effet, notre but est idéalement de protéger les renseignements sensibles contenus dans notre ensemble de données. En 1977, Dalenius a proposé la définition suivante, que nous appelerons *confidentialité parfaite* :

Aucune information sur un individu ne devrait être apprise en consultant la base de données $BD(A)$ *qui ne pourrait pas être apprise sans la consulter.*

Dwork [18] a prouvé que cette contrainte est impossible et cette inéquation a motivé la création de la définition de la confidentialité différentielle.

2.3.7 Confidentialité différentielle : exemple d'application

Tout comme pour le *k*-anonymat, la confidentialité différentielle ne prescrit aucun algorithme servant à sa mise en place. Toutefois, sur un ensemble de données limité, nous pouvons plus facilement illustrer un exemple d'application d'un algorithme naïf.

Reprenons l'ensemble de données *k*-anonymisé de la Section 2.3.3. Nous utiliserons la méthode simple de la *table de contingence*. Pour cette application, nous devons avoir les données regroupées en bloc. Par souci de simplicité, nous avons conservé uniquement les paramètres ayant des données.

Tableau 2.VI – Ensemble de données-type regroupé

Date de naissance	Sexe	Cause de sélection	Nombre
$\leq$ 1970-12-31	F	IMC hors normes	3
$\leq$ 1970-12-31	M	Hypertension	2
$\leq$ 1970-12-31	F	Autre	2
$\leq$ 1970-12-31	M	Autre	2

Afin d'obtenir un ensemble de données différentiellement confidentiel, nous ajouterons du bruit au nombre d'individus dans chaque groupe. La sélection de la distribution pour la quantité de bruit diffère selon le but et les caractéristiques souhaitées, mais l'utilisation de bruit laplacien [42] ou alors issu d'une distribution géométrique [28] sont les méthodes les plus fréquemment rencontrées. Afin de garder l'exemple simple et facile à comprendre, nous utiliserons le résultat d'un dé non pipé[11].

Simulons quatre valeurs de notre distribution : nous obtenons par exemple 1, 4, 4, 6. Ajoutons les valeurs dans l'ordre du vecteur `Nombre` de notre tableau pour obtenir

$$4, \qquad 6, \qquad 6, \qquad 8.$$

Advenant que l'un des individus présent dans la table d'origine n'aurait pas conscenti à partager ses données, il nous suffirait de supprimer son tuple, aggréger de nouveau et d'appliquer une perturbation similaire. Le domaine de $n + U(1,6)$ et $n - 1 + U(1,6)$ *i.i.d.* diffère d'une valeur ($n + 6$ est absent de la seconde et n de la première). Nous aurons confirmation de la présence du tuple uniquement dans le cas suivant :

[11] ou, pour être plus formel, le résultat d'une distribution uniforme sur le support $\{1,2,\ldots,6\}$.

- le premier jet de dé a donné 6 (donc nous avons $n+6$ pour le tuple dans le paramètre nombre) ;
- le second jet de dé a donné 1 (donc nous avons $n-1+1=n$ pour le tuple dans le paramètre nombre).

Encore une fois, la taille de notre exemple mène à des résultats inintéressants. Notons par exemple la difficulté de calculer les paramètres. En effet, le calcul de ε (et de δ par le fait même) dépend de nombreux facteurs, notamment le nombre de requêtes à l'ensemble de données et leur nature. Pour cette raison, il nous est impossible de dériver facilement les paramètres dans un exemple. Une preuve pour l'utilisation d'un bruit Laplacien est disponible en [16] et [19].

La confidentialité différentielle présente un modèle novateur de mesure de l'anonymat d'un ensemble de données. Toutefois, il est important de reconnaitre les limites de celui-ci [10]. De par sa nature, le but de la confidentialité différentielle n'est pas de préserver le secret sur un individu mais bien de limiter le risque de divulgation lors de son inclusion dans la table. Les garanties sont également tributaires du type de requête utilisé et se prêtent mal à un partage d'ensemble de données, puisque nous n'avons plus le contrôle une fois les données divulguées.

L'ajout de bruit est également un problème inhérent à l'application de la confidentialité différentielle et préviendra son application dans la majorité des cas étudiés ici. Plusieurs ensembles de données sont affublés de tuples considérés *extrêmes* et peuvent augmenter de façon ridicule la sensiblité d'une requête. Les auteurs en [59] montrent l'exemple avec une requête simple (la moyenne du revenu annuel aux États-Unis, utilisant $\varepsilon = 0.25$) : la réponse dévie de la valeur réelle de 10 000$ ou moins seulement 3% du temps ! Les données d'assurance sont généralement très sensibles aux paramètres de classification et de mesure et une divergence par rapport aux valeurs réelles peut poser un réel préjudice.

Finalement, les fondements plus théoriques de la confidentialité différentielle compliquent la mise en place auprès d'une clientèle plus orientée affaires. Bien que la recherche soit encore effervescente dans le domaine [32] [20], les limitations encore pré-

sentes ont le malheur d'effrayer les preneurs de décisions. Les méthodes syntaxiques, avec le k/LKC-anonymat en tête, présentent une courbe de familiarisation plus intuitive et sont plus faciles d'adoption.

2.4 La mesure de l'utilité d'un ensemble de données

Suite à une opération d'anonymisation, il convient de mesurer si notre nouvel ensemble de données fournira des résultats satisfaisants pour la tâche demandée. À cette fin, certaines métriques font souvent partie intégrante de la présentation d'une méthode d'anonymisation. L'utilisation de certaines mesures permet également d'orienter les efforts d'anonymisation tout au long du processus, si celui-ci est interactif.

À chaque méthode sa métrique. En se basant sur les deux grandes familles de méthodes d'anonymisation (voir la Section 2.3), une proposition d'amélioration pour un problème bien défini consiste souvent en l'application du même algorithme sur l'ensemble de données original et anonymisé. Les résultats de l'algorithme seront alors comparés et les mesures émanant de l'analyse seront utilisés pour mesurer l'utilité d'une technique pour un algorithme défini.

Cette approche cause une multiplication de méthodes d'utilité, toutes similaires dans leur but mais différentes dans leur application. Nous nous limiterons à celles d'intérêt général par souci de pertinence, sans toutefois passer sous silence celles susceptibles de s'appliquer à notre domaine. Nous nous limiterons également aux mesures appliquées aux méthodes syntaxiques (dérivées du k-anonymat) étant donné que celles appliquées à la confidentialité différentielle dépendent de la requête posée[12].

Une première mesure d'optimalité de notre opération d'anonymisation est le nombre d'opérations de suppression et de généralisation. Le problème défini de cette façon a été prouvé NP-difficile tant sur la généralisation/suppression des attributs que des tuples [49]. Contrairement aux autres mesures, celle-ci reste indépendante du type de données et de l'analyse faite. Dans le cas de données hautement qualitatives où les hiérarchies de généralisation sont bien définies, cependant, cette mesure peut être une ex-

[12]Dans le cas d'une requête de comptage, par exemple, le biais entraîné est facilement calculable par une différence.

cellente façon de comparer deux itérations.

Notons également la mesure d'utilité prédictive de [12], séparant son ensemble de données (2/3 pour l'entraînement du modèle et 1/3 pour la prédiction) de façon répétée afin de mesurer l'impact sous différents paramètres K (sous un dérivé du k-anonymat, le LKC-anonymat). Sur un modèle de classification [50], l'*erreur de classification*[13] versus l'*erreur de base*[14] est utilisée.

Dans le cas où l'algorithme n'est pas immédiatement connu, le score de discernabilité pour les ensembles de données généralisables semble être un proxy populaire [62]. L'idée générale est de donner une pénalité à un tuple pour être indistinguable des autres tuples. Pour chaque tuple dans un groupe équivalement *qid* (où tous les QID sont égaux à *qid*), la pénalité est de $|T[qid]|$. Nous pouvons alors facilement voir que la pénalité au sein d'un groupe est $|T[qid]| \times |T[qid]| = |T[qid]|^2$. Il est possible d'optimiser une généralisation en minimisant la somme de ce score sur l'ensemble des groupes formés. La borne inférieure de ce score est bien entendu $|T|$, soit le nombre de tuples dans la tables.

$$Score(t) = \sum_{qid_t} |T[qid_t]| \ \forall t \in T \tag{2.4}$$

Équation d'optimisation : score de discernabilité

Similaire au score de discernabilité et toujours pour les dérivés du k-anonymat, nous avons à notre disposition la *normalized average equivalence class size metric* que nous résumerons par C_A[15]. Cette métrique, voulue minimisée, favorise aussi un grand nombre de classes d'équivalence mais tient en compte le paramètre K utilisé lors de l'anonymisation. Nous favorisons de cette façon un K plus élevé et compensons ainsi la perte de précision entraînée par ce choix.

[13] de l'anglais *classification error (CE).*

[14] de l'anglais *baseline error (BE).* Il s'agit de l'erreur de classification du modèle sur les données originales.

[15] Une traduction aurait inutilement alourdi le texte.

$$C_A = \frac{K \times \#\text{ de tuples}}{\#\text{ de classes d'équivalence}} \tag{2.5}$$

Équation d'optimisation : C_A

2.5 Ensemble : le compromis confidentialité-utilité

On présente souvent les conflits inhérents à la confidentialité des microdonnées comme une dualité entre l'utilité et la confidentialité. La polarisation est évidente : nous pouvons d'un côté obtenir une utilité maximale en faisant fi de la confidentialité et en publiant les tuples bruts. De l'autre, nous pouvons émettre des tuples aléatoires afin d'obtenir une confidentialité espérée maximale, au détriment de l'utilité ici nulle [28].

En entreprise, il convient d'ajouter un élément financier à ce compromis : une stratégie de confidentialité doit, en plus de conjuguer les deux pôles précédemment mentionnées, être efficace sur le plan financier. En considérant les données comme un actif distinct, on peut plus facilement intégrer une stratégie de divulgation confidentielle au sein d'une entreprise.

La quantification, financière ou empirique, de la valeur résiduelle des données suite à une opération d'anonymisation est un sujet fréquemment considéré dans la littérature. Une quantification bien orchestrée dépend souvent d'une connaissance intime du domaine d'application des données. On peut toutefois rassembler les méthodes de tarification sous deux grandes familles :

- celle *constructive*, où l'on bâtit le prix selon les caractéristiques et la quantité de données ;
- celle *prospective* où l'on donne plutôt un prix à la divulgation ou un coût à la protection desdits renseignements.

Pour les acteurs en préservation de la confidentialité, la seconde approche parait plus sensée dans une optique de protection des renseignements des usagers. On parlera alors de *coût de la confidentialité* [2]. Notons qu'une tarification selon le prix à la divulgation est généralement plus élevée que lorsque nous utilisons l'approche basée sur un coût sur

protection. Dans une approche de gestion de risques, l'aversion face à une perte semble expliquer cette dichotomie.

Cette méthode suppose généralement une relation de gré-à-gré entre l'utilisateur des données et les fournisseurs (celles et ceux fournissant leurs renseignements personnels). Dans notre cas, il s'agit d'une hypothèse impossible à combler : le propriétaire des données n'a souvent aucun lien personnel avec les renseignements contenus. Les principes béhavioraux (sic) des travaux cités plus haut ont alors une portée limitée. Toutefois, cette approche nous servira lors de la construction des hypothèses de base pour un partage réussi.

Les auteurs en [39] ont présenté pour les détenteurs d'informations médicales un modèle combinant différents facteurs de coûts associés au cycle de vie des données pour fournir un comprimis utilité-confidentialité en terme de valeur monétaire. Le modèle analytique regroupe l'impact financier sous deux forces opposées : la valeur monétaire de l'ensemble de données anonymisé et le coût potentiel des dommages.

La valeur monétaire de l'ensemble de données est définie sous la base constructive de la façon suivante :

$$\left.\begin{array}{ll} & \text{coût par attribut} \\ \times & \text{nombre d'attributs} \\ \times & \text{nombre de tuples} \\ \times & \text{sensibilité} \end{array}\right\} - \text{Coût d'anonymisation} = \text{Valeur de l'ensemble anonymisé} \tag{2.6}$$

s'appuyant sur les caractéristiques économiques les plus élémentaires des données. Le coût par attribut, puisque lié directement à la nature des données collectées, est bien entendu le facteur le plus déterminant quant à la valeur constructive de l'ensemble de données brutes.

Le coût sur les dommages est de son côté basé sur une approche probabiliste quant à

une fuite sur les données. Il est construit par les auteurs comme

$$\begin{aligned}\text{Coût sur dommages} = \Pr[\text{fuite}] \times \text{compensation sur fuite} \\ + \text{Coûts fixes}\end{aligned} \tag{2.7}$$

La valeur nette est alors la différence entre les deux, et les auteurs maximisent cette différence pour obtenir la solution optimale.

L'utilisation d'une théorie basée sur le prix des données apparait séduisante et permet de mettre un contexte financier sur les opérations de préservation de l'anonymat lors d'un partage. Cependant, il nous apparait peu commode d'utiliser une méthode purement économique afin de mesurer l'utilité d'un ensemble de données : l'absence de concensus quant au juste prix d'un type de paramètre, un tuple donné, etc. nous empêche d'utiliser cette technique de façon robuste. De plus, cette approche balaie le concept d'utilité basée sur la capacité d'analyse, qui est au centre des efforts de préservation de l'utilité. Cette technique prend tout son sens lorsque nous devons communiquer les risques inhérents à une mauvaise gestion des renseignements confidentiels et peut néanmoins servir comme outil de promotion.

2.6 Les systèmes de divulgation existants

Il est difficile de traiter de confidentialité pratique sans mentionner quelques systèmes œuvrant déjà dans le domaine. Hélas, peu de joueurs sont encore actifs et ARX (ci-bas) jouit d'une popularité éclipsant les autres. Nous avons limité notre recherche aux systèmes disponibles sous source libre.

μ-ARGUS et τ-ARGUS [31] sont issus du domaine de la statistique, notamment du *Statistical Disclosure Control (SDC)*. Plusieurs algorithmes sont disponibles d'office ainsi que de nombreuses manipulations sur les données (fusion, suppression, etc.). Le logiciel se spécialise dans le partage de données *many-to-one*, où plusieurs contributeurs versent leurs données dans une table commune (qui servira à l'analyse). Le logiciel tente de prévenir le risque qu'un contributeur important puisse inférer la contribution

des autres en faisant la différence de l'ensemble résultant sur les données qu'il a contribué. Nous verrons en Section 3.1 que ce scénario est un cas spécifique de ce que nous appelerons un partage à *récipiendaires discrets*.

Le UTD Anonymization Toolbox [67] implémente 6 algorithmes selon 3 définitions différentes (notamment le k-anonymat multidimensionnel). Le programme consiste en plusieurs paquets Java liés ensemble par un fichier de configuration XML. La méthode de paramétrisation est à notre avis compliquée, peu flexible et empêche un travail interactif.

Le *Cornell Anonymization Toolkit* [71] (CAT) a été développé dans le but de fournir un système utilisant la l-diversité [47]. Le système procède de façon similaire à notre modèle (anonymisation à priori, puis ajustement selon les risques et l'utilité). Le type d'analyse est limité aux données de recensement et l'utilité semble être basée sur une comparaison des densités. Le développement est inactif et aucune activité n'a été publiée sur le dépôt de code depuis le début de l'an 2013.

ARX [22] se démarque par une interface graphique polie et facile d'utilisation. Le programme se spécialise dans les données biomédicales. ARX est largement documenté et possède une API[16] intuitive qui a guidé nos premiers tests en anonymat appliqué. L'absence de *LKC*-anonymat et une hiérarchisation des données numériques (plus spécialement, les dates) inégale nous a toutefois empêché de l'utiliser à son plein potentiel.

Finalement, quelques logithèques sont disponibles pour des langages de programmation. Étant donné l'omniprésence du langage R [65] en statistiques, la bibliothèque la plus populaire est `sdcTable` [4]. Encore une fois, les méthodes employées sont surtout issues de la SDC : ici, les algorithmes proposés se résument à supprimer des tuples considérés sensibles. Le développement semble se limiter à des corrections de bogues depuis un peu plus d'un an.

[16] *Application Programming Interface*

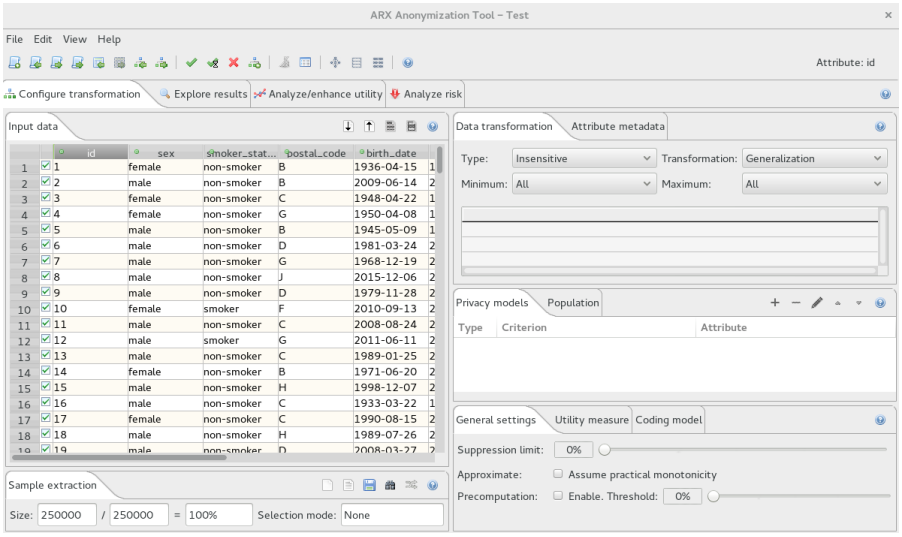


Figure 2.2 – Capture d'écran de l'outil ARX en action.

CHAPITRE 3

MÉTHODOLOGIE

Le chapitre précédent nous a permis de jeter les bases des problématiques rencontrées, des solutions proposées et des compromis faits dans le domaine du partage anonymisé des données. Nous commencerons par présenter les fondations sur lesquelles notre système s'appuie, avant de s'attaquer aux hypothèses sous-jacentes et à son fonctionnement. Le but est de comprendre adéquatement la problématique rencontrée ainsi que les étapes menant à sa résolution.

3.1 Hypothèses et scénario type

Avant de s'embarquer dans la présentation exhaustive du système, il convient de définir quelques hypothèses-clés. Nous commencerons par des scénarios de transmission et d'attaques ainsi que par les principaux acteurs. Nous nous inspirerons des scénarios de transmission de données issues du domaine de la cryptographie [38], par leur complétude et leur facilité de compréhension.

Dans un scénario classique de transmission de données bipartie, nous avons traditionnellement trois acteurs :**Alice** (A) qui tente de communiquer avec **Bob** (B) souhaite également éviter qu'un espion (ici appelé **Eve** (E)) puisse lire la communication.

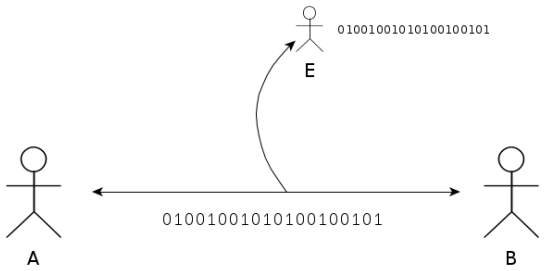


Figure 3.1 – Alice et Bob communiquant sans cryptage

Dans ce scénario, le contenu du message importe peu : nous souhaitons que le récipiendaire soit le seul capable de le lire.

Dans un scénario de divulgation bipartie, la première différence notable est que *Bob et Eve sont la même entité*. Cela implique automatiquement un changement d'objectif : nous souhaitons en tant qu'émetteur que le récipiendaire puisse inférer les bons résultats de la divulgation, sans toutefois pouvoir identifier les individus contenus dans l'ensemble de données. Formellement, nous souhaitons que la connaissance *a priori* sur chacun des individus présents ou non dans l'ensemble de données divulgué reste environ identique[1] à la connaissance *a postériori* du récipiendaire-adversaire (que nous appellerons désormais simplement *récipiendaire*).

Nos divulgations seront rassemblées en deux types : une divulgation dite *à récipiendaires discrets* (où l'on connait leur nombres) et une divulgation *publique*. Le tableau 3.I résume les principales différences entre les deux types.

Tableau 3.I – Différences entre les deux types de divulgation

Récipiendaires discrets (Figure 3.2)	**Public** (Figure 3.3)
Nous pouvons plus facilement[2] faire des hypothèses sur la connaissance des récipiendaires.	Nous ne pouvons pas trivialement limiter la connaissance des récipiendaires.
L'échange est discret : nous savons quand les récipiendaires acquièrent les données.	Les récipiendaires peuvent acquérir les données n'importe quand suite à la divulgation.
La connaissance des récipiendaires est limitée à leurs ensembles de données $BD(B_i)$ internes (aucun partage).	La connaissance des récipiendaires peut potentiellement être combinée.
Suite à une fuite publique, la connaissance des récipiendaires est augmentée.	

Afin de rester réaliste avec le contexte économique dans lequel les acteurs évoluent, nous considérerons que les récipiendaires discrets se comportent en **optimiseurs égoïstes** : ceux-ci tentent de maximiser l'inférence qu'ils peuvent obtenir sur l'ensemble de données sans toutefois aider les autres récipiendaires. Le partage de connaissance est alors limité au minimum. Il est facile de voir qu'il s'agit de la situation pour un bon

[1] Rappelons-nous qu'il est impossible de garantir qu'elle reste identique à 100% (voir la section 2.3.6)
[2] Cela ne veut pas dire qu'il est facile, juste plus aisé.

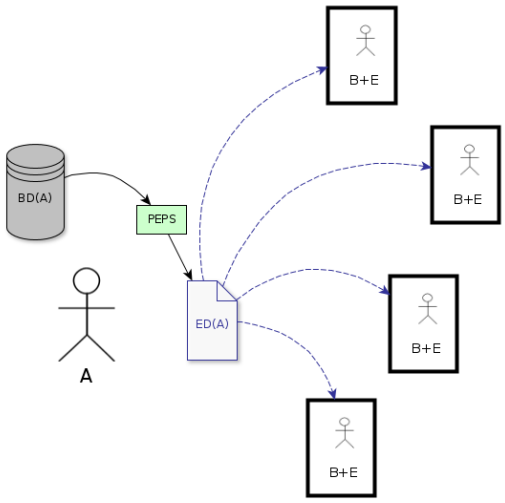


Figure 3.2 – Alice divulgue un ensemble de données anonymisé à quatre récipiendaires.

nombre d'industries classiques de notre jour : votre assureur ne partage pas votre prime et votre questionnaire de sélection à moins d'y être absolument contraint ! Dans le cadre d'une divulgation publique, cette hypothèse tombe pour faire place à celle de *communauté intelligente* : la connaissance acquise par un des participants peut devenir celle initiale d'un autre. Il nous sera alors impossible d'utiliser les primitives de confidentialité qui dépendent de la connaissance du récipiendaire. Cette hypothèse se prête bien aux divulgations similaires à celle que l'on a vu en Section 1.3.

Il arrive souvent que le protocole de divulgation passe par un intermédiaire agrégateur. Nous pouvons penser aux études d'industrie, où les participants peuvent partager leurs données à un tiers parti qui procède avec l'analyse avant de publier les résultats. Selon le récipiendaire des données finales — si elles sont partagées[3] — nous pouvons le poser sous l'ombrelle d'une ou l'autre des techniques susmentionnées.

[3] Si uniquement les résultats d'analyse sont publiés, nous espérons que les résultats sont suffisamment agrégés !

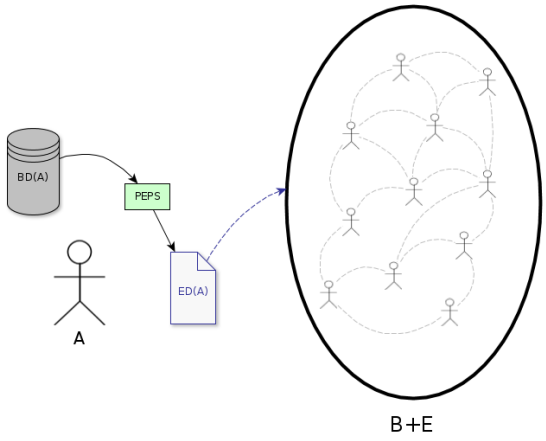


Figure 3.3 – Alice divulgue un ensemble de données anonymisé publiquement.

3.2 Architecture proposée

À la lumière des objectifs proposés en section 1.1, nous avons conçu le système PEPS[4]. Étant donné l'effervescence de la recherche dans le domaine de la confidentialité et de l'économie numérique, nous avons privilégié un système le plus modulaire possible. Cela facilite l'audit général du tout : chaque composante est responsable d'un nombre limité d'opérations et est remplaçable sans trop de difficultés si nous jugeons pertinent de le faire. À cette fin, il nous semble approprié de décrire le système d'abord dans son ensemble, puis d'énumérer les éléments en les caractérisant notamment par leurs intrants et extrants (voir la section 3.3).

Le système s'inspire du principe de la boucle de rétroaction, couramment vue dans les modèles d'optimisation. La figure 3.4 présente les trois modules principaux (octogones bleus) ainsi que les intrants-extrants (rectangles verts).

Chaque nœud est présenté dans la section suivante, en débutant par la caractérisation des intrants-extrants puis celle des modules.

[4]Pour Privacy Experimentative Profiling System

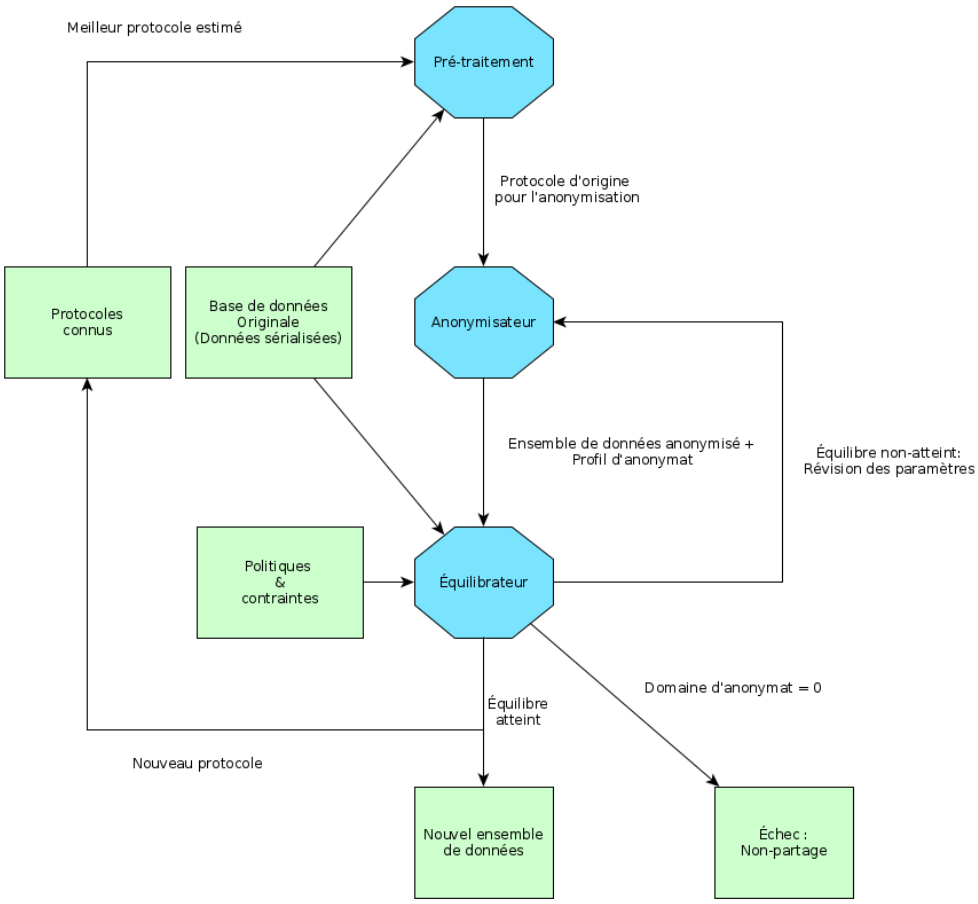


Figure 3.4 – PEPS : Aperçu du système de base, sans perturbation

3.3 Description des nœuds

3.3.1 Base de données originale

Intrants : Données collectées

Extrants : Données sérialisées

Nous considérons ici *base de données originale* l'ensemble des sources – le plus souvent internes – disponibles à l'entreprise. Par souci de simplicité, nous considérons toutes les sources, leur croisement ainsi que l'inférence émanant d'autres mécanismes que le système ici décrit comme étant sous l'égide de cet élément.

Nous supposons que les données sont immuables à un moment précis : il nous est possible de savoir l'état de $BD(A)$ pour chacune des divulgations précédentes. Nous éviterons de prescrire une méthode précise pour respecter cette condition, étant donné la variété des environnements propres aux entreprises. Dans notre cas, un champ du type *ajouté le* nous sera suffisant. Comme pour la plupart des bases de données auditables, il est hélas jugé mauvaise pratique de supprimer des tuples. Nous aurons plutôt tendance à étiqueter les tuples inactifs et les dater.

Les données originales alimentent le **module de prétraitement** ainsi que celui d'**équilibration.**.

3.3.2 Politiques et contraintes

Intrants : Politiques et contraintes
Extrants : Codification des contraintes externes

Les *politiques et contraintes* font partie, avec les attaques (voir section 3.4), des forces dites externes : celles-ci ne dépendent pas de l'usage du système. Elles aident au démarrage du système, permettant d'avoir un premier ensemble de règles sur lesquelles les protocoles originaux peuvent se baser. L'encodage des politiques et contraintes, de par leur nature, doit se faire manuellement par les principaux concernés.

Bien que les politiques et contraintes ont pour la grande majorité un impact sur les finances des acteurs concernés, il convient de sous-diviser celles-ci en plusieurs groupes. Nous avons ici choisi de catégoriser les politiques et contraintes selon leur provenance : cela permet de déterminer la façon de les traiter plus rapidement.

Parmi les principaux types de politiques, nous retrouvons :

- les projets et lois concernant la protection de la vie privée (niveau législatif) ;
- les règlements internes et industriels quant à la collecte, l'utilisation et la sauvegarde des données personnelles (niveau entreprises et organisations) :
- la nature des contrats de partage des données (niveau gré à gré).

Parmi les principales contraintes, nous pouvons voir

- les contraintes physiques (espace disque, performance des analyses, etc.) ;
- les contraintes financières directes (coût de la conservation des données, services d'audit, etc.) ;
- les contraintes réputationnelles (opinion du public quant aux politiques de confidentialité).

Les politiques et contraintes posent le plus souvent un seuil minimal quant à la confidentialité à observer ; plus rarement, nous verrons un seuil minimal quant à l'utilité. Puisque nous tentons de maximiser à la fois la confidentialité et l'utilité, il ne nous apparait pas logique d'avoir un seuil maximal de confidentialité[5] ou d'utilité : les forces économiques se chargent de cet élément.

3.3.3 Protocoles connus

Intrants : Protocoles réussis.
Extrants : Protocole optimal pour un attribut et/ou un contrat donné.

Nous considérons comme protocole l'ensemble des spécifications d'attributs menant à une sortie réussie du système. Un protocole type comporte :

- une codification de la demande (pour la retrouver facilement lors d'une divulgation subséquente) ;
- les paramètres choisis pour l'algorithme d'anonymisation ;
- les hiérarchies et/ou transformations ;
- le score d'utilité (si disponible lors de la divulgation) ;

En fonctionnement normal, nous gardons simplement un dictionnaire (ou table de correspondance) des protocoles : lors d'une nouvelle demande au système, nous démarrons avec le ou les protocoles le plus adaptés pour la situation.

[5] Du moins, pas explicitement...

3.3.4 Module de prétraitement

Intrants : Données sérialisées $ED(A)$, protocoles applicables, règles

Extrants : Protocole retenu et données prétraitées $ED'(A)$

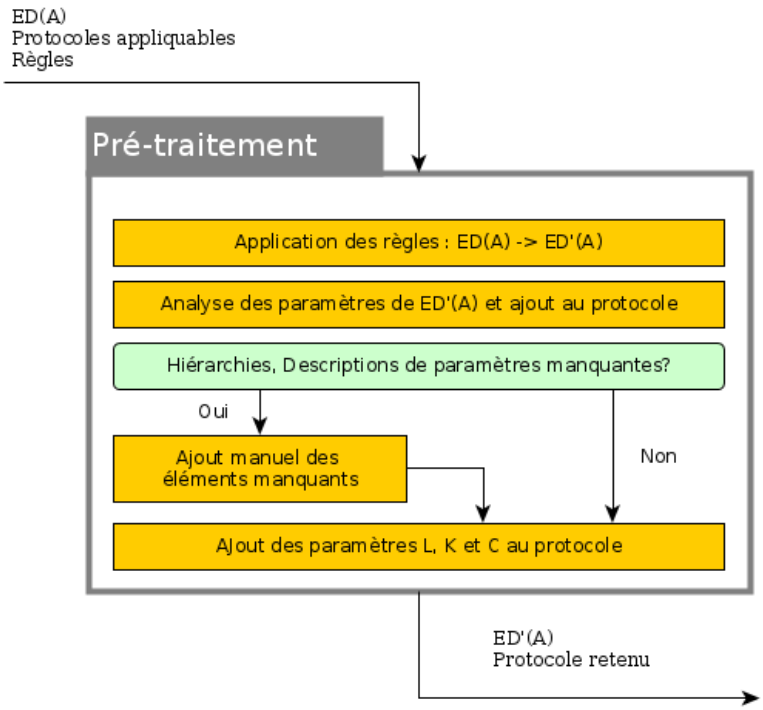


Figure 3.5 – PEPS : Module de prétraitement

Le prétraitement est le premier module de PEPS. Celui-ci permet à l'usager de spécifier les types d'attributs de la table $ED(A)$ (ceux-ci ont été décrits en Section 2.1), ainsi que leurs hiérarchies. Cette étape nettoie également les données des tuples inutiles ou trop sensibles, par exemple suite à une attaque.

En premier lieu, les règles prescrites, provenant de source externe ou rendues nécessaires suite à une attaque réussie (voir la Section 3.4) sont appliquées à $ED(A)$, qui devient $ED'(A)$. Ensuite, nous passons en revue les attributs. Une fois ceux-ci identifiés, les protocoles connus serviront à la proposition du ou des meilleurs candidats pour chaque attribut, ainsi que les hiérarchies précédemment utilisées. Si le système rencontre un at-

tribut pour la première fois, l'usager devra codifier manuellement le type ainsi qu'une hiérarchie si nécessaire. Il est possible, si la hiérarchie n'est pas connue à priori ou si l'usager est indifférent, de laisser le système décider de lui même : sans contextualisation, notre algorithme reste *LKC*-anonyme (voir la Section 3.7.1).

3.3.5 Module d'anonymisation

Intrants : Données prétraitées, protocole retenu.
Extrants : Données anonymisées, protocole utilisé et (facultatif) score préliminaire.

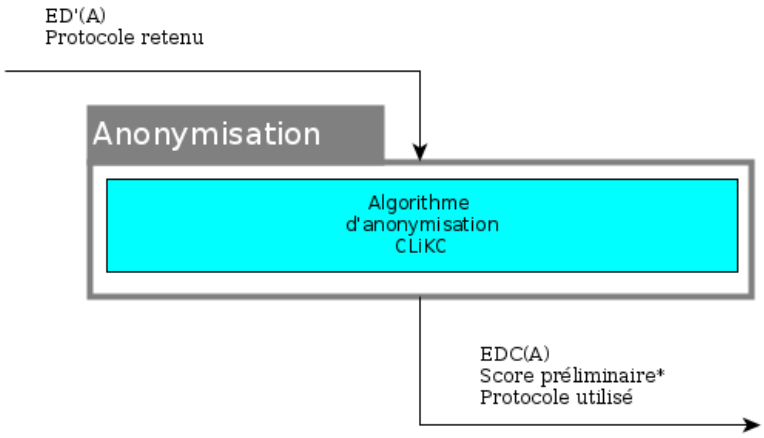


Figure 3.6 – PEPS : module d'anonymisation

Le module d'anonymisation est central au bon fonctionnement de notre système. Il est en effet difficile de concevoir un système de confidentialisation des données sans cette étape cruciale !

L'algorithme sélectionné est un dérivé de l'algorithme PAIP [50], adapté à une forte contextualisation des données passées en entrée. Celui-ci, dénommé CLiKC, est présenté en Section 3.7.

Nous avons simplifié au maximum les intrants et extrants de ce module. Nous voulons en effet permettre à l'usager de faire évoluer le système en modifiant l'algorithme ou en l'améliorant au besoin. Bien que toutes les étapes soient conçues pour être facile-

ment substituables, la priorité de PEPS est d'anonymiser un ensemble de données en vue de le partager et nous sommes sûrs que la recherche future nous contraindra à améliorer ce module.

Il fut brièvement discuté de s'appuyer sur un système d'anonymisation existant afin de faciliter l'intégration et *s'appuyer sur le travail des géants*, en quelque sorte. Cependant, nous avons fait face à deux problèmes majeurs :

- la majorité des systèmes connaissent un développement très lent, voire inactif ;
- le principal candidat, ARX [22], n'a pas implémenté un algorithme de *LKC*-anonymat à ce jour[6].

L'impossibilité d'avoir un algorithme satisfaisant parmi les solutions implémentées a motivé la création du nôtre.

3.3.6 Module d'équilibration

Intrants : Données anonymisées $EDC(A)$, données sérialisées $ED(A)$, score préliminaire (facultatif)
Extrants : Selon le cas

L'équilibrateur tente de fournir un compromis entre l'anonymat et l'utilité résiduelle de l'ensemble de données.

Trois situations peuvent émaner de l'étape d'équilibrage :

- le domaine conjoint d'anonymat et d'utilité est nul : le cas dégénéré se présente lorsqu'il est impossible, considérant les contraintes d'anonymat et d'utilité, d'avoir une sortie satisfaisante. Le système alors échoue ;
- une première sortie remplit le contrat, mais est susceptible d'être améliorée : l'équilibrateur enregistrera alors la tentative si elle est la meilleure rencontrée à

[6]Un billet datant du 24 novembre 2014 est toutefois ouvert. Voir `https://github.com/arx-deidentifier/arx/issues/27`

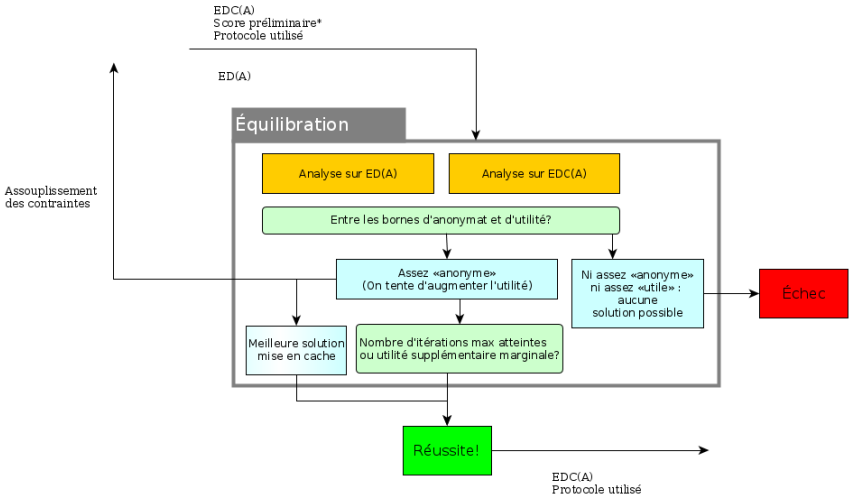


Figure 3.7 – PEPS : module d'équilibration

présent. Ensuite, il tentera d'ajuster les paramètres afin de maximiser l'utilité sans faire basculer le score d'anonymat en deçà du minimum permis ;

- le nombre maximal d'itérations est atteint ou une solution satisfaisante est atteinte : advenant que les itérations améliorent le profil de l'ensemble de données de façon marginale, le système cessera les itérations et émettra une solution dite *en production*. Le protocole retenu sera alors ajouté aux protocoles connus.

3.4 Attaques possibles et PEPS-ajusté

Un système d'anonymisation incapable de s'adapter à la réalité dans lequel il évolue est voué à ne vivre sa courte existence que sur le disque dur de son créateur. Nous souhaitons que PEPS puisse réagir adéquatement face aux perturbations pouvant le toucher.

Par souci de complétude, nous avons préféré approcher les attaques de façon plus générale, en les caractérisant par le vecteur maléfique employé. Nous définirons alors deux types d'attaques.

L'attaque directe ou bris d'anonymat sur données internes Cette attaque a lieu lorsqu'un attaquant complète l'une des situations décrites en Section 2.3.1. Nous avons

donc une brèche sur notre ensemble de données.

Exemple En partageant mes données, un attaquant a été en mesure de déterminer que Monsieur Personne a fait faillite lors de la dernière année. Cette information a été obtenue par une attaque via lien identitaire.

Protocole d'action Il est difficile de déterminer un protocole strict pour ce genre d'attaque. Au minimum, il faudrait hausser les paramètres d'anonymat (dans le cas du k/LKC-anonymat, hausser le K d'au moins 1 pour compenser cet individu désormais connu de tous[7]) ou alors, le supprimer complètement de toutes nos analyses subséquentes (ce qui influencerait notre utilité de façon permanente).

L'attaque indirecte ou l'inférence via un bris externe Celle-ci a lieu lorsqu'une attaque est réussie sur un ensemble de données *dans lequel nous n'avons pas participé*.

Exemple Un compétiteur a partagé un ensemble de données. Suite à ce partage, un attaquant a été en mesure de déterminer que Madame Personne a souffert d'une maladie X en 2014. Cette information a été obtenue par une attaque via lien attributif total (Probabilité = 100%).

Protocole d'action *Pourquoi réagir à une attaque où nous ne sommes pas concernés ?* Dans un domaine où nous traitons avec des individus, il est courant qu'une même personne se retrouve chez plusieurs compagnies[8]. Si une attaque est réussie dans une autre table et que cet individu se retrouve dans la (ou les) nôtre, il faut déterminer si le risque de réidentification est augmenté. Advenant le cas, il est de bon ton de procéder à un *empoisonnement volontaire* du ou des tuples.

[7] Si on peut le réidentifier dans une étude subséquente, il est primordial que l'on ne puisse pas inférer encore plus d'informations sur lui ! À cette fin, on considère son tuple comme *empoisonné* et celui-ci n'aide pas à obtenir un k/LKC-anonymat

[8] Avez-vous votre assurance auto, habitation, collective, et vie au même endroit ?

Exemple alternatif : attaque différentielle Un tiers-parti agrégateur collecte les données de n participants et publie les résultats ainsi qu'un ensemble de données anonymisé. L'un de ces participants fait la différence entre les données anonymisées et les données publiées et publie le résultat. Certains tuples sont alors facilement identifiables.

Protocole d'action Cette attaque découle d'une mauvaise interprétation de la connaissance *a priori* des participants-adversaires. Nous verrons en Section 5.2 comment s'en protéger. Selon le cas, il peut s'agir d'une attaque externe ou interne (selon le propriétaire des tuples identifiés) : les réactions appropriées sont identiques à celles susmentionnées.

La réaction face aux attaques se matérialise dans l'étape du prétraitement. Avant même de démarrer la catégorisation des attributs, nous appliquons un ensemble de règles. Les actions choisies suite à une ou plusieurs attaques se retrouvent dans ces règles.

3.5 Démonstration du système : Ensemble de données synthétique

Tester notre système nécessite de pouvoir comparer l'ensemble de données original et celui anonymisé. Étant donné la nature hautement sensible des données que nous traitons, il s'avère impossible de publiquement utiliser une table provenant d'une compagnie d'assurance. D'un autre côté, nous souhaitons avoir un ensemble de données plausible, avec des champs permettant de bien cerner les avantages et limitations du modèle proposé.

Cette problématique se prête spécialement bien à un outil informatique. À cette fin, nous avons créé l'outil *Synthure*. Celui-ci, selon les paramètres fournis, génère un ensemble de données d'une population fictive d'assurés présentant une corrélation avec les distributions sous-jacente. Une courte présentation de l'outil est disponible en Annexe I et nous présenterons les hypothèses sous-jacentes à l'ensemble simulé dans les sections suivantes.

Finalement, un test est fait sur un ensemble de données public, similaire à celui présenté en Section 1.3. Les champs disponibles seront toutefois plus restreints et les arbres taxonomiques moins élaborés.

3.5.1 Présentation de l'ensemble de données d'assurés

La table `DONNEES_ASSURES`[9] illustre l'information que nous utiliserons pour la démonstration. Nous avons sélectionné les paramètres de sorte à démontrer différentes taxonomies et paramétrisations.

Dans le tableau suivant, un astérisque coiffant le nom du champ indique que celui-ci est une clé primaire (unique). Pour les champs de type *énumération*, ceux-ci sont listés dans les tableaux 3.III et 3.IV :

Tableau 3.II – Champs de la table `DONNEES_ASSURES`

Champ	**Type**	**Description**
ID*	Numérique	Champ d'identification de l'assuré (p. ex. NAS)
Date de naissance	Date	Date de naissance de l'assuré
Date d'assurance	Date	Date de prise d'assurance
Type de produit	Énumération	Type de produit d'assurance souscrit
Sexe	Booléen	`True` pour Femme, `False` pour Homme
Statut fumeur	Booléen	`True` pour Fumeur, `False` pour Non-Fumeur
Score d'assuré	Numérique 2 décimales	Multiple de la prime de base chargée
Score de crédit	Numérique 1 décimale	Score de crédit inspiré du score Equifax/FICO
IMC	Numérique 2 décimales	Indice de masse corporelle
Date d'événement	Date	Optionel : Date de l'événement enregistré
Type d'événement	Énumération	Optionel : Type de l'événement enregistré

Tableau 3.III – Types de produits possibles

Code de type de produit	**Type de produit**
TEMP05	Temporaire 5 ans
TEMP10	Temporaire 10 ans
TEMP20	Temporaire 20 ans
VE	Vie entière
VU	Vie universelle

Les types de produits possibles n'influencent pas directement la mortalité. Cependant, puisque le comportement du consommateur est différent selon la structure du produit vendu, nous avons pris 5 des plus populaires afin d'apporter un peu de variété à

[9] Disponible sur le dépôt de code de l'outil *Synthure*.

notre ensemble. On remarquera l'impact uniquement dans la date de terminaison.

Tableau 3.IV – Codes d'événements possibles

Code d'événement	**Type d'événement**
DEC	Décrément principal : paiement de la police
TER	Terminaison d'assurance
`null`	Par défaut : En vigueur

Nous nous concentrerons sur deux types d'événements possibles en assurance-vie. Le premier est bien entendu le paiement de la police. Le second est la terminaison d'assurance. Les assurés ont l'option de terminer leur couverture à leur convenance : l'occurrence de ces terminaisons a souvent un impact sur la profitabilité des produits vendus. Les polices non touchées par l'un ou l'autre des événements susmentionnés sont alors considérées en vigueur.

Bien que plusieurs autres types d'événements soient possibles, leur étude est souvent plus marginale. La pratique courante consiste souvent à rassembler les événements restants dans l'un ou l'autre :

- Les paiements de montants assurés, qu'ils soient complets ou partiels, seront codés `décrément principal`.
- Les événements menant à un arrêt de réception des primes, sans paiement par rapport au montant assuré, seront codés `terminaison`.

3.5.2 Hypothèses quant à l'ensemble de données

La construction d'un ensemble de données synthétique est tributaire de plusieurs hypothèses. Nous en ferons l'énumération, leur justificatif et leurs compromis dans cette section. Les tables de mortalité et de terminaison utilisées sont disponibles sur le dépôt de code de l'outil *Synthure* (voir en Annexe I). Il est également possible de voir le fonctionnement basique de l'outil à cet endroit.

Pour tous les champs de type date, nous n'avons besoin que du jour, du mois et de l'année : l'heure de l'événement (naissance, prise d'assurance, terminaison et/ou décès)

n'a pas d'importance ici et n'est pas encodée. Nous évitons par le fait même les problèmes de changement de date dûs aux fuseaux horaires (voyage, déménagement, etc.)

3.5.2.1 Hypothèse de distribution des naissances

Nous considérons par simplicité que les naissances sont uniformément distribuées sur le domaine d'étude. En plus d'ignorer la saisonnalité des naissances et le problème du 29 février, cette hypothèse mène à une surreprésentation des assurés plus jeunes[10].

3.5.2.2 Hypothèse de distribution de prise d'assurance

Encore une fois, nous considérons que la date de prise d'assurance est uniforme sur le domaine possible, dans notre cas, des 18es anniversaires jusqu'au 65es (nous excluons les polices juvéniles, prises pour des enfants parfois peu après la naissance). Cette hypothèse contre-balance partiellement la précédente en sous-représentant la prise d'assurance plus élevée à partir de la trentaine jusqu'à environ 60 ans[11].

3.5.2.3 Hypothèse de décrément

Nous utilisons comme hypothèse de décrément principal la dernière table publiée par l'Institut Canadien des Actuaires (ICA), appelée dans ce mémoire `CIA9704` [8].

Il existe 8 tables `CIA9704` : une pour chaque combinaison de sexe (homme/femme), statut fumeur (fumeur/non-fumeur) et méthode de calcul d'âge (âge atteint/âge arrondi). Nous utiliserons les 4 tables « âge atteint » puisque nous avons la date de naissance à notre portée et nous n'avons pas besoin de poser d'hypothèse d'âge arrondi.

Bien que le point milieu des tables analysées remonte à il y a presque 15 ans, nous n'utiliserons pas d'hypothèse d'amélioration des décréments : étant donné que la prise d'assurance est répartie sur un long horizon, ajuster le vecteur de décrément à chaque police nous semblait contre-intuitif et le gain de réalisme nous apparait négligeable.

[10] Le taux de natalité a en effet chuté de 59% dans la période 1960-2013 [74]

[11] La prise d'assurance est souvent corrélée avec un événement de vie marquant, comme l'achat d'une propriété ou la naissance d'un enfant. Cela est également en accord avec l'âge des assurés retrouvé en [34]

3.5.2.4 Hypothèse de terminaison par produit et de renouvellement

Nous utiliserons les études les plus récentes de l'ICA [33] [54] [7] disponibles pour dériver des tables basiques de terminaison des produits retenus.

La plupart des produits temporaires sont renouvelables pour minimalement un second terme : par exemple, une police d'assurance Temporaire 10 ans offrira la possibilité à l'assuré de renouveler pour un second terme de 10 ans, pour une prime plus élevée[12]. Dans le souci de comparer des risques similaires, nous nous limiterons à l'étude des risques pré renouvellement. Nos tables de terminaison pour les produits temporaires auront alors un taux de 100% pour la dernière durée du premier terme. Les produits permanents n'ont évidemment pas de structure de renouvellement et ne seront pas modifiés.

3.5.2.5 Distribution fractionnaire des décès et terminaisons

Les tables de décès et de terminaison sont définies sur bases annuelles. Afin de calculer le jour exact de l'événement, nous faisons l'hypothèse la plus courante en industrie, soit celle d'une *distribution uniforme des décréments* [6].

Définition 3.5.1. Distribution Uniforme des Décréments (DUD) : Posons x variable aléatoire uniforme $[0,1)$ simulée. Considérons que la fonction sur laquelle nous souhaitons inverser u est monotone croissante. Soient alors l la valeur plancher et u a valeur plafond de x et pour lesquelles $l^* = F^{-1}(l)$ et $u^* = F^{-1}(u)$ ont une valeur connue. $x^* = F^{-1}(x)$ sera alors égale à

$$x^* = l^* + \frac{x-l}{u-l} \times (u^* - l^*) \tag{3.1}$$

Par exemple, si nous avons la distribution illustrée en table 3.V et que nous avons simulé la valeur 0.57, nous aurons donc

[12] En plus de considérer la mortalité supplémentaire due à l'âge, un assuré en bonne santé pourra souvent se requalifier pour une nouvelle police en repassant les tests médicaux nécessaires. Considérant que le renouvellement est offert sans tests supplémentaire, les assurés qui renouvellent ont dans l'ensemble une expérience significativement moins bonne que le reste du porte-feuille. Ce phénomène — appelé *antisélection* — est considéré très sérieusement par les assureurs. Pour une discussion de l'antisélection et une tentative de mesure de l'impact de celle-ci, voir [44].

$$4 + \frac{0.57 - 0.5}{0.6 - 0.5} \times (5 - 4) = 4.7$$

Tableau 3.V – Distribution simple pour un exemple de DUD

Valeur	$F(x)$
...	...
4	0.5
5	0.6
...	...

3.5.2.6 Hypothèse de distribution par sexe, statut fumeur

Les paramètres suivants servent à déterminer la table de mortalité utilisée. Les hypothèses ont été choisies selon la répartition typique d'une population d'assurés :

- 50% hommes / 50% femmes ;
- 85% non-fumeurs / 15% fumeurs.

3.5.2.7 Hypothèses sur le score d'assurabilité

Plutôt que de refuser les personnes qui ne se conforment pas parfaitement aux critères idéaux d'assurabilité, les compagnies auront plutôt tendance à les *sur-primer*. Une surprime est souvent calculée comme un pourcentage de la prime initiale et fonctionne par incréments de 25%.

La distribution des scores dépend des règles de tarification et de sélection des risques et est souvent considérée élément concurrentiel des compagnies d'assurance. À cette fin, peu d'information circule à ce sujet à l'extérieur des cercles d'industrie. Nous avons pris une hypothèse respectant l'intuition à ce sujet :

- la grande majorité des risques sont standards (0% de surprime) ;
- la majorité des surprimes sont en deçà de 100% ;

- la surprime maximale est aux alentours de 500%.

Nous avons retenu une distribution exponentielle « plancher » pour notre hypothèse de score. Celle-ci présente l'avantage d'être très facile à inverser (pour les besoins de simulation), en plus d'avoir une queue assez fine pour éviter les valeurs extrêmes.

$$X = 25\% \times (\lfloor Y \rfloor - 1), \;\; Y \sim \exp(\lambda = 0.85) \tag{3.2}$$

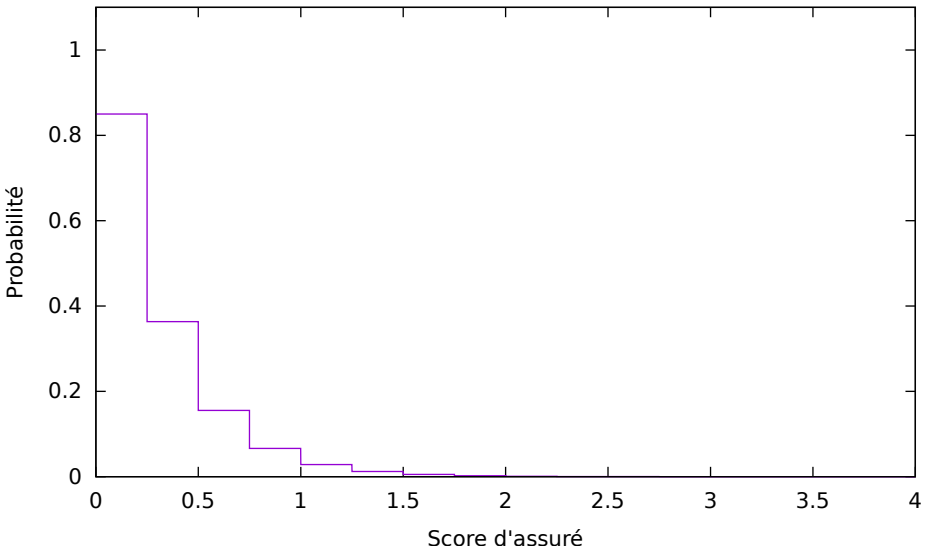


Figure 3.8 – Distribution du score d'assuré

3.5.2.8 Hypothèses sur le score de crédit

De prime abord, un champ tel que le score de crédit peut paraître surprenant pour des données d'assurance de personne. La recherche est cependant très active sur l'utilisation d'un score de crédit ajusté comme mandataire pour la mortalité des assurés [43]. L'utilisation d'un score de ce genre a déjà fait ses preuves en assurance IARD. Le score de crédit présente en outre, tout comme l'IMC (voir la section 3.5.2.9), une excellente opportunité de tester notre algorithme et nos mesures d'utilité sur des données à seuil.

Nous utilisons une distribution issue d'une communication du Gouvernement du Canada [24]. Le but est d'avoir une idée générale de la distribution des scores FICO/Equifax

de la population canadienne. Comme les scores sont présentés sous forme d'histogramme (voir la figure 3.9), nous discrétiserons via une distribution uniforme entre les intervalles.

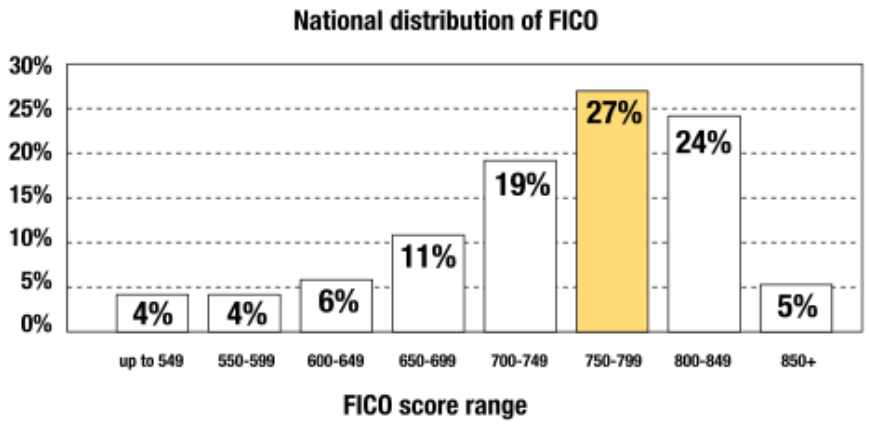


Figure 3.9 – Distribution du score FICO/Equifax dans la population canadienne

3.5.2.9 Hypothèses sur l'IMC

L'indice de masse corporelle (IMC) est un prédicteur classique de la mortalité chez les assurés. Il s'agit d'ailleurs de l'indice à seuil le plus connu auprès de la population. Un IMC élevé est généralement prédicteur de troubles de la santé aux âges plus avancés.

Nous avons simulé l'IMC selon la courbe fournie par Statistiques Canada lors de leur dernière étude complète sur l'obésité [66]. La figure 3.10 provenant de cette étude illustre notre distribution. Encore une fois, nous avons pris les mesures aux valeurs entières (de 15 à 44), puis avons interpolé linéairement entre celles-ci. Nous ne retenons que la première décimale.

3.5.3 Limitations connues

Le but est de produire un ensemble de données qui se conforme aux principes que nous avons identifiés et non de reproduire verbatim ce qui peut se retrouver dans les fichiers d'une compagnie d'assurance. Plusieurs limitations par rapport à un ensemble type se retrouvent volontairement dans notre ensemble de données simulé et nous ferons la discussion des principales dans cette section. Notons cependant que les limitations de

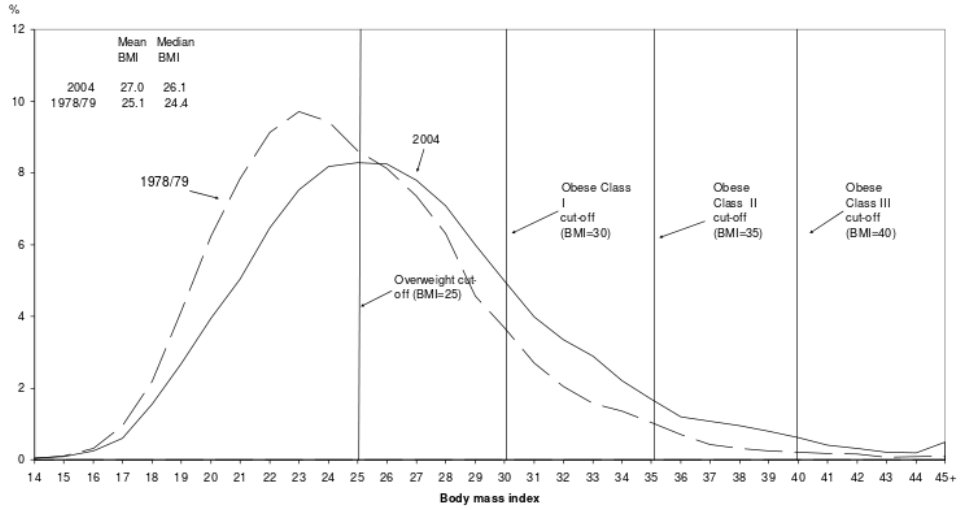


Figure 3.10 – Distribution de l'IMC dans la population canadienne

l'ensemble de données simulé sont mitigées par la nature indépendante ou comparative de nos scores d'utilité.

Corrélation entre les QID et la mortalité ou entre QID Une corrélation peut être présente entre l'un ou l'autre des QID ou entre un ou plusieurs QID et la mortalité. Les facteurs de mortalité sont toutefois construits en tenant compte de la mortalité moyenne des paramètres susmentionnés. De plus, bien que nous puissions tenter d'améliorer progressivement notre « prédiction » de la mortalité selon les différents scores, une simulation parfaitement fidèle reviendrait à dire que nous pouvons parfaitement expliquer notre expérience : un rêve pour les assureurs !

Il est toutefois facile d'ajouter une corrélation simple si nous le souhaitons.

Uniformité des naissances et prises d'assurance Bien que l'uniformité de ces deux variables s'oppose en partie, nous avons une distribution de la population inexacte. Dans le cas des naissances, une diminution graduelle du taux de natalité (dans le monde occidental) aurait pu être représentée.

Toutefois, puisque nous souhaitons mesurer l'utilité d'un ensemble de données anonymisé relativement à son homologue sérialisé, l'hypothèse d'uniformité est à notre sens

suffisante.

Modélisation de l'antisélection basique Nous simulons les dates de décès et de terminaison de façon indépendante. Dans une population réelle, il est raisonnable de croire qu'un assuré dont la santé a diminué depuis sa sélection aura tendance à conserver sa police plus longuement. Nous ne pouvons pas simuler le comportement d'un assuré de façon réaliste tout en conservant une performance acceptable. Nous ignorons alors le problème complètement.

Omission du montant d'assurance Le plus souvent, les analyses de performance d'un portefeuille d'assurés portent sur le montant total payé en comparaison aux projections. La distribution des montants d'assurance selon les QID que nous avons présentement est un problème difficile et le gain serait marginal par rapport à l'utilisation des projections sur le nombre. Les deux analyses étant identiques sauf pour une répartition des poids différente, nous pouvons sans perte de généralité utiliser celle par nombre comme mandataire plutôt que celle par montant dans notre comparaison d'utilité.

3.6 Information propre à la problématique : assurance

En précisant notre domaine d'application lors de l'anonymisation des données, nous profitons de connaissance propre à ce domaine. Celle-ci peut nous guider dans nos choix d'opérations (suppression/généralisation) afin d'améliorer l'utilité de l'ensemble de données résultant.

Prenons en exemple l'IMC, rencontré en Section 3.5.2.9 : le domaine des données collectées est souvent entre 14 et 45[13]. En se référant à la Figure 3.10, on voit toutefois que la valeur numérique est souvent agglomérée dans un intervalle (dans ce cas, $[14,25[$, $[25,30[$, $[30,35[$, $[35,40[$, $[40,\infty)$. Il importe plus, dans une analyse, de savoir dans quel intervalle un individu se retrouvera.

[13] En assurance, on voit parfois des données au-delà de 50, mais à ce niveau-là, on agglomère souvent à 45+ ou 50+.

Cette classification disponible *a priori* fait partie de ce que nous appellerons des données *à seuil*. Posons la définition suivante :

Définition 3.6.1. Donnée à seuil : Soit x une donnée numérique dans le domaine $\mathbb{X}$. Soient les valeurs $x_1, x_2, \ldots, x_n$ des bornes connues *a priori*. x sera une donnée dite *à seuil* si nous pouvons interchanger x par i (ou un nom pour l'intervalle numéro i) où $x_i \leq x < x_{i+1}$ sans perte de prédiction pour nos analyses.

En pratique, les intervalles auront une définition textuelle ou numérique alternative. Dans notre exemple d'IMC, on peut voir les intervalles illustrés dans la table 3.VI

Tableau 3.VI – Intervalles pour l'IMC

Intervalle	Définition
$[14,25]$[14]	*Aucun*
$[25,30[$	Sur-poids
$[30,35[$	Obèse classe I
$[35,40[$	Obèse classe II
$[40,\infty$	Obèse classe III

Plusieurs modèles prédictifs fonctionnent bien lorsque les données sont regroupées en intervalles : un exemple appliqué à l'assurance est disponible en [57]. Afin d'avoir confiance en nos données, on tente traditionnellement de conserver un nombre minimal de sinistres dans chaque cellule utilisée pour l'analyse. L'opération d'anonymisation n'est donc plus vue comme une opération nuisant à l'analyse, mais plutôt comme une extension de la généralisation déjà faite pour obtenir des cellules suffisamment peuplées. Dans une optique d'optimisation d'utilité basée sur le domaine d'application, il est alors impensable de séparer ces deux étapes.

3.7 CLiKC : Un algorithme de *LKC*-anonymat contextualisé

À la lumière de la situation présentée, il nous apparait peu commode de reléguer l'opération d'anonymisation à un algorithme générique : nous avons préféré améliorer

[14] Ordinairement, l'intervalle pour l'IMC d'une personne en bonne santé se situe entre 18.5 et 25 [52]. Toutefois, comme le bris n'est pas fait sur la figure, nous avons gardé pour l'exemple que les intervalles illustrés.

un algorithme existant en le contextualisant selon le domaine d'application. Celui-ci guidera les hiérarchies de généralisations ainsi que les contraintes prétraitement à partir duquel nous ne pouvons plus considérer l'ensemble de données utile.

Nous nous rappelons (voir la Section 2.3.5) que l'algorithme classique du *LKC*-anonymat nécessite un score qui sert à guider les choix de généralisations. Nous étendrons le score proposé par [50], soit le score de discernabilité, en utilisant l'information supplémentaire à notre portée. Nous aurons également un avantage par rapport à l'algorithme original du fait que nos hiérarchies de classification seront ordinairement bien définies : leurs définitions sont disponibles en Section 3.8.

L'algorithme *CLiKC* (pour *Contextualised LKC-private algorithm*) utilise les différences par rapport au cas général pour améliorer les opérations et améliorer l'utilité résultante. Dans le système de PEPS (Figure 3.4), il s'insère dans le module *Anonymisation*. *CLiKC* s'inspire de l'algorithme *PAIS* [50], qui lui-même est un dérivé de l'algorithme menant au k-anonymat TDR [27].

L'idée générale est d'emprunter une approche *top down* (ou de haut en bas). Nous initialisons d'abord l'ensemble des attributs de chacun des tuples à la valeur la plus générale. Par la suite, tant qu'une spécialisation Cut_i (la procédure opposée d'une généralisation) est possible, nous sélectionnons celle qui est la plus prometteuse (voir la Section 3.7.2) puis appliquons la spécialisation à la table T. Une fois $\cup Cut_i$ épuisé ou que le score des spécialisations restantes n'atteint pas une cible M définie, nous retournons l'ensemble de données ainsi que le score atteint.

Algorithme 1 Contextualised LKC-Private Algorithm (CLiKC)

1: Initialiser toutes les valeurs en T à la valeur la plus générale
2: Initialiser Cut_i pour inclure la valeur la plus générale
3: Tant qu'un $x \in \cup Cut_i$ est valide
4: Trouver la spécialisation B de $\cup Cut_i$ qui possède le plus haut score combiné
5: Appliquer B sur T et mettre à jour $\cup Cut_i$
6: Mettre à jour $ScoreC(x)$ pour $x \in \cup Cut_i$
7: Fin Tant que
8: Retourner T et $\cup Cut_i$

Nous souhaitons nous prémunir principalement des deux attaques vues en Section 2.3.1. Puisque notre algorithme n'introduit pas de bruit, il est important de noter qu'un participant à une étude commune peut comparer ses propres résultats à ceux de l'ensemble différencié (dans lequel il a enlevé sa propre contribution). Nous verrons les impacts en Section 5.2.

3.7.1 Preuve du LKC-anonymat de *CLiKC*

Nous souhaitons démontrer que notre algorithme mène à un ensemble de données *LKC*-anonyme. Puisque notre algorithme est une spécialisation d'un existant, il nous suffit de prouver que nos spécialisations n'entraînent pas un non-respect des conditions d'anonymat.

Notre preuve est en quatre parties :

Énoncé 1 : Une spécialisation x n'augmente jamais le nombre de tuples pour un vecteur de QID qid_i donné

Preuve : Prenons l'exemple dégénéré où nous sélectionnons une spécialisation qui ne fait absolument rien, c'est-à-dire que tous les QID de chaque tuple restent exactement identiques. Dans ce cas, la transformation $i \rightarrow i$ ne change rien et le nombre de tuples pour chaque combinaison de QID reste identique pour toutes les combinaisons.

Une spécialisation effective (ayant un effet sur au moins une combinaison de QID)

va transformer un attribut A en deux attributs A_1 et A_2. Cette scission va mener à une séparation des ensembles de tuples possédant A dans leur vecteur QID. Si un ensemble a n tuples, cette spécialisation va séparer le groupe en deux sous-groupes de n_1 et n_2 tuples, $n_1 + n_2 = n$. Comme n_1 et $n_2 > 0$, alors $n_1 < n$ et $n_2 < n$. On peut voir que n est égal à $|T[qid_i]|$ pour les qid_i contenant A, ce qui conclut la preuve.

Énoncé 2 : La propriété de *LKC*-anonymat de *CLiKC* est monotone. Un ensemble de données T non-*LKC*-anonyme suite à la généralisation (étape 1) ne peut le devenir suite à l'algorithme et, inversement un ensemble de données *LKC*-anonyme suite à la généralisation le restera.

Preuve : La première partie de l'énoncé découle de la preuve de l'énoncé 1. Puisque nous ne pouvons pas augmenter $|T[qid_i]|$ pour n'importe quel tuple i, la définition 2.3.5 nous confirme que nous échouons la définition de *LKC*-anonymat.

La preuve de la seconde partie de l'énoncé est tributaire des conditions dans la boucle **Tant que**. Nous retenons les spécialisations valides uniquement, prévenant par le fait même de perdre le *LKC*-anonymat en sélectionnant l'une d'entre elles. Un ensemble de données *LKC*-anonyme suite à la première étape le reste donc jusqu'à la fin de l'algorithme.

Énoncé 3 : Un ensemble de données complètement généralisé de $\max(K, 1/C)$ tuples est *LKC*-anonyme

Preuve : Un ensemble de données complètement généralisé ne contient aucune information : il est équivalent à un ensemble de données avec 1 seul attribut et 1 valeur sensible, contenant $\max(K, 1/C)$ tuples avec la valeur $(*, *)$ à l'intérieur. $L = 1$ (car un seul attribut).

Nous respectons la première condition du *LKC*-anonymat, car nous avons au moins K tuples distincts pour toutes les combinaisons de QID (dans notre cas, $*$).

Nous respectons la deuxième condition, puisque $P(*|*) = \min(1/K, C) \leq C$.

Énoncé final : l'algorithme *CLiKC* retourne un ensemble de données *LKC*-anonyme (si celui-ci contient au moins $\max(K, 1/C)$ tuples).

Preuve : La preuve découle des énoncés 2 et 3. L'algorithme part toujours d'un ensemble complètement généralisé qui lui est *LKC*-anonyme. De par la monotonicité du *LKC*-anonymat de notre algorithme, la table garde cette propriété jusqu'à la sortie. Nous avons donc un ensemble de données *LKC*-anonyme garanti.

3.7.2 Score contextualisé

La pierre angulaire de l'algorithme *CLiKC* est le choix du score permettant d'orienter les prochaines spécialisations. Nous utiliserons une extension du score de discernabilité, rencontré en Section 2.4, qui implémente un système de bonus-malus représentant notre contextualisation connue. Le but d'un tel score est de permettre de profiter de l'information propre au domaine d'application tout en permettant de retomber élégamment sur un algorithme *LKC*-anonyme classique si nous n'avons (ou ne souhaitons) pas à s'en prévaloir.

Le score se calcule sur l'ensemble des n combinaisons de QID qid. Nous sélectionnerons à chaque tour de boucle de l'algorithme la spécialisation $A \rightarrow \vec{A}$ qui maximise :

$$ScoreC(v) = \sum_{qid} |T[qid]|^2 \times e^{b_{A \rightarrow \vec{A}}} \tag{3.3}$$

Nous appliquons le bonus ou le malus b uniquement aux vecteurs de QID qui furent impactés par la transformation : lorsque A n'est pas un attribut du vecteur, le bonus $b = 1$. Cela explique la notation $b_{A \rightarrow \vec{A}}$ et permet d'éviter qu'une spécialisation ayant un haut score pour peu de tuples soit nécessairement priorisée.

Le bonus b peut prendre une valeur sur l'ensemble des réels : un b positif augmentera le score d'une spécialisation tandis qu'une valeur négative diminuera celui-ci. Un bonus de 0 signifie une position neutre. En pratique, nous ne rencontrerons pas vraiment de b négatif dans une spécialisation que nous avons créée. Il est toutefois possible que, si nous prenons une taxonomie issue de l'industrie et que certaines spécialisations nous laissent circonspects, nous puissions les tempérer avec un bonus négatif (que nous

appellerons malus).

3.8 Hiérarchies de classification : table `donnees_assures`

Afin que notre algorithme fonctionne sans problème, nous avons besoin de définir nos hiérarchies de classification (ou taxonomies) pour notre table. Dans l'ensemble de données simulé en Section 3.5, nous avons les paramètres suivants qui peuvent être généralisés :

- les dates de naissance, de prise d'assurance et d'événement ;
- le sexe ;
- le statut fumeur ;
- le type de produit ;
- le score d'assuré ;
- le score de crédit ;
- l'IMC.

Nous allons proposer des hiérarchies types pour l'ensemble de ces paramètres et indiquer les bonus applicables. Dans les schémas présentés dans cette section, le nœud terminal de l'arbre, identifié par un astérisque, représente la valeur la plus générale. Advenant que nous n'effectuions aucune spécialisation sur cet attribut, cela équivaut à une *suppression* pure et dure. Les bonus données pour une spécialisation sont sur les branches du sous-arbre formé. L'absence de bonus équivaut à $b = 0$.

Notons que l'ensemble des hiérarchies proposées forme des arbres binaires. Cela nous permet de les afficher sous forme de graphiques plus facilement. Toutefois, nous pourrions définir plusieurs arbres pour un même attribut et évaluer toutes les spécialisations possibles lors d'un tour de boucle. Si, par exemple, l'intervalle $[a,b)$ est séparé en $[a,c),[c,b)$, alors nous éliminerons toutes les spécialisations de ce paramètre n'œuvrant pas sur l'un ou l'autre de ces intervalles.

3.8.1 Hiérarchisation des dates

Nous pouvons considérer les trois dates (naissance, prise d'assurance, décrément) de la même façon puisqu'elles diffèrent peu l'une de l'autre par leur nature et leur usage. Celles-ci constituent des QID de choix pour un attaquant puisque l'information sur la naissance et le décès d'une personne est facilement obtenable. La date de prise d'assurance peut être obtenable par différents moyens que nous n'exposerons pas ici.

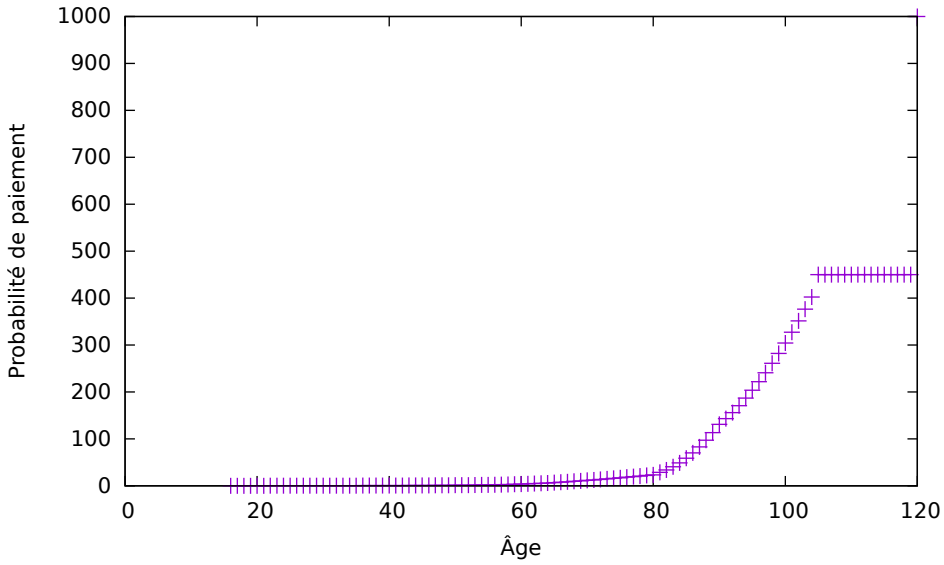


Figure 3.11 – Probabilités de décrément CIA9704 ultime : Femme, non-fumeur, âge atteint. La probabilité est multipliée par 1000.

En regardant une courbe de mortalité, nous pouvons voir que celle-ci varie peu pour les premières années avant de rapidement grimper. La sensibilité de la date de naissance et de la date de paiement augmentent *plus celles-ci sont distancées*. Comme nos sinistres sont uniquement entre le 1er janvier 2005 et le 1er janvier 2015, il nous est plus aisé de prioriser nos généralisations en privilégiant celles sur les assurés plus jeunes.

3.8.2 Hiérarchisation du sexe et du statut fumeur

Ces paramètres sont hautement prédictifs, si prédictifs en fait que nous avons des tables de décrément séparées pour chaque combinaison. Nous les illustrons à titre indi-

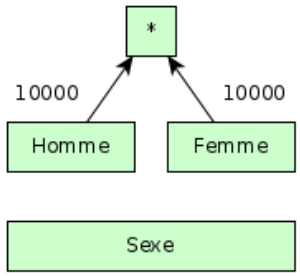


Figure 3.12 – Hiérarchie type pour le sexe

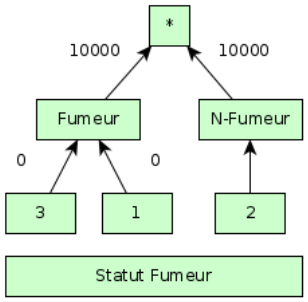


Figure 3.13 – Hiérarchie type pour le statut fumeur

catif, mais il va sans dire que ces généralisations sont uniquement à utiliser en dernier recours

3.8.3 Hiérarchisation de l'IMC

Nous avons hiérarchisé l'IMC comme sur la figure 3.10 en ajoutant deux bornes intermédiaires :

- Une à 18,5, le seuil de maigreur dit *anormal* ;
- Une à 33, qui est selon les manuels de tarification en assurance-vie le seuil à partir duquel les risques de maladie sont augmentés de façon significative.

En bonne compagnie d'assurance fictive, notre première spécialisation consiste en effet de séparer les IMC sur la borne de 33.

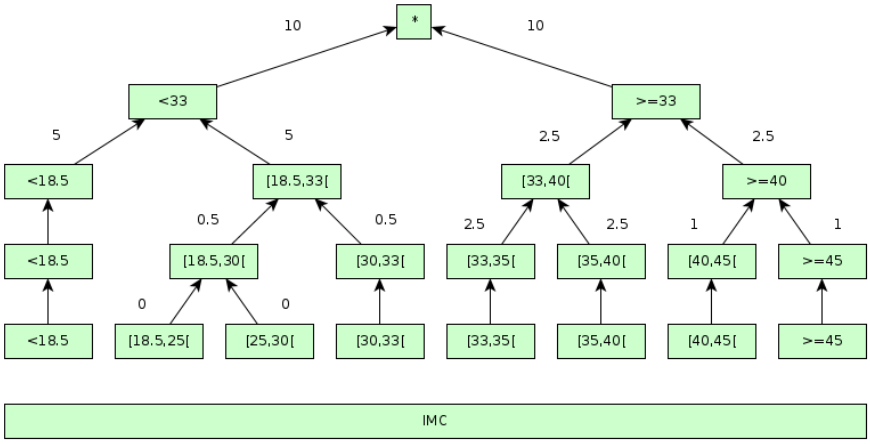


Figure 3.14 – Hiérarchie type pour l'IMC

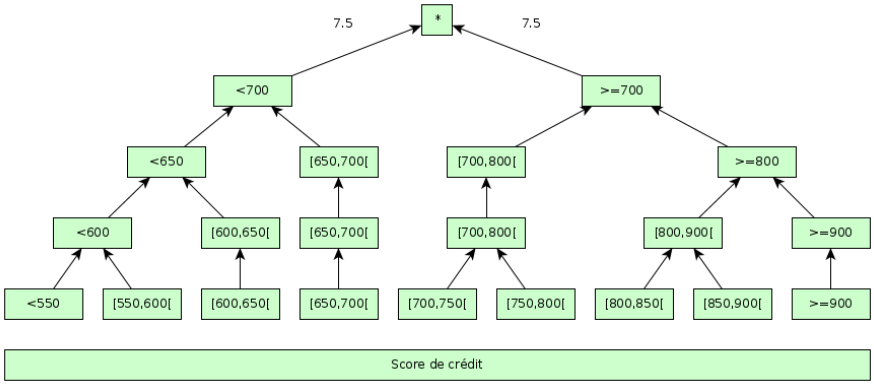


Figure 3.15 – Hiérarchie type pour le score de crédit

3.8.4 Hiérarchisation du score de crédit

Peu d'options nous sont possibles sur la hiérarchisation du score de crédit. Les fournisseurs de cet indice sont notoirement secrets sur la méthode employée et fournissent eux-mêmes les intervalles de risque. Nous nous sommes contentés de reproduire ceux-ci, en hiérarchisant sur les seuils plus généraux (700 est généralement le seuil pour les produits de crédit standards, 800 pour un excellent crédit, 650 pour un crédit pauvre, etc.)[15]

[15]Ces renseignements sont issus du site d'Equifax, bien qu'ils utilisent une terminologie plus enthousiaste (*fair, good, very good, great, excellent*. Le redouté *Poor* semble être réservé aux faillites et collections multiples.

3.9 Mesure de l'utilité

Que souhaitons-nous mesurer ? Cette question est au centre de notre système. Dans la plupart des cas, l'analyse souhaitée sera définie avant l'anonymisation, parfois même avant la collecte de données. Les métriques d'utilité peuvent donc tirer avantage de cette contextualisation.

3.9.1 Ratio ajusté sur perte de QID

La première mesure d'intérêt est le nombre de combinaisons différentes de QID que nous retrouvons dans l'ensemble anonymisé vs l'ensemble original. Nous souhaitons avoir un premier score remplissant les contraintes suivantes :

- celui-ci devrait être facile à comprendre et facile à calculer ;
- celui-ci doit tenir compte que plus le paramètre L est élevé, plus il est difficile de conserver des QID ;
- être entre 0 et 1, où 0 est la pire valeur et 1 la meilleure.

Nous proposons là *le ratio ajusté sur perte de QID*

Définition 3.9.1. Ratio ajusté sur perte de QID : Soit $|qid|$ le nombre de combinaisons distinctes de QID dans l'ensemble de données originales, et $|qid_E|$ le nombre dans l'ensemble de données anonymisé E. Le *ratio ajusté sur perte de QID* $R_A(E)$ est alors défini par la formule suivante :

$$R_A(E) = \left(\frac{|qid_E|}{|qid|} \right)^{1/L} \tag{3.4}$$

Nous ajustons le quotient entre le nombre de QID dans l'ensemble anonymisé et celui original par une puissance $1/L$. Empiriquement, le nombre de QID chute de façon dramatique en fonction de L. Puisque notre but est de comparer deux algorithmes n'ayant pas nécessairement ce même paramètre, nous avons proposé une correction qui semble graphiquement linéariser le score. Nous pouvons alors utiliser celui-ci pour déterminer rapidement la diversité des QID restants suite à une opération d'anonymisation. Il est

important de noter que ce score ne différencie pas une combinaison de QID utile d'une autre qui ne nous sert à rien. Au besoin, on pourra prendre uniquement les combinaisons de QID pour les champs qui nous sont critiques afin d'avoir un score reflétant mieux la réalité.

Exemple Posons $L = 5$, $|qid| = 25000$ et $|qid_E| = 1800$. Le score sera alors :

$$R_A(E) = \left(\frac{1800}{25000}\right)^{1/5} = 0.591$$

3.9.2 Amplitude sur champ numérique

Dans les grands ensembles de données, la distribution des valeurs finit par prendre des allures de courbe normale (par le théorème de la limite centrale). À cette fin, les intervalles de queue (ceux aux extrémités du domaine) des attributs numériques seront généralement les premiers généralisés.

Nous pouvons calculer ce que nous appelons l'*amplitude sur champ numérique*. Nous la définissons comme suit :

Définition 3.9.2. Amplitude sur champ numérique : Soit un champ numérique C généralisé en n intervalles $y_1, y_2, \ldots, y_n$. Prenons $\lceil y_1 \rceil$ la borne supérieure de l'intervalle contenant les plus petites valeurs et $\lfloor y_n \rfloor$ la borne inférieure de l'intervalle contenant les plus grandes. L'*amplitude sur le champ C* sera alors égal à

$$\lfloor y_n \rfloor - \lceil y_1 \rceil \tag{3.5}$$

Dans le cas d'un ensemble de données où des valeurs extrêmes sont présentes, cet indice sera particulièrement d'intérêt.

Exemple Prenons l'âge d'un assuré. Dans la base de données originale, nous avons les âges 18 à 96. Dans notre ensemble anonymisé, le plus petit et plus grand intervalle sont $[18, 25)$ et $[65, \infty)$. Nous pouvons comparer les deux amplitudes sur l'âge :

$$96 - 18 = 78 \text{ vs. } 65 - 25 = 40$$

Nous voyons ici que nous perdons beaucoup d'information sur les valeurs éloignées de la médiane.

3.9.3 Précision sur transformation

Le troisième et dernier score proposé permet de tirer avantage du fait que nous avons l'ensemble de données original en notre possession. Comme nous le verrons à la Section 4.2, nous devrons transformer notre ensemble de données afin de pouvoir calculer nos statistiques d'intérêt. Comme la majorité des statistiques calculées seront globales, nous ne souhaitons pas introduire une trop forte déviation lors de l'opération de généralisation-discrétisation[16]. Nous définissons la *précision sur transformation* comme :

Définition 3.9.3. Précision sur transformation : Soit une statistique X calculée sur les deux ensembles de données (X_O pour l'ensemble original et X_A pour l'ensemble anonymisé). La précision sur transformation est égale à :

$$P(X,A) = \left| \frac{X_A}{X_O} - 1 \right| \tag{3.6}$$

Exemple Prenons comme statistique X la moyenne d'âge chez les hommes fumeurs. Si $X_O = 45.2$ et que $X_A = 44.8$ alors nous aurons une variation de :

$$P(X,A) = \left| \frac{44.8}{45.2} - 1 \right| = 0.0088$$

Dans plusieurs cas, nous pourrons quantifier l'impact des variations. Un exemple sera montré en Section 4.3.

[16] D'où l'intérêt que la discrétisation soit contextualisée autant que possible.

Conclusion

Nous avons couvert les concepts nécessaires afin d'anonymiser un ensemble de données et d'évaluer sa performance lors d'une analyse prédéfinie. Le prochain chapitre est dédié à la mise en application et fournira des outils supplémentaires pour les embûches rencontrées.

CHAPITRE 4

EXPÉRIMENTATION

Nous avons un système prêt à être utilisé et des données qui ne demandent qu'à servir. Ce chapitre sera dédié à deux implémentations, basées sur deux situations possibles :

- à partir des données simulées, utilisation des taxonomies définies en Section 3.8 afin de démontrer le fonctionnement du système et la simulation de deux attaques différentes ;
- utilisation du système pour anonymiser adéquatement un partage public retrouvé sur le site de l'Institut Canadien des Actuaires.

4.1 Démonstration à partir des données simulées

Dans ce scénario, nous souhaitons éviter de divulguer directement le risque associé à l'IMC ou le score de crédit ainsi que le score de l'assuré. La date de prise d'assurance, de terminaison et de décrément sont considérés comme des attributs QID. Nous reprendrons les taxonomies discutées en Section 3.8 afin de *LKC*-anonymiser notre ensemble de données. L'étape de prétraitement est déjà quasiment complétée : nous avons l'ensemble des hiérarchies définies. Nous considérerons par simplicité n'avoir aucune règle[1]. Il nous reste uniquement de déterminer les paramètres L, K et C.

Paramètre L : taille du vecteur de la connaissance *a priori* Rappelons-nous que l'ensemble des champs, à l'exception du score d'assuré, peuvent être utilisés comme QID. Afin d'établir le paramètre L, il nous suffit d'identifier le plus long vecteur d'information disponible à l'attaquant potentiel, en éliminant les identificateurs (dans notre cas, `id`). Nous supposerons ici que le plus long vecteur est :

```
Date de naissance, Sexe, Statut Fumeur
```

[1] Il s'agit de notre premier partage !

donc $L = 3$. Nous avons volontairement pris un L plus petit que $|qid|$ pour démontrer la différence entre le *LKC*-anonymat et le k-anonymat, où $L = |qid|$.

Paramètre K : taille minimale des groupes de QID identiques Nous poserons $K = 20$: il s'agit d'un seuil suffisant pour démontrer l'algorithme.

Paramètre C : probabilité de réidentifier un individu étant donné qu'il est dans un groupe Nous choisirons $C = 1/4$. Il s'agit d'un bon début pour tester notre système.

Notons que les paramètres pourront être révisés au besoin suite à la boucle de rétroaction située dans le module d'équilibration. Notre but ici est de choisir des valeurs correctes, qui seront remises en question plus tard.

4.1.1 Premiers pas à travers l'algorithme *CLiKC*

À la lumière des hiérarchies proposées, il est évident que les deux premières spécialisations à faire sont selon le sexe et le statut fumeur. Dans les deux cas, nous avons mis b arbitrairement élevé puisque nous avons impérativement besoin de cette information. Calculons les scores respectifs.

$$\begin{aligned} |homme| &= 124242 \\ |femme| &= 125758 \\ ScoreC(v) &= (124242^2 + 125758^2) \times e^{10000} \\ &\approx 3,1 \times 10^{10} \times e^{10000} \end{aligned}$$

$$
\begin{aligned}
|non-fumeur| &= 212928 \\
|fumeur| &= 37072 \\
ScoreC(v) &= (212928^2 + 37072^2) \times e^{10000} \\
&\approx 4,6 \times 10^{10} \times e^{10000}
\end{aligned}
$$

Notons que dans tous les cas, la somme des différentes agrégations devra toujours donner 250000, soit le nombre de tuples. Dans notre cas, la première spécialisation sera sur le statut fumeur et la seconde sur le sexe. En nous assurant d'abord que les deux spécialisations appliquées de façon successive soient *valides* (*c.-à-d.* qu'elles ne rendent pas notre ensemble de données non-*LKC*-anonyme), nous finissons donc nos deux premières spécialisations avec un ensemble résumé à la figure 4.I

Tableau 4.I – Ensemble de données simulé après deux tours de boucle

Sexe	Statut fumeur	Autres paramètres	Nombre de tuples
Femme	Non-Fumeur	*	107056
Femme	Fumeur	*	18702
Homme	Non-Fumeur	*	105872
Homme	Fumeur	*	18370

Suite à ces deux spécialisations, nous avons un score d'environ $2,3 \times 10^{10}$. Il est normal d'avoir un score plus bas à chaque fin de tour de boucle : puisque nous prenons le carré de la taille de chacun des groupes, le score optimal est dans notre cas de 250000^2, soit le score du départ ! Rappelons toutefois que ce score doit être recalculé à chaque étape et qu'il ne sert qu'à guider notre prise de décision *pour cette étape uniquement*.

Maintenant que nous avons les principaux critères de classification, nous aurons besoin des dates.

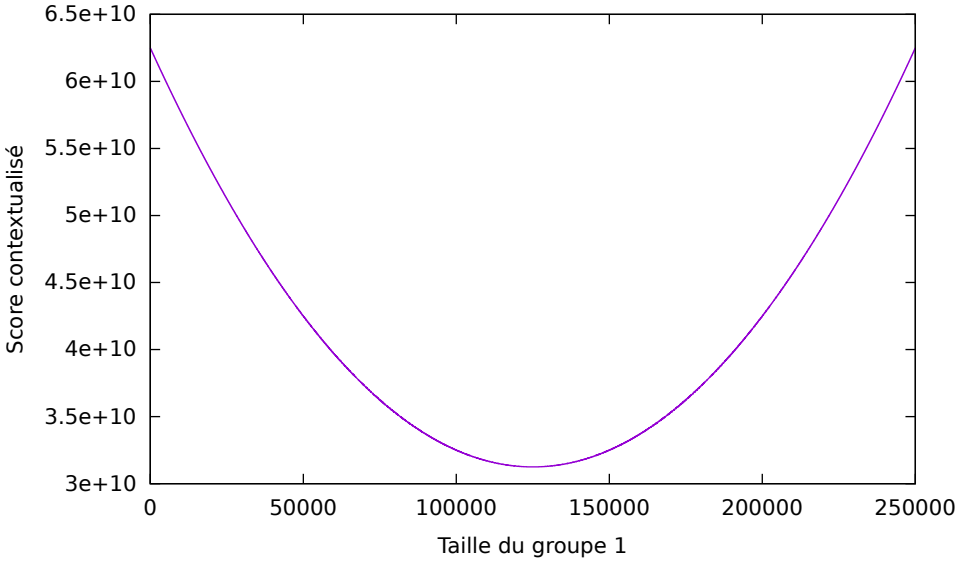


Figure 4.1 – Distribution du score contextualisé (sans bonus) selon la taille de l'un des groupes

4.1.2 Gestion des dates et données à forte spécialisation

Une des particularités de notre algorithme est la finesse et la facilité avec laquelle nous pouvons ajuster la granularité du partage des dates. En assurance, nous nous trouvons face à un dilemme : nous savons d'une part que la date de naissance est l'un des quasi-identificateurs les plus faciles à obtenir, mais avons besoin d'une bonne précision pour évaluer notre risque.

En regardant la figure type de la courbe de mortalité (voir la Section 3.8.1), la sensibilité est nettement différente selon l'âge de l'assuré. Nous pouvons donc tirer avantage de cette asymétrie pour améliorer notre utilité. Toutefois, il faut agir avec prudence : un ensemble de données d'assurance possède moins d'individus plus l'âge est élevé, à cause des paiements et terminaisons passées. En définissant une hiérarchie adaptée à nos données, nous pouvons tenter une spécialisation satisfaisante tout en respectant les promesses de notre algorithme.

Nous pouvons également prendre avantage de la connaissance propre à notre contexte pour déterminer comment les données seront utilisées. La date de naissance et de prise d'assurance fournissent deux paramètres d'intérêt : l'âge atteint lors de la prise d'as-

surance et le nombre d'années depuis celle-ci. Ce sont ces paramètres qui sont utilisés pour calculer la probabilité de paiement (statut DEC) et de terminaison (statut TER). Nous pouvons alors utiliser ces informations pour déterminer l'écart raisonnable des intervalles que nous souhaitons utiliser.

Tentons d'abord une spécialisation basée sur l'âge atteint lors du début de la collecte de données (2005-01-01). Avec les intervalles spécifiés dans le Tableau 4.II, nous obtenons trois classes avec un compte plus petit que K : (femme, fumeur, 1944-1945), (homme, fumeur, 1925-), (homme, fumeur, 1944-1945). Nous devrons donc ajuster notre taxonomie pour ces âges pour l'ensemble des sous-groupes.

Tableau 4.II – Première tentative d'intervalle pour les âges

Intervalle d'âge atteint	Équivalent en année de naissance
$[0,25[$	[1980, 2005]
$[25,30[$	[1975, 1979]
$[30,35[$	[1970, 1974]
$[35,40[$	[1965, 1969]
$[40,45[$	[1960, 1965]
$[45,50[$	[1955, 1960]
$[50,52[$	[1953, 1954]
$[52,54[$	[1951, 1952]
$[54,56[$	[1949, 1950]
...	...
$[80,120]$[2]	[1925 et -]

Un autre danger de procéder aussi rapidement avec les dates est le grand nombre de sous-groupes résultant. Les hiérarchies vues précédemment privilégient une segmentation en deux groupes à chaque fois. Nous pouvons toutefois émuler une division en n sous-groupes par $n-1$ divisions binaires.

Nous nous exposons également au risque d'obtenir une solution fortement sous-optimale en créant trop rapidement des groupes avec un peu plus de K individus. Comme nous devons conserver un minimum de K tuples dans chaque sous-ensemble de QID,

[2]Comme la prise d'assurance se termine à 65 ans, nous ne pouvons avoir de nouvelles personnes dans notre échantillon arrivant après cet âge. Nous faisons la première hypothèse qu'après 80 ans, les groupes formés seront trop petits pour garantir un anonymat.

nous devons nous assurer que tous nos groupes aient au moins $2K$ tuples avant de procéder. Nous pouvons inférer par la même occasion une propriété importante de la spécialisation :

Corolaire : Supposons que chaque opération de spécialisation sépare un groupe en au moins n sous-groupes, et soit g le plus petit groupe formé par les spécialisations précédentes. Nous pouvons faire au plus

$$\left\lfloor \log_n \left(\frac{|g|}{K} \right) \right\rfloor \tag{4.1}$$

opérations de spécialisation tout en conservant la propriété K du LKC-anonymat.

Notre score contextualisé est adapté à cette réalité. La spécialisation par date souffre également d'un malus interne : comme nous prenons une somme de carrés, un grand terme est plus avantageux qu'une somme de plus petits. Nous ignorerons donc plus longtemps les spécialisations possédant de nombreuses dimensions.

Ces raisons, ainsi que la forte contextualisation des dates, expliquent pourquoi nous avons considéré les dates comme une extension de notre algorithme plutôt que de les intégrer via une hiérarchie statique. Nous avons donc, une fois les paramètres principaux spécifiés, obtenu la spécialisation suivante pour les dates de naissance :

Tableau 4.III – Classification des dates de naissance

Champ
entre 1900 et 1939 inclusivement
entre 1940 et 1944 inclusivement
entre 1945 et 1949 inclusivement
entre 1950 et 1954 inclusivement
entre 1955 et 1959 inclusivement
entre 1960 et 1964 inclusivement
entre 1965 et 1969 inclusivement
entre 1970 et 1974 inclusivement
entre 1975 et 1979 inclusivement
entre 1980 et 2005 inclusivement

Nous ne pouvons malheureusement pas tirer avantage de la courbe de mortalité pour la distribution de notre spécialisation : nous n'arrivons pas à obtenir suffisamment de

tuples dans chacun des groupes en les sous-divisant selon l'espérance de décrément sous la table CIA9704. Nous discuterons plus en détail des problématiques et opportunités liées aux paramètres dates en Section 4.2.1.

Sortie de l'algorithme après les taxonomies statiques

L'algorithme *CLiKC* a obtenu un ensemble de données préliminairement anonyme en effectuant les spécialisations suivantes :

Tableau 4.IV – Champs de la table `DONNEES_ASSURES` *LKC*-anonymisée

Champ	**Traitement**
ID*	Supprimé
Date de naissance	Voir le Tableau 4.III
Date d'assurance	*Avant 1985*, sinon par année
Type de produit	TEMP(oraire) ou PERM(anent)
Sexe	Homme / Femme
Statut fumeur	Fumeur / Non-Fumeur
Score d'assuré	Standard quand 0%, Surprimé sinon
Score de crédit	Moins de 650, Entre 650 et 799, 800 et plus
IMC	Moins de 33, Entre 33 et 39.9, 40 et +
Date d'événement	Année de l'événement
Type d'événement	Tel quel

Nous avons profité de la modularité de notre algorithme pour faire une première spécialisation (sexe, statut fumeur, événement) dans un sous-ensemble de nos attributs. Ensuite, nous avons tenté différentes heuristiques pour les dates. Cette approche est selon nous *a priori*ser lors de la première utilisation du système ou pour un tout nouvel ensemble de données. En définissant clairement nos hiérarchies lorsqu'applicable et en utilisant la connaissance propre au domaine d'application, nous sommes désormais propriétaires d'un ensemble de données prêt à être partagé.

Mais est-il *utile* ?

4.2 Discrétisation des valeurs de l'ensemble anonymisé

Le but de cet ensemble de données est de constituer la base d'une étude d'expérience sur 10 ans. Comparer les principaux résultats par rapport à l'étude non anonymisée nous semble particulièrement à propos.

La première étape est de transformer ces données brutes en données d'expérience. Une *donnée d'expérience* est la représentation temporelle d'un tuple durant la durée active de la police. Prenons par exemple un individu souscrivant à un produit d'assurance le 3 janvier 2007 et terminant celle-ci le 7 février 2008 (à 23h59). Pour l'année d'expérience 2007, celui-ci aura contribué 363 jours (avec 0 décrément) et pour l'année 2008, 38 jours. La conversion de données sérialisées en données d'expérience permet d'avoir un attendu annuel des réclamations et ainsi de mesurer la performance du portefeuille d'assurés par rapport à nos projections. La construction complète d'une étude d'expérience est un sujet bien en dehors de notre champ, mais les bases sont présentées en [6] et [13]. Un exemple spécifique de données d'expérience est présenté en Section 4.5.1.

Pour la transformation des données sérialisées en données d'expérience, nous utiliserons un programme nommé *Archipel* [56] que j'ai développé lors de mon emploi en réassurance-vie[3]. Le programme possède de nombreux modules, mais nous utiliserons uniquement celui pour l'assurance-vie.

4.2.1 Granularité au plan des dates

Un des prérequis pour la création d'une étude d'expérience est d'avoir des dates granulaires sur le plan du jour. Or, notre algorithme nous propose une granularité sur le plan de l'année, voire sur une période de plusieurs années ! Nous devrons donc discrétiser les données que nous avons. Conformément à notre hypothèse rencontrée en Section 3.5.2.1, nous interpolerons linéairement entre les bornes fournies.

Cette approche ne fonctionne malheureusement pas pour les intervalles terminaux. Il est en effet bien rare qu'une distribution se termine de façon linéaire ! De plus, plusieurs

[3]Le droit d'auteur a malheureusement été cédé à mon employeur. Nous sommes en discussion pour publier certaines fonctionnalités sous licence libre.

intervalles seront définis de la forme $[x,\infty$ et il est bien peu pratique d'interpoler avec l'infini. Nous pouvons toutefois nous tourner vers une courbe exponentielle basée sur l'estimateur de Nelson-Aalen [1] [40]

Définition 4.2.1. Estimateur de Nelson-Aalen sur données modifiées : Pour un ensemble de taille n, soit $y_1 < y_2 < \cdots < y_k$ k valeurs distinctes apparaissant dans l'échantillon, où $k \leq n$. Posons alors s_j le nombre de fois que nous observons la valeur y_j dans l'échantillon, donc $\sum_{j=1}^{k} s_j = n$. Posons également l'**ensemble de risque**[4] $r_j = \sum_{i=j}^{k} s_i$ qui correspond aux observations plus grandes ou égales à y_j. En utilisant cette notation, l'estimateur de Nelson-Aalen pour la fonction de hasard $\hat{H}(x)$ est alors

$$\hat{H}(x) = \begin{cases} 0, & x < y_1, \\ \sum_{i=1}^{j-1} \frac{s_i}{r_i}, & y_{j-1} \leq x < y_j, j = 2, \ldots, k, \\ \sum_{i=1}^{k} \frac{s_i}{r_i}, & x \geq y_k. \end{cases} \tag{4.2}$$

L'estimateur de la fonction de survie est alors $\hat{S}(t) = e^{-\hat{H}(t)}$. Pour les valeurs de t où l'intervalle n'est pas borné, nous pouvons utiliser l'estimateur $\hat{S}(t) = \hat{S}(y_k)^{t/y_k}$.

Cet estimateur permet de régler la problématique des intervalles non bornés que nous avons créés lors de l'anonymisation de la table `DONNEES_ASSURES`. Nous ne discrétiserons ici que les intervalles de dates. En exemple, prenons les dates de naissance de notre ensemble anonymisé. Afin de rester réaliste, notons que nous plafonnerons l'âge maximal à 120 ans (la limite de notre table de décrément). On censure volontairement le dernier intervalle (65-105) pour obtenir la bonne fonction de survie.

[4] de l'anglais *risk set*

Tableau 4.V – Estimateur de Nelson-Åalen pour les années de naissance

Années	t	i	s_i	r_i	$\hat{H}(x)$
1940-1944	60-65	9	17056	17056	2.0793+1=3.0793
1945-1949	55-60	8	23939	40995	1.4954+0.5839=2.0793
1950-1954	50-55	7	26982	67977	1.0984+0.3969=1.4954
1955-1959	45-50	6	26945	94922	0.8146+0.2839=1.0984
1960-1964	40-45	5	26426	121348	0.5968+0.2178=0.8146
1965-1969	35-40	4	25749	147097	0.4218+0.175=0.5968
1970-1974	30-35	3	24876	171973	0.2771+0.1447=0.4218
1975-1979	25-30	2	22518	194491	0.1613+0.1158=0.2771
1980-2005	0-25	1[5]	37414	231905	0.1613

Notre estimateur pour l'intervalle non borné sera alors de $\hat{S}(t) = \hat{S}(y_9)^{t/65}$. Notre fonction de distribution $\hat{F}(t)$ (dont nous aurons besoin pour la simulation) est alors de

$$\hat{F}(t) = 1 - \hat{S}(t) = 1 - (e^{-3.0793})^{t/65} = 1 - 0.0460^{t/65} \tag{4.3}$$

Avec l'estimateur de Nelson-Åalen pour la fonction de hasard, nous obtenons donc une distribution complètement définie pour les années de naissance. Cependant, comme nous avons déjà un intervalle défini pour chaque tuple, nous interpolerons linéairement lorsque l'intervalle sera borné et garderons l'estimateur défini plus haut pour l'intervalle non borné. Nous procéderons de façon identique pour les années de prise d'assurance. Afin de rester cohérents avec la littérature en assurance et nos hypothèses précédentes, nous utiliserons l'hypothèse de distribution uniforme des décréments (DUD) pour répartir les événements sur les années d'observation.

[5]Comme notre intervalle non borné est à gauche, nous commençons le compte à l'intervalle le plus récent.

4.2.2 Granularité du score d'assuré

Comme celui-ci sert à la détermination de la probabilité finale de paiement, un score d'assuré discret est également nécessaire pour le bon fonctionnement du programme *Archipel*. Nous reprendrons la distribution discutée en Section 3.5.2.7 en tronquant la valeur 1.0 (que nous avons déjà identifiée dans notre ensemble anonymisé comme `Std`). Le choix d'utiliser une distribution exponentielle est particulièrement pratique puisque cette distribution est dite *sans mémoire* [40], c'est à dire que

$$P(X > s+t|X > s) = P(X > t), \forall s,t > 0 \tag{4.4}$$

Dans notre cas, nous avons que $P(X > 25\% + t|X > 25\%) = P(x > t)$. Nous n'avons alors qu'à simuler de la même façon qu'avec l'outil *Synthure*, en ajoutant 25% au résultat.

Avec ces outils, nous sommes en mesure de re discrétiser l'ensemble de données anonymisé et de pouvoir comparer les résultats avec celui original.

4.3 Mesure de l'utilité de l'ensemble anonymisé

Nous commencerons notre mesure de l'utilité par le calcul du ratio ajusté sur perte de QID (voir la Section 3.9.1). Afin d'éviter de polluer notre ratio, nous allons ignorer les champs utilisant des dates[6]. Nous avons $L = 3$, $|QID| = 249709$ et $|QID_E| = 773$.

$$R_A(E) = \left(\frac{773}{249709}\right)^{1/3} = 0.1457 \tag{4.5}$$

Le résultat n'est pas très satisfaisant. Rappelons-nous cependant que l'IMC et le score de crédit sont représentés dans l'ensemble de données original par des nombres à deux et une décimales, respectivement[7]. Nous avons toutefois défini en Sections 3.8.4 et 3.8.3 les taxonomies *idéales* pour notre analyse. Puisqu'il s'agit de données à seuil, nous pouvons utiliser la généralisation la plus spécifique (le dernier étage sur les figures 3.14 et 3.15) pour calculer notre ratio. Le $|QID|$ ajusté est alors de 15 624 pour un

[6] $365 \times 365 \times 365$ et notre ratio est ruiné.

[7] Le score d'assuré prend uniquement des multiples de 25% et n'est pas affecté

score normalisé de :

$$R_A(E) = \left(\frac{773}{15624}\right)^{1/3} = 0.3671 \tag{4.6}$$

Le score, bien que nettement meilleur, illustre bien la perte dans la diversité des QID. Nous pourrions difficilement réduire le paramètre L : à 3, le descendre plus bas serait critiquable sur le plan de notre perception des attaques possibles. Un adversaire incapable d'obtenir au moins 3 pièces d'information dans les QID de la table ne serait pas très dangereux ! Le paramètre K, présentement à 20, pourrait être descendu à 15 ($R_A(E) = 0.4127$) ou même 10^8 ($R_A(E) = 0.4458$) si nous souhaitions avoir une utilité selon ce score plus grande. Toutefois, nous ne pouvons pas y échapper : puisque nos données sont concentrées auprès des valeurs centrales (prise d'assurance en milieu de vie, probabilité de paiement plus élevée à partir de 65 ans), une perte d'information est pratiquement inévitable pour les valeurs de queue.

Pour l'amplitude sur champ numérique, nous allons observer que les indices pour l'IMC, le score de crédit et d'assuré. Le tableau 4.VI illustre nos résultats. Notons que nous prendrons encore une fois les intervalles idéaux pour notre ensemble de données original.

Tableau 4.VI – Mesure de l'amplitude sur champ numérique pour trois QID

QID	$\lceil y_1 \rceil$	$\lfloor y_n \rfloor$	Amplitude	Ratio sur Orig.
IMC (Orig.)	18.5	45	26.5	100%
IMC (Anon.)	33	40	7	26.4%
Score de crédit (Orig.)	550	900	350	100%
Score de crédit (Anon.)	650	800	150	42.9%
Score d'assuré (Orig.)	0%	400%	400%	100%
Score d'assuré (Anon.)	0%	25%	25%	6.3%

On voit que la répartition plus normale de l'IMC et du Score de crédit font que l'amplitude est moins taxée par l'anonymisation. Le score d'assuré est littéralement réduit à

[8]Notons toutefois qu'un $K = 20$ nous apparaissait plutôt agressif même en considérant la nature de nos données. La théorie en Section 2.5 nous rappelle qu'il existe une asymétrie des risques perçus lorsque nous ne sommes pas l'individu à risque.

sa plus simple expression (100% ou *autre*). Le calcul de l'amplitude ne donne cependant pas d'indication sur le choix des intervalles : il est tout à fait possible que le pouvoir prédictif du score d'assuré soit préservé par cette généralisation[9].

Finalement, nous prendrons les statistiques suivantes afin de mesurer la précision sur transformation de notre ensemble de données. La notation est non-canonique et uniquement pour garder le tableau des résultats (Tableau 4.VII) plus léger.

- La moyenne d'âge des hommes $(\bar{x_H})$
- La moyenne d'âge des non-fumeurs $(\bar{x_{NS}})$
- La variance sur l'année de police des fumeurs $(Var(d_F))$
- Le 40e centile des terminaisons pour l'année 2008 $P_{0.4}(2008)$
- L'exposition totale pour l'année 2007 $Expo_{2007}$

Tableau 4.VII – Mesure de la précision sur transformation pour 5 statistiques sommaires

Statistique	X_0	X_A	$P(X,A)$
$(\bar{x_H})$	40,10	39,85	0,62%
$(\bar{x_{NS}})$	40,18	40,04	0.35%
$(Var(d_F))$	1,69	1,66	1,80%
$P_{0.4}(2008)$	2008-05-22 (142 jrs.)	2008-05-26 (146 jrs.)	2.81%
$Expo_{2007}$	127065	127106	0,03%

Nous obtenons ici des résultats enthousiasmants, mais peu surprenants. Les méthodes d'anonymisation issues de la famille du k-anonymat brillent particulièrement bien sur ce genre de requêtes. De plus, des méthodes similaires sont employées avec succès en se basant sur la théorie de la crédibilité. L'algorithme *CLiKC*, de même que les algorithmes de *LKC*-anonymat semblent être particulièrement robustes lorsque nous avons besoin de statistiques sommaires.

[9]En pratique, comme les assurés sur-primés sont rares, nous les regroupons de toute façon pour avoir un nombre suffisant dans chaque groupe afin de rester crédible [41]

4.4 Boucle de rétroaction et sortie des résultats

Pour le besoin de la démonstration, nous jugerons que la première itération présente un profil utilité-confidentialité satisfaisant. Si toutefois nous souhaitions ajuster la paramétrisation passée au module d'anonymisation — l'équivalent de faire une boucle de rétroaction — nous pourrions procéder comme suit.

4.4.1 Réajustement du paramètre K

Le paramètre K a l'impact le plus clair sur l'ensemble de données résultant et les premiers ajustements seront certainement sur ce paramètre. Un paramètre plus élevé réduira la diversité des QID de façon dramatique et il convient de le garder à une petite fraction du nombre de tuples (appelé ici n).

N'importe quel algorithme issu de la famille du k-anonymat aura tendance à agglomérer rapidement les valeurs de queue[10], spécialement si la distribution de l'attribut suit une courbe normale[11]. Des mesures présentées plus bas peuvent tenter d'empêcher cette généralisation "triangulaire" (où les valeurs centrales sont plus granulaires que celles de queue) avec différents niveaux de succès.

Nous avons discuté en Section 2.3.5 que lorsque le nombre de QID croît, un ensemble de données k-anonyme devient rapidement inutile : la malédiction de la dimensionnalité frappe rapidement. Cependant, baisser K sans retenue n'est pas la solution : fort heureusement, nous avons d'autres stratégies de mitigation dans notre sac.

4.4.2 Réajustement du paramètre L

Le paramètre L, soit la taille des vecteurs de connaissance préalable de l'adversaire, permet de s'échapper — ne serait-ce que temporairement — de la malédiction liée à la forte dimensionnalité des QID d'un ensemble de données. Il ne faut cependant pas succomber à l'appel de réduire trop violemment ce paramètre : rappelons-nous qu'une fois un partage effectué, celui-ci est permanent, tandis que les connaissances de l'adversaire

[10] Il s'agit des valeurs les plus petites et les plus grandes d'une distribution sur un attribut donné.
[11] Lorsque n est grand, ceci est une hypothèse raisonnable à faire.

augmenteront très certainement dans le futur. Nous suggérons de prendre L égal au plus long vecteur de connaissance présumé de l'adversaire. Si, dans le pire cas, L est aussi long que l'ensemble des QID, nous perdons cet avantage par rapport au k-anonymat classique. Dans le k-anonymat, en effet, l'adversaire est réputé avoir une connaissance égale à l'ensemble des QID présents dans la table pour un individu donné (voir la Section 2.3.2).

4.4.3 Réajustement du paramètre C

Le paramètre C nécessite que nous approchions le problème un peu différemment. Celui-ci mesure, *une fois les données généralisées puis respécialisées*, la probabilité que nous puissions correctement identifier un attribut sensible pour un individu donné. Celle-ci est tributaire non seulement de la diversité des attributs sensibles, mais également de leur distribution. Nos exemples furent plutôt cléments en ce qui a trait à ce paramètre, mais en multipliant les attributs sensibles, nous pouvons rapidement nous trouver contraints de spécialiser plus prudemment.

Un élément souvent oublié lors de la détermination de ce paramètre est que *celui-ci s'applique sur tous les attributs sensibles, sans discernement*. En exemple, supposons que nous avons 8 tuples dans notre sous-groupe spécialisé (7 attributs sensibles A, 1 attribut sensible B) et que l'attribut sensible B est critique à nos yeux. Un observateur serait tenté de déclarer que la probabilité d'associer un attribut sensible à un tuple de ce sous-groupe est de $1/8$. Cependant, nous devons considérer l'attribut A également et sa probabilité de $7/8$, qui est nettement plus contraignante sur le choix de notre paramètre C. Une solution serait de supprimer ce paramètre ou de l'invalider (par exemple en mettant "Autre").

4.4.4 Réajustement des hiérarchies et des bonus

Finalement, un élément décisif d'un partage couronné de succès est le choix de nos hiérarchies et l'ordre dans lequel nous souhaitons les appliquer, modélisé par le choix de nos paramètres b. Celles-ci n'ont pas d'impact sur l'anonymat de l'ensemble de données

(un ensemble *LKC*-anonyme l'est ou ne l'est pas). Dans nos deux exemples, étant donné que les tables de décrément dépendent substantiellement du sexe et du statut fumeur, nous avons littéralement imposé ces bris en premier lieu. Un autre type d'étude pourrait privilégier d'autres bris (l'âge, l'IMC, etc.) afin d'obtenir un ensemble plus approprié pour l'information qu'il souhaite partager.

Une approche alternative, bien que tout aussi valable, est d'accorder un bonus de 0 à toutes les spécialisations et d'affubler explicitement un malus ($b < 0$) aux spécialisations qui ne nous intéressent pas. Un b légèrement au-dessous de zéro gardera la spécialisation comme possible, bien que handicapée par rapport aux autres, tandis qu'un b fortement négatif éliminera cette spécialisation des candidats.

Contrairement au choix des paramètres L, K et C, il nous est difficile de déterminer à l'avance l'impact que la modification des taxonomies peut avoir sur l'ensemble résultant. La bonne connaissance du contexte, de l'ensemble de données ainsi qu'une touche de créativité sera définitivement utile pour naviguer à travers les nombreux choix. N'oublions pas que, bien souvent, les principales hiérarchies et bonus seront prescrits dans le but explicite de faire sortir certaines informations spécifiques.

Nous passerons à un second exemple, issu de données partagées publiquement. Les enjeux et attributs sont similaires, mais la structure de l'ensemble de données est différente.

4.5 Données publiques : Étude de cas sur l'étude de l'ICA

Comme nous l'avons mentionné quelques fois auparavant, les données publiques en assurance sont très rares. De plus, il est évidemment impossible d'obtenir un ensemble de données de qualité (nombreux paramètres QID discrets et nombreuses valeurs sensibles) : le contraire serait inquiétant !

Nous pouvons toutefois nous rabattre sur les études annuelles de décrément publiées par l'Institut Canadien des Actuaires. Celle que nous utiliserons est intitulée *Risques normaux grande branche au Canada 2012-2013 à l'aide des tables 97-04* et l'ensemble

de données est disponible publiquement sur Internet[12]. Comme l'ensemble de données montré en Section 1.3, cet ensemble de données est 1-anonymisé[13] et, si l'information sensible présente nous intéresse, nous pouvons facilement lancer une des attaques définies en Section 3.4.

4.5.1 Présentation de l'ensemble de données

Puisque les champs sont similaires à ceux de notre ensemble simulé, nous ne présenterons que sommairement l'ensemble de données. Les paramètres sont présentés dans le Tableau 4.VIII.

Tableau 4.VIII – Description des paramètres de la table IndLifeMDB.1213.v2

Champ	Type	Définition
Year	Numérique	Année d'observation (2013)
Sex	Numérique	Sexe de l'assuré-e
Smoker	Numérique	Statut fumeur
TypeofIns	Numérique	Produit assuré
FaceSize	Numérique	Bande de taux[14]
PreferredClass	Numérique	Classe préférentielle (analogue au score)
DbDuration	Numérique	Durée de police
DbIssueAge	Numérique	Âge à l'émission
NExposed	Numérique	Nombre d'années-exposition
AExposed	Numérique	Montant exposé
NDeaths	Numérique	Nombre de réclamations
AClaims	Numérique	Montant total réclamé
8692NExpDeaths	Numérique	Attendu du nombre de réclamations (table CIA8692)
8692AExpClaims	Numérique	Attendu du montant de réclamations (table CIA8692)
9704NExpDeaths	Numérique	Attendu du nombre de réclamations (table CIA9704)
9704AExpClaims	Numérique	Attendu du nombre de réclamations (table CIA9704)

Nous pouvons voir que nous avons affaire à un ensemble de données d'expérience. Ordinairement, sur une étude d'expérience, il est plus compliqué d'assurer un k/LKC-

[12] `http://www.cia-ica.ca/fr/publications/détails-de-publication/215062`

[13] Comme nous l'avons vu en Section 2.3.5, cela est environ équivalent au *Confidence Bounding*.

[14] Les taux de prime dépendent souvent du montant assuré. Cet identifiant permet d'identifier l'ajustement dû à ce montant.

anonymat puisque chaque individu peut se retrouver plusieurs fois (avec les même QID) dans l'ensemble de données. Les garanties du k-anonymat et de ses extensions dépendent de la prémisse que chaque tuple représente un individu différent. Dans notre cas, puisque l'étude dure uniquement une période d'un an, nous supposerons que chaque individu ne possède qu'une seule police et donc ne se retrouve qu'une fois maximum dans la table.

L'information sensible n'est pas légion dans notre ensemble de données. Le seul élément apprenable via une attaque est le montant assuré pour une police. Nous tenterons donc de préserver cette information hors d'un attaquant.

Les 8 premiers champs, à l'exception de FaceSize, sont des quasi-identificateurs, mais nous pouvons déjà ignorer le champ Year, toujours égal à 2013. Nous poserons $L = 3$, $K = 50$ et $C = 10$. Puisque les données sont déjà semi-agrégées, nous utiliserons la somme du champ NExposed pour déterminer le respect du paramètre K.

Attaque contextuelle sur montant assuré Nous avons été en mesure d'identifier une attaque propre à cet ensemble de données en utilisant la bande de taux. Le Tableau 4.IX, issu du document d'étude [34], donne les intervalles de montant assuré pour chacune des bandes. Le niveau de granularité est le dollar.

Tableau 4.IX – Bandes de taux selon le montant assuré

Bande	Montant
1	< 10 000 $
2	10 000 $ à 49 999 $
3	50 000 $ à 99 999 $
4	100 000 $ à 249 999 $
5	250 000 $ à 499 999 $
6	500 000 $ à 999 999 $
7	1 000 000 et plus

Prenons un exemple pour la bande 4, soit de 100 000$ à 249 999$. Si nous avons 50 assurés (la condition sur K est respectée), mais que le montant assuré total est de 5 000 000$, nous pouvons clairement voir que tous les 50 assurés de cette bande le sont pour 100 000$ (le montant minimal 100 000$ $\times$ 50 = 5 000 000$). Nous avons donc inféré avec succès le montant assuré de *tous les assurés de cette bande*.

En supposant que nous souhaitons protéger le montant assuré par chaque individu, il est complètement saugrenu de fournir un QID tel que la bande de taux ! Dans plusieurs cas, l'attaquant sera parfaitement satisfait de cette information et notre partage sera aussi anonyme qu'un partage sérialisé.[15]. Nous commencerons donc par supprimer cet attribut de notre ensemble de données. Ceci est un exemple flagrant de l'importance de bien contextualiser notre partage confidentiel. Les promesses d'anonymat tombent bien rapidement lorsque nous ne caractérisons pas promptement nos données.

4.5.2 *LKC*-anonymat et utilité sommaire

La première étape est de définir les hiérarchies. Nous ne détaillerons pas nous choix et utiliserons celles issues de notre interprétation de chacun des champs détaillés dans l'annexe du document d'étude cité au début de cette section.

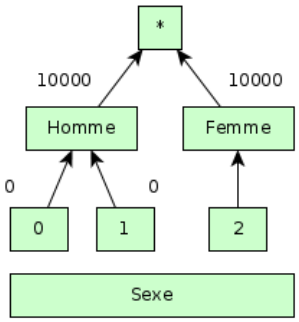


Figure 4.2 – Hiérarchie type pour le sexe

Notons la forte réduction de dimensionnalité pour la durée (voir la Figure 4.6). Nous partons des entiers de 1 à 101 pour obtenir uniquement deux valeurs (avec 16 comme valeur de bris). La contextualisation de l'étude publiée nous confirme que ce bris est le plus important pour nous et que la division quinquennale est de second ordre. Ceci est un exemple de valeur à seuil que nous n'aurions pas nécessairement identifié sans bien connaître le domaine d'application. Notons également l'absence de hiérarchie pour l'âge : les bris les plus granulaires sont identifiés ci-bas. Cependant, un peu à l'instar des

[15]Nous espérons que les auteurs de l'étude ne considéraient pas le montant d'assurance comme sensible !

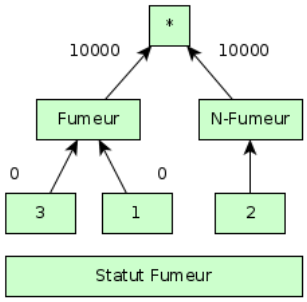


Figure 4.3 – Hiérarchie type pour le statut fumeur

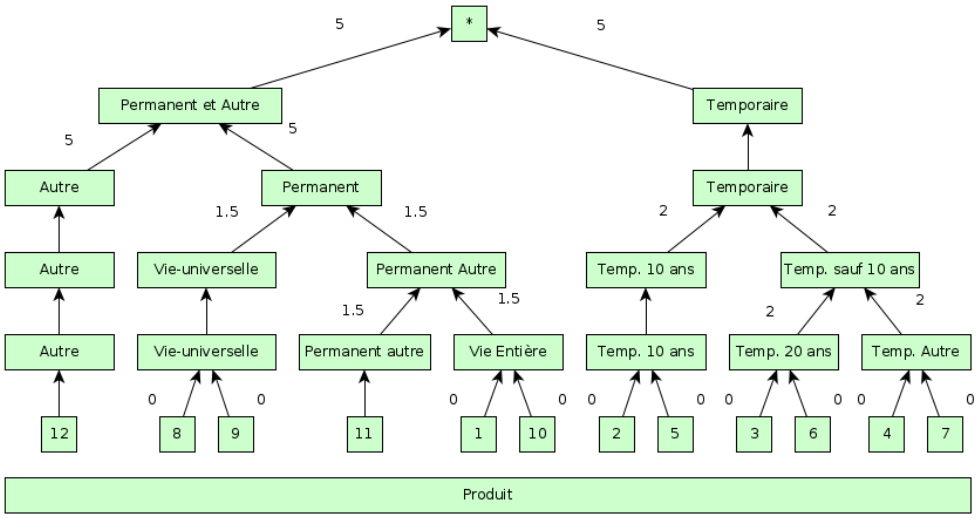


Figure 4.4 – Hiérarchie type pour le produit

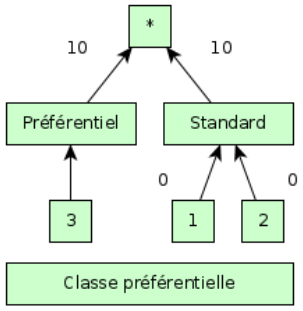


Figure 4.5 – Hiérarchie type pour la classe préférentielle

dates, nous souhaitons avoir le plus de précision possible et n'avons pas de hiérarchie préférée. Nous utiliserons alors une approche similaire aux dates en section 3.8.1 pour

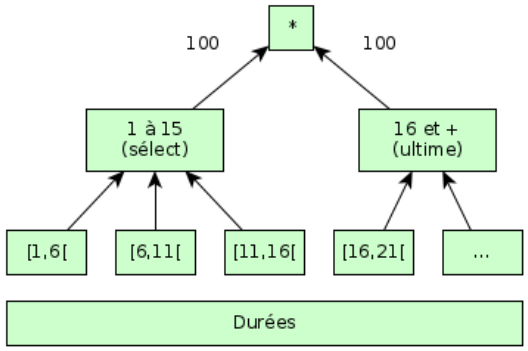


Figure 4.6 – Hiérarchie type pour la durée

les âges.

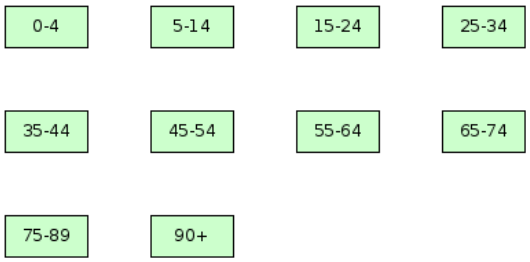


Figure 4.7 – Groupements souhaités pour l'âge

Les résultats de l'algorithme sont présentés dans le tableau 4.X.

Tableau 4.X – Résultats de l'algorithme *CLiKC* sur la table IndLifeMDB

Champ	Valeurs possibles
Sexe	Homme / Femme
Statut fumeur	Fumeur / Non-Fumeur
Produit	PERM(anent) / TEMP(oraire)
Age	0-34, 35-44, 45-54, 55+

L'algorithme arrive avec des bornes similaires pour le sexe, le statut fumeur ainsi que le produit. Cela est sans surprise étant donné le bonus important dans nos taxonomies. Le niveau de granularité de l'âge est insatisfaisant : on voit clairement que la majorité des polices sont vendues entre 35 et 55 ans. Nous pouvons toutefois reproduire la majorité

des tableaux publiés dans le document d'étude[16], à l'exception de ceux demandant le statut préférentiel ou le l'absence de classement fumeur.

Sur le plan de l'utilité, nous avons 32 combinaisons de QID différentes. En ne conservant que les tuples possédant une valeur > 0.0 en NExposed, nous en avions 20979. Notre ratio de discernabilité primitif est alors de 0,103, un score reflétant la perte de nombreux QID de notre ensemble de données anonymisé. En prenant les taxonomies idéales, nous obtenons 14160 QID, pour un ratio ajusté de 0,131. Dans les tableaux que nous pouvons reproduire, nous avons bien entendu une précision sur transformation de 100% : contrairement au premier cas, nous n'avons appliqué aucune transformation à nos données.

Nous avons dû faire preuve d'imagination pour appliquer notre système à un ensemble de données réel. Toutefois, cela nous a permis de voir quelques lacunes que nous discuterons dans le chapitre suivant.

[16]Sans toutefois être en mesure de reproduire ceux de l'annexe Excel retrouvée sur la page web du document. La niveau de granularité pour l'âge est insuffisant.

CHAPITRE 5

DISCUSSION

Le système *PEPS*, en plus des exemples précédents, a été employé dans le partage de données avec différents acteurs en assurance. Au fil de nos expériences, nous avons été en mesure de relever quelques difficultés et opportunités que nous présenterons ici.

Nous commencerons par discuter de la différence d'approche entre une divulgation discrète (un nombre fini de récipiendaires) et générale (publiquement, le plus souvent sur internet). Nous évoquerons également les problématiques liées à un partage de type *mise en commun* où un contributeur potentiellement malicieux contribue la majorité des données.

Nous proposerons ensuite un concept de *hiérarchies généralisées*, où nous pouvons émuler des hiérarchies plus complexes via d'autres binaires. Nous ferons alors un tour sur les difficultés rencontrées avec certains ensembles de données problématiques, avant de traiter du risque de partages répétés. Nous tracerons un parallèle entre le concept de généralisation-discrétisation et l'introduction de bruit, avant de voir si notre concept peut s'appliquer ailleurs. Nous conclurons ce chapitre par le résultat que personne ne souhaite obtenir, le partage *impossible*.

5.1 Divulgation discrète vs. générale

Lors des démonstrations du Chapitre 4, nous avons peu traité des différences entre une divulgation à récipiendaires discrets à une divulgation générale. Celle-ci à pourtant un impact important sur les choix des paramètres[1]. Nous ferons la comparaison entre le partage discret — le cas de base que nous avons considéré depuis le début de ce mémoire — et celui général.

Dans le cas d'un partage général, les pressions pour que L tendent vers K sont très fortes : nous pouvons difficilement justifier *connaître l'information disponible à l'en-*

[1] Elle a cependant moins d'impact sur les taxonomies, puisque l'utilité n'est pas modulable selon la définition de celles-ci. Voir la section 4.4.4.

semble des récipiendaires potentiels. Le choix de ce paramètre doit être plutôt fait en fonction de l'information sur les QID qui est propre à nous. Pensons à l'utilisation de classificateurs internes (*i.e.* Groupe A, B, C, D, E, F, sans fournir d'information permettant de recréer les groupes). De façon générale, un ensemble de données partagé publiquement sera nettement plus généralisé qu'un autre partage discrètement, justement par une moins grande différence entre le paramètre L et K.

5.2 Mise en commun avec contributeur dominant

Considérons désormais le scénario de partage via *mise en commun* : dans celui-ci, les individus envoient les données sérialisées à un tiers de confiance, qui procédera à une analyse puis une anonymisation subséquente avant de repartager les données. Dans notre cas, le récipiendaire importe peu.

Rappelons que, dans notre scénario, le récipiendaire est également un attaquant potentiel. Un attaquant pourrait, en acceptant de saboter ses propres individus, lancer une *attaque différentielle* (à ne pas confondre avec l'intimité différentielle, vue à la Section 2.3.6). Cette attaque consiste simplement à prendre l'ensemble de données anonymisé, supprimer ses tuples et publier l'ensemble différentié.

Cette attaque est à double tranchant puisque nous obtenons deux ensembles aux garanties de *LKC*-anonymat brisées : l'ensemble différencié et l'ensemble supprimé[2]. Cela est pourquoi nous disons que cette attaque constitue un sabotage. Les raisons motivant ce genre de comportement sont au-delà de notre compréhension : notons toutefois qu'un contributeur particulièrement malicieux pourrait simuler des données au lieu de partager les siennes, supprimant son risque.

Cette attaque est particulièrement efficace lorsque nous avons un contributeur dominant. Dans le domaine de l'assurance, c'est une réalité dans toutes les mises en commun [36] : en assurance-vie, les trois plus gros assureurs (Manuvie, SunLife et Great-West) se partagent plus de 60% du marché. En reprenant les données utilisées en Section 4.5, nous pouvons voir la répartition des contributeurs en Figure 5.1.

[2]Nous pouvons faire la différence entre les données anonymisées et l'ensemble différencié pour obtenir les données supprimées par l'attaquant.

Société	Comprend	Contribution	
		2011-2012	2012-2013
Desjardins Laurentienne Vie	Imperial Vie; Laurier	3,12 %	3,07 %
Équitable		0,00 %	3,76 %
Great-West Life	London Life; Canada-Vie	24,53 %	23,80 %
Industrielle Alliance		11,75 %	11,50 %
Manuvie	La Maritime	23,48 %	21,84 %
Banque Royale du Canada		6,28 %	6,64 %
Sun Life		19,58 %	18,84 %
Transamerica		11,26 %	10,55 %
Exposition totale		100,00 %	100,00 %

Figure 5.1 – Contributeurs à l'étude de l'ICA utilisée

Comment se prémunir d'un potentiel contributeur dominant ? Si nous avons suffisamment de données, nous pouvons limiter l'impact de celui-ci en anonymisant ses données séparément de celles des autres. De cette façon, les garanties de *LKC*-anonymat seront préservées même s'il décide d'attaquer. Il nous est également possible d'anonymiser l'ensemble de données comme à l'habitude et ensuite de tester l'impact de ce genre d'attaque pour tous les contributeurs. Ce genre de test entrainera une baisse du nombre de spécialisations possibles et limitera par le fait même l'utilité.

Si le mal est déjà fait, rappelons que *PEPS* gère efficament les attaques, peu importe leur nature. Il suffira d'ajouter les règles nécessaires pour les tuples touchées et de s'engager un excellent avocat...

5.3 Hiérarchies généralisées

Que faire si nous voulons absolument séparer les valeurs attribut en trois ? En quatre ? En n ? Nous avons depuis le début considéré que les spécialisations binaires du style $A \rightarrow A_1, A_2$. Voyons maintenant comment émuler une spécialisation en n champs avec $n-1$ spécialisations binaires.

Émulation d'une spécialisation d'un groupe en n sous-groupes Soit b^* le bonus que nous souhaitons appliquer à la spécialisation en n sous-groupes. Pour l'émuler, il suffit de sélectionner une première spécialisation que nous insérerons dans notre taxonomie avec $b = b^*$. Par la suite, spécialisons les deux feuilles de l'arbre généré avec $b = \infty$ récursivement jusqu'à obtenir les intervalles souhaités.

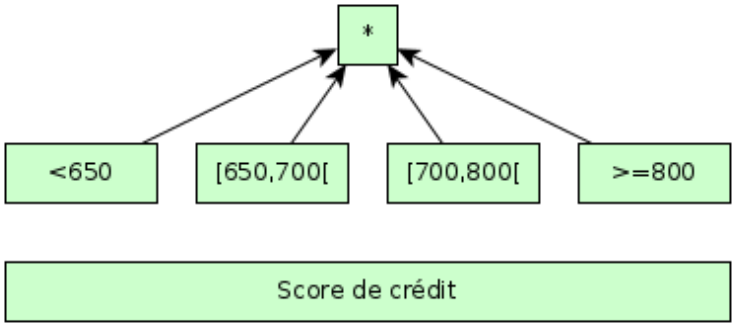


Figure 5.2 – Spécialisation du score de crédit souhaitée

Prenons en exemple le score de crédit. Supposons que nous souhaitons sous-diviser en 4 sous groupes. La spécialisation souhaitée est illustréé en Figure 5.2. Nous pouvons la reproduire avec 3 spécialisations binaires, tel qu'illustré à la Figure 5.3.

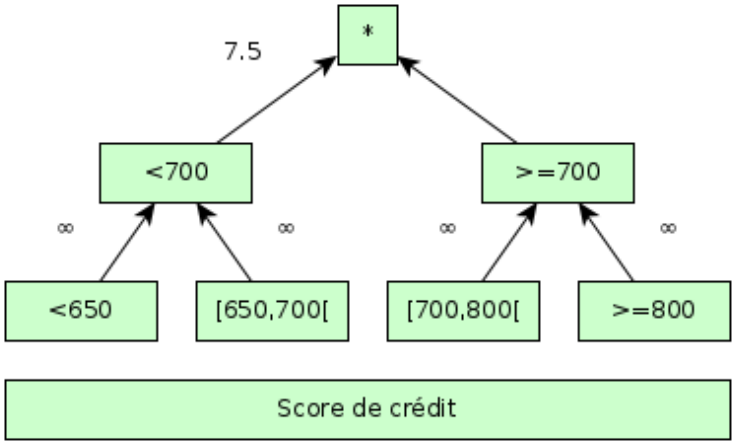


Figure 5.3 – Spécialisation du score de crédit émulée

Un point important : si l'une des spécialisations mène à un non-respect du *LKC*-anonymat, *les spécialisations faites précédemment seront conservées*. Il est une bonne idée de veiller à une hiérarchie logique.

5.4 Difficultés envisageables sur les ensembles de données

À travers nos partages réussis (et nos échecs), nous avons rencontré des ensembles de données nettement plus difficiles que d'autres. Plus spécifiquement, deux conditions ont été rencontrées qui ont compliqué notre travail. Nous les présentons sans plus attendre.

5.4.1 Forte corrélation entre un vecteur de QID et un attribut sensible

La plus difficile des conditions à traiter est une forte corrélation entre un vecteur de QID et un attribut sensible. Supposons que l'attribut A est hautement sensible et que nous souhaitons protéger les personnes le possédant. Le pire cas possible arrive lorsqu'une combinaison de QID donne $n \geq K$ tuples distincts, *possédant tous l'attribut A*. Ceci nous donne un $C = 1$ pour cet attribut. L'algorithme *CLiKC* va alors éviter les spécialisations menant à cette situation, livrant un ensemble de données fortement sous-optimal, specialement si C est proche de 0.

La réponse la plus facile est de remplacer l'attribut A par quelque chose de non-sensible, du style *Autre*. Une autre approche, à utiliser avec précaution, est d'utiliser l'algorithme avec $C = 1$ (éliminant les vérifications sur la distribution des valeurs sensibles), puis de traiter les attributs sensibles de tous les sous-groupes de $l \leq L$ manuellement. Nous ne pouvons pas recommander systématiquement cette approche : bien qu'elle permet d'obtenir des résultats satisfaisants, mais prend un temps prohibitif sur une base de données de grande taille, étant donné qu'il faut regarder chaque groupe aux QID identiques formé et calculer manuellement les probabilités d'identification. On peut finalement se retrouver dans un état inconsistant, où l'attribut A est risqué selon le critère C demandé pour un sous-groupe mais pas pour un autre.

5.4.2 Nécessité de conserver l'amplitude d'un attribut

Dans les deux exemples que nous avons présentés dans le Chapitre 4, l'âge fut fortement généralisé dans les deux cas. Il nous est arrivé d'avoir besoin d'une précision beaucoup plus granulaire que la période de 5 ans rencontrée dans la Section 4.1.2.

La première étape est de définir des hiérarchies sur les dates avec des bonus-malus appropriés afin d'obtenir le niveau de granularité disponible. Il est également nécessaire de jouer avec les paramètres L, K et C pour obtenir un bon compromis. Toutefois, la meilleure façon reste de s'assurer d'avoir un échantillon suffisament volumineux. Il n'existe pas de règle formelle sur la taille minimale pour obtenir une granularité intéressante, mais nous croyons en la maxime du *plus y'en a, mieux c'est*.

5.5 Partages répétés

Le partage répété d'un ensemble de données pose également son lot de problèmes et gérer cette réalité n'est pas chose facile. Nous identifions deux problèmes liés à de multiples partages et proposons des pistes de discussion pour s'en prémunir.

5.5.1 Partage selon deux taxonomies différentes

Ici, nous partageons deux ensembles de données issus de la même population, mais selon des taxonomies différentes. Notre ensemble de données sera alors *LKC*-anonyme que pour l'union de ces deux ensembles anonymisés.

La justification est très simple : Si les QID sont séparés différemment mais que les attributs sensibles restent les mêmes, il est possible de reséparer les sous-groupes formés dans le premier ensemble de données en utilisant les données du second. Prenons en exemple les données dans les Tableaux 5.I et 5.II réputés $(2;20;0,5)$-anonymes. En joignant les deux ensembles, nous en obtenons un illustré dans le Tableau 5.III. L'ensemble résultant n'est, si le tableau est complet, que $(1;1;0,8)$-anonyme, une garantie qui ne vaut plus rien.

Tableau 5.I – Ensemble de données A

Âge	Score	Attribut sensible	Nombre de tuples
18-25	500-600	Attribut A	30
18-25	500-600	Attribut B	25
18-25	500-600	Attribut C	20

Tableau 5.II – semble de données B

Âge	Score	Attribut sensible	Nombre de tuples
20-25	500-600	Attribut A	26
20-25	500-600	Attribut B	24
20-25	500-600	Attribut C	20

Tableau 5.III – Ensemble de données A et B fusionnés

Âge	Score	Attribut sensible	Nombre de tuples
18-19	500-600	Attribut A	**4**
18-19	500-600	Attribut B	**1**
20-25	500-600	Attribut A	26
20-25	500-600	Attribut B	24
20-25	500-600	Attribut C	20

La contextualisation du domaine d'application est une piste de solution pour résoudre ce problème. En définissant des taxonomies statiques sur les champs critiques et en s'y tenant, nous pouvons éviter de rencontrer ce genre de problème. Le compromis est une perte de flexibilité sur les intervalles souhaités.

5.5.2 Partage temporel de données

Une base de données d'assurance n'est pas une entité statique : de nouvelles polices sont vendues, d'autres sont payées, certains assurés terminent leur contrat, etc.

Supposons que nous souhaitons identifier un individu dont nous savons la date d'entrée dans la base de données en plus d'un vecteur L de QID. En accumulant les partages chronologiques de l'ensemble de données, nous pouvons différencier celui précédant l'inscription de l'individu ainsi que celui suivant et tenter de l'identifier. Ceci est une dérivée de l'attaque différentielle illustré en Section 5.2. Cette attaque n'est cependant pas sans faille : dans le cas où les produits vendus sont temporaires[3], les terminaisons abondantes entraînent un changement plus important dans les données.

Deux réactions sont possibles sur ce scénario :

[3]La très grande majorité des produits d'assurance-vie vendus sont temporaires, ainsi que la quasi-totalité des produits d'assurance auto et habitation, eux sur base annuelle

- espacer les partages chronologiques afin de s'assurer que les mouvements d'individus soient suffisament significatifs pour éviter de causer préjudice. Cela peut être fait en testant le *LKC*-anonymat de la différence des deux ensembles chronologiques ou en testant les tuples existant dans les deux ensembles de données ;

- si cela n'est pas possible, limiter les spécialisations de sorte que la différence des deux ensembles de données soit toujours *LKC*-anonyme. Sous ce scénario, plus nous progressons, plus l'ensemble de données sera généralisé.

5.6 Généralisation-discrétisation ou ajout de bruit ?

Un parallèle important peut être tracé entre le concept de généralisation-discrétisation (voir la Section 4.2) et l'introduction de bruit. Prenons en exemple une date de naissance : 25 mars 1998. Supposons que nous généralisons au plan de l'intervalle 1995-2000, avant de discrétiser pour obtenir le 28 mai 1999. Ne serait-ce pas plus simple d'introduire directement un bruit avec moyenne et variance contrôlée pour obtenir directement une date proche ?

La différence réside dans le concept de la *véracité de l'information divulguée* et est discutée sommairement en [25]. Dans le premier cas, l'information que nous donnons est "1995-2000", ce qui est un intervalle plus général mais reste une information *toujours vraie* sur la vraie date de naissance. L'utilisateur possède la liberté de discrétiser selon ses besoins. Toutefois, dans le second cas, nous livrons dès le départ une information *fausse*. Si l'utilisateur décide de discrétiser selon un intervalle plus petit (par exemple l'année), l'intervalle sera potentiellement faux. Tel est notre cas, 1999 n'est évidemment pas égal à 1998.

Un risque inusité est également présent avec des données sérialisées bruitées. Un utilisateur non-conscient du bruit appliqué peut inférer une mauvaise information sur un individu et cela peut entraîner des conséquences fâcheuses. Prenons un exemple bidon où nous tentons d'identifier le score de crédit d'un individu né le 3 mars 1994 selon sa date de naissance. Notre base de données avant bruit possède un seul tuple $(1994-03-03,727)$. Suite au bruitage, nous obtenons deux tuples pour cette date :

$$(1994-03-03,410) \text{ et } (1994-03-03,446)$$

Nous apprenons donc *à tort* que cet individu a un (très) mauvais score de crédit. Nous ne pouvons pas toujours contrôler l'usage de nos données une fois hors de notre portée et n'avons pas la garantie que l'ensemble de données, une fois partagé, gardera la mention de bruit appliqué.

5.7 Généralisation

PEPS est facilement généralisable à un domaine d'affaire différent si nous prenons la peine d'adapter les hypothèses sous-jacentes. Le concept de *LKC*-anonymat est facilement portable dans plusieurs domaines [12] [50] [26]. Notre extension de ce concept via une contextualisation des taxonomies est également facilement transportable, comme nous le verrons prochainement.

Le concept de généralisation-discrétisation, quant à lui, peut être appliqué ailleurs selon certaines conditions. Afin de pouvoir discrétiser les attributs généralisés, nous devons avoir une hypothèse sur la distribution de la répartition inter-bornes. Plus celle-ci sera précise, plus notre utilité sera augmentée.

Nous terminerons cette section avec un exemple ludique d'application de *PEPS* a un domaine différent de l'assurance[4] : le commerce électronique.

Supposons que nous sommes des employés de POStify, une compagnie fournissant des terminaux de point de vente (*Point Of Sale terminal*) ainsi qu'une boutique en ligne permettant d'acquérir différents items répartis dans différentes catégories. Notre entonnoir de ventes comporte 7 étapes :

- la page de l'item en question : le client ajoute l'item à son panier ;
- la visualisation du panier : le client confirme ;
- l'entrée des renseignements de livraison et de paiement ;
- la proposition d'items supplémentaires ;

[4] Et hors de ma zone de confort !

- le choix du moyen de livraison ;
- la confirmation finale ;
- l'affichage du bon de commande et la confirmation du paiement.

Nous collectons de nombreux éléments d'information : l'item regardé, l'étape d'abandon, le contenu final de la commande, le choix de la livraison, l'adresse IP, le fournisseur d'accès internet, la localisation géographique, le navigateur, le type d'appareil, la date et l'heure, le nom, l'adresse, le moyen de paiement, etc. Nous pouvons également déterminer le nombre de commandes passées dans un intervalle donné.

Nous voyons ici que l'étape d'abandon suit une fonction de survie, avec ici $S(1) = 1$ (tout le monde dans la base de données se rend au moins à l'étape 1) et (7) équivaut à la probabilité qu'un consommateur démarrant le processus se rende à un achat. Un parallèle évident peut être tracé avec la table CIA9704, bien que celle de notre exemple soit beaucoup moins compliquée.

Sans rentrer dans les détails, nous pouvons voir qu'il est tout à fait possible de déterminer des taxonomies afin de pouvoir publier un ensemble de données anonymisé. Le processus reste le même : en déterminant le genre d'analyse souhaitée, nous attribuerons des bonus aux spécialisations avant de procéder avec l'ensemble du système *PEPS*.

5.8 Et si un partage est impossible ?

Que faire si, malgré nos précautions, il nous est impossible de partager selon des critères que nous jugeons raisonnables ? La réponse selon le *PEPS* est de tout simplement ne pas partager.

D'accord, mais je n'ai pas le choix. Bien qu'il nous apparaît caduc d'utiliser un système pour ensuite ignorer le travail accompli, nous sommes conscients que la réalité est différente des modèles théoriques. À cette fin, il convient de mesurer adéquatement les risques en utilisant la théorie décrite dans la Section 2.5 et de trouver des stratégies de mitigation autres. De l'assurance confidentialité existe déjà[5], alors pourquoi pas de l'assurance réidentification ?

[5] `https://www.avivacanada.com/privacybreach`

CHAPITRE 6

CONCLUSION

La problématique approchée dans ce mémoire est liée au cycle complet de partage confidentiel des données. Nous avons pu voir que celle-ci dépasse la simple étape d'anonymisation, comme vue dans les systèmes cités en Section 2.6.

Nous rappellerons les grandes étapes empruntées par *PEPS* avant de proposer des pistes d'extensions. La première étape est la caractérisation du domaine d'application et des analyses performées, illustrées par les *Protocoles* ainsi que les *Politiques et Contraintes*. Un travail préalable modeste doit également être entrepris pour la gestion des données, représenté par la *base de données originale*. Le coeur du processus réside dans trois étapes, liées ensemble par une boucle de rétroaction :

- l'étape de *prétraitement* permet d'adapter l'ensemble de données original suite aux règles et contraintes et permet de contextualiser les attributs en vue d'une anonymisation efficace ;
- l'étape d'*anonymisation* assure le respect du modèle de confidentialité proposé, soit le *LKC*-anonymat, tout en prenant avantage de la connaissance propre au domaine ;
- l'étape d'*équilibration* (ou l'équilibrateur) conjugue utilité et confidentialité afin de permettre une solution satisfaisante au problème rencontré.

Notre système en outre sait s'ajuster aux forces externes, notamment sous la forme des politiques susmentionnées, mais également en proposant des stratégies de réaction aux *attaques*. En comprenant la nature de celles-ci, nous pouvons éviter de causer un préjudice supplémentaire et bien déterminer le risque de divulgation auquel nous nous exposons. Finalement, notre système conserve en mémoire les paramétrisations des partages passés, de sorte qu'il est possible de reproduire un partage passé et de s'en inspirer pour ceux futurs.

PEPS est à notre avis une excellente première étape pour la gestion de la confidentialité pour les propriétaires de données. Cela signifie que de nombreux travaux futurs peuvent s'en inspirer. Impossible de passer sous silence l'importance d'avoir une interface soignée et intuitive : une opportunité de développer une expérience utilisateur rehaussée est bien présente. Nous croyons également que notre algorithme, *CLiKC*, serait fort bienvenu dans une plateforme d'anonymat générale du style ARX [22]. Nous soulignons par le fait même l'importance du développement libre de solutions d'anonymisation : tout comme pour la cryptographie, il nous est impossible de faire confiance aveuglément à un programme présenté comme une *boîte noire*. Finalement, le perfectionnement d'un langage de définition pour les hiérarchies nous apparait comme une opportunité de développement : un langage de description de graphes (à la yED[1]) nous semble comme une bonne avenue, en plus de fournir un résultat visuel immédiat.

Nous finirons ce mémoire par une contribution inusitée de la rédaction de celui-ci. Ayant la chance d'oeuvrer dans le domaine de l'assurance, les nombreuses discussions que j'ai pu avoir avec de nombreux collègues les ont sensibilisés à la protection des renseignements confidentiels. La mission d'éducation étant éternelle, nous sommes très heureux de cet effet collatéral et souhaitons que la vulgarisation de la recherche en confidentialité ne cesse jamais.

[1] `https://www.yworks.com/products/yed`

BIBLIOGRAPHIE

[1] Odd Aalen. Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics*, 6(4):701–726, 1978.

[2] Alessandro Acquisti, Leslie John et George Loewenstein. What is privacy worth. *The Journal of Legal Studies*, 42, 2013-05.

[3] Charu C. Aggarwal. On k-anonymity and the curse of dimensionality. Dans *Proceedings of the 31st international conference on Very large data bases*, pages 901–909. VLDB Endowment, 2005.

[4] Bernhard Meindl. CRAN - Package sdcTable, 2016. URL `https://cran.r-project.org/web/packages/sdcTable/`. [Accédé le 2016-04-16].

[5] Besurance Corporation. Besurance Corporation, 2015. URL `http://www.besurance.ca/`. [Accédé le 2016-04-09].

[6] N.L. Bowers. *Actuarial Mathematics*. Society of Actuaries, 1997. ISBN 978-0-938959-46-5.

[7] Canadian Institute of Actuaries. Lapse Experience Under Lapse-Supported Policies, 1999-10. URL `https://www.cia-ica.ca/docs/default-source/1999/9954e.pdf?sfvrsn=0`. [Accédé le 2016-04-17].

[8] Canadian Institute of Actuaries. Construction of CIA9704 Mortality Tables for Canadian Individual Insurance based on data from 1997 to 2004, 2010-05-01. URL `http://www.actuaries.ca/members/publications/2010/210028e.pdf`. [Accédé le 2016-03-13].

[9] Canadian Institute of Actuaries. Lapse Experience under Universal Life Level Cost of Insurance Policies, 2015-09. URL `http://www.cia-ica.ca/docs/default-source/2015/215076e.pdf`. [Accédé le 2016-04-09].

[10] Chris Clifton et Tamir Tassa. On Syntactic Anonymity and Differential Privacy. Dans *Transactions on Data Privacy*, volume 6, pages 161–183, 2013.

[11] Communitech. Manulife opens RED Lab in Communitech Hub – Communitech, 2016. URL `https://www.communitech.ca/press-release/manulife-opens-red-lab-in-communitech-hub/`. [Accédé le 2016-04-09].

[12] Gaby G. Dagher, Farkhund Iqbal, Mahtab Arafati et Benjamin C. M. Fung. Fusion : Privacy-Preserving Distributed Protocol for High-Dimensional Data Mashup. Dans *2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS)*, pages 760–769. IEEE, 2015-12. ISBN 978-0-7695-5785-4.

[13] David C. M. Dickson, Mary R. Hardy et Howard Richard Waters. *Actuarial mathematics for life contingent risks*. International series on actuarial science. Cambridge Univ. Press, 2. ed édition, 2013. ISBN 978-1-107-04407-4.

[14] Cynthia Dwork. Differential privacy. Dans *Automata, languages and programming*, pages 1–12. Springer, 2006.

[15] Cynthia Dwork. Differential privacy : A survey of results. Dans *Theory and applications of models of computation*, pages 1–19. Springer, 2008.

[16] Cynthia Dwork. A Firm Foundation for Private Data Analysis, 2011. URL `http://research.microsoft.com/pubs/116123/dwork_cacm.pdf`. [Accédé le 2016-04-06].

[17] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov et Moni Naor. Our data, ourselves : Privacy via distributed noise generation. Dans *Advances in Cryptology-EUROCRYPT 2006*, pages 486–503. Springer, 2006.

[18] Cynthia Dwork, Frank McSherry, Kobbi Nissim et Adam Smith. Calibrating noise to sensitivity in private data analysis. Dans *Theory of cryptography*, pages 265–284. Springer, 2006.

[19] Cynthia Dwork et Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3-4):211–407, 2013. ISSN 1551-305X, 1551-3068.

[20] Hamid Ebadi, David Sands et Gerardo Schneider. Differential Privacy : Now it's Getting Personal. Dans *Proceedings of the 42nd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, pages 69–81. ACM Press, 2015. ISBN 978-1-4503-3300-9.

[21] Khaled El Emam, Fida K. Dankar, Régis Vaillancourt, Tyson Roffey et Mary Lysyk. Evaluating the risk of re-identification of patients from hospital prescription records. *The Canadian journal of hospital pharmacy*, 62(4):307, 2009.

[22] Fabian Prasserer, Florian Kohlmay, Ronald Lautenschläger et Klaus A. Kuhn. ARX - A Comprehensive Tool for Anonymizing Biomedical Data. *Proceedings of the AMIA 2014 Annual Symposium*, 2014. Washington D.C., USA.

[23] Gregory Ferenstein. Misc : Why "Data Scientist" Is The Best Job To Pursue In 2016, 2016-01-20. URL `http://www.forbes.com/sites/gregoryferenstein/2016/01/20/misc-why-data-scientist-is-the-best-job-to-pursue-in-2016/`. [Accédé le 2016-03-09].

[24] Financial Consumer Agency of Canada. *Understanding your credit misc and credit score.* Financial Consumer Agency of Canada, 2008. ISBN 978-1-100-10209-2.

[25] Benjamin C. M. Fung, Ke Wang, Rui Chen et Philip S. Yu. Privacy-preserving data publishing : A survey of recent developments. *ACM Computing Surveys*, 42(4): 1–53, 2010-06-01. ISSN 03600300.

[26] Benjamin C.M. Fung, Thomas Trojer, Patrick C.K. Hung, Li Xiong, Khalil Al-Hussaeni et Rachida Dssouli. Service-Oriented Architecture for High-Dimensional Private Data Mashup. *IEEE Transactions on Services Computing*, 5(3):373–386, 2012-23. ISSN 1939-1374.

[27] Benjamin C.M. Fung, Ke Wang et Philip S. Yu. Anonymizing Classification Data for Privacy Preservation. *IEEE Transactions on Knowledge and Data Engineering*, 19(5):711–725, 2007-05. ISSN 1041-4347.

[28] Arpita Ghosh, Tim Roughgarden et Mukund Sundararajan. Universally Utility-maximizing Privacy Mechanisms. *SIAM Journal on Computing*, 41(6):1673–1693, 2012-01. ISSN 0097-5397, 1095-7111.

[29] Henk C. A. van Tilborg et Sushil Jajodia. *Encyclopedia of Cryptography and Security*. Springer, 2 édition, 2011.

[30] R. Hickey. The Clojure programming language. Dans *Proceedings of the 2008 symposium on Dynamic languages*. ACM New York, NY, USA, 2008.

[31] Anco Hundepool et Peter-Paul de Wolf. Methods Series : Statistical disclosure control. *Statistics Netherlands*, 2011.

[32] ICML 2016. TPDP 2016 – Theory and Practice of Differential Privacy, 2016. URL `http://tpdp16.cse.buffalo.edu/`. [Accédé le 2016-04-16].

[33] Individual Life Experience Subcommittee. Lapse Experience Study for 10-Year Term Insurance, 2014-01. URL `http://www.cia-ica.ca/docs/default-source/2014/214011e.pdf`.

[34] Institut Canadien des Actuaires. Risques normaux grande branche au Canada 2012-2013 à l'aide des tables 97-04, 2015-07. URL `http://www.cia-ica.ca/docs/default-source/2015/215062f.pdf?sfvrsn=0`. [Accédé le 2016-04-19].

[35] International Organization for Standardization. *ISO/IEC 9075 :1992 : Information technology — Database languages — SQL*. pub-ISO, 1992.

[36] Journal de l'assurance. Parts de marché des principaux assureurs vie au Canada et au Québec, 2009. URL `http://journal-assurance.ca/media/docs/extra-page-52-tableaux.pdf`. [Accédé le 2016-04-24].

[37] Kanetix. Compare Car Insurance Quotes & Save - Kanetix.ca, 2016. URL https://www.kanetix.ca/auto-insurance. [Accédé le 2016-03-09].

[38] Jonathan Katz et Yehuda Lindell. *Introduction to Modern Cryptography (Chapman & Hall/Crc Cryptography and Network Security Series)*. Chapman & Hall/CRC, 2007. ISBN 1-58488-551-3.

[39] Rashid Hussain Khokhar, Rui Chen, Benjamin CM Fung et Siu Man Lui. Quantifying the costs and benefits of privacy-preserving health data publishing. *Journal of biomedical informatics*, 50:107–121, 2014.

[40] Stuart A. Klugman, Harry H. Panjer et Gordon E. Willmot. Continuous Models. Dans *Loss Models*, pages 61–100. John Wiley Sons, Inc., 2008. ISBN 978-0-470-39134-1.

[41] Stuart A. Klugman, Harry H. Panjer et Gordon E. Willmot. Credibility. Dans *Loss Models*, pages 555–640. John Wiley Sons, Inc., 2008. ISBN 978-0-470-39134-1.

[42] Aleksandra Korolova. Protecting Privacy when Mining and Sharing User Data. *Thèse de doctorat - Stanford University*, 2012.

[43] Derek Kueker. The Power of Big Data : An RGA Case Study, 2015-12. URL http://www.rgare.com/offices/germany/Documents/Europe%20Quarterly_bigdata.pdf. [Accédé le 2016-04-16].

[44] Derek Kueker, Tim Rozar, Michael Cusumano, Suzan Willeat et Richard Xu. Misc on the Lapse and Mortality Experience of Post-Level Premium Period Term Plans (2014), 2014-05. URL https://www.soa.org/Files/Research/Exp-Study/research-2014-post-level-shock-misc.pdf. [Accédé le 2016-03-25].

[45] K. LeFevre, D. J. DeWitt et R. Ramakrishnan. Mondrian Multidimensional K-Anonymity. Dans *22nd International Conference on Data Engineering (ICDE'06)*, pages 25–25, 2006-04.

[46] Ninghui Li, Tiancheng Li et Suresh Venkatasubramanian. t-closeness : Privacy beyond k-anonymity and l-diversity. Dans *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106–115. IEEE, 2007.

[47] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke et Muthuramakrishnan Venkitasubramaniam. l-diversity : Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3, 2007.

[48] David J. Martin, Daniel Kifer, Ashwin Machanavajjhala, Johannes Gehrke et Joseph Y. Halpern. Worst-case background knowledge for privacy-preserving data publishing. Dans *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 126–135. IEEE, 2007.

[49] Adam Meyerson et Ryan Williams. On the Complexity of Optimal K-anonymity. Dans *Proceedings of the Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '04, pages 223–228. ACM, 2004. ISBN 1-58113-858-X.

[50] Noman Mohammed, Benjamin Fung, Patrick CK Hung et Cheuk-kwong Lee. Anonymizing healthcare data : a case study on the blood transfusion service. Dans *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1285–1294. ACM, 2009.

[51] Arvind Narayanan et Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. Dans *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125. IEEE, 2008.

[52] National Heart, Long and Blood Institute. Assessing Your Weight and Health Risk, 2016. URL `http://www.nhlbi.nih.gov/health/educational/lose_wt/risk.htm`. [Accédé le 2016-04-17].

[53] Peter-Paul de Wolf. mu-ARGUS, 2015. URL `http://neon.vb.cbs.nl/casc/mu.htm`. [Accédé le 2016-04-16].

[54] Research Committee Individual Life Subcommittee. Lapse Experience under Universal Life Level Cost of Insurance Policies, 2007-10. URL http://www.cia-ica.ca/docs/default-source/2007/207086e.pdf. [Accédé le 2016-04-17].

[55] RGAx. rgax, 2015. URL http://www.rgax.com/. [Accédé le 2016-04-09].

[56] Jonathan Rioux. Archipel : Un logiciel d'études d'expérience pour l'assurance-vie. *RGA Canada*, 2015.

[57] Tim Rozar, Scott Rushing et Suzan Willeat. Misc on the Lapse and Motality Experience of Post-Level Premium Period Term Plans, 2010-07.

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

[59] Rathindra Sarathy et Krishnamurty Muralidhar. Evaluating Laplace Noise Addition to Satisfy Differential Privacy for Numeric Data. *Trans. Data Privacy*, 4(1):1–17, 2011-04. ISSN 1888-5063.

[60] Sim Segal. *Corporate value of enterprise risk management : the next step in business management*. Wiley corporate F&A. Wiley, 2011. ISBN 978-0-470-88254-2.

[61] Erez Shmueli, Tamir Tassa, Raz Wasserstein, Bracha Shapira et Lior Rokach. Limiting disclosure of sensitive data in sequential releases of databases. *Information Sciences*, 191:98 – 127, 2012. ISSN 0020-0255. Data Mining for Software Trustworthiness.

[62] Andrzej Skowron et Cecylia Rauszer. Intelligent Decision Support : Handbook of Applications and Advances of the Rough Sets Theory. Dans Roman Słowiński, éditeur, *Handbook of Applications and Advances of the Rough Sets Theory*, pages 331–362. Springer Netherlands, 1992. ISBN 978-94-015-7975-9.

[63] Latanya Sweeney. How Unique are You ?, 2013. URL http://aboutmyinfo.org/index.html. [Accédé le 2015-09-06].

[64] Term4sale. Term4Sale - Term Life Insurance Quotes and Comparisons, 2016. URL https://www.term4sale.ca/. [Accédé le 2016-04-09].

[65] The R Foundation. R : What is R ?, 2016. URL https://www.r-project.org/about.html. [Accédé le 2016-04-16].

[66] Michael Tjepkema et others. Adult obesity in Canada : Measured height and weight. *Statistics Canada*, 1:1–32, 2005.

[67] University of Texas as Dallas Data and Privacy Lab. UTD Anonymization ToolBox, 2010. URL http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php?go=home. [Accédé le 2016-04-16].

[68] Ke Wang et Benjamin C. M. Fung. Anonymizing Sequential Releases. Dans *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 414–423. ACM, 2006. ISBN 1-59593-339-5.

[69] Ke Wang, Benjamin C. M. Fung et Philip S. Yu. Handicapping attacker's confidence : an alternative to k-anonymization. *Knowledge and Information Systems*, 11(3):345–368, 2006. ISSN 0219-3116.

[70] Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu et Ke Wang. (alpha, k)-anonymity : an enhanced k-anonymity model for privacy preserving data publishing. Dans *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 754–759. ACM, 2006.

[71] Xiaokui Xiao, Guozhang Wang et Johannes Gehrke. Interactive anonymization of sensitive data. Dans *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 1051–1054. ACM, 2009.

[72] Abdulsalam Yassine et Shervin Shirmohammadi. Privacy and the market for private data : a negotiation model to capitalize on private data. Dans *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on*, pages 669–678. IEEE, 2008.

[73] Marek P. Zielinski et Martin S. Olivier. On the use of economic price theory to find the optimum levels of privacy and information utility in non-perturbative microdata anonymisation. *Data & Knowledge Engineering*, 69(5):399–423, 2010-05. ISSN 0169023X.

[74] École de politique appliquée, Université de Sherbrooke. Canada - Taux de natalité brut (par 1000 habitants) | Statistiques, 2016. URL `http://perspective.usherbrooke.ca/bilan/tend/CAN/fr/SP.DYN.CBRT.IN.html`. [Accédé le 2016-04-19].

Annexe I

Synthure, un outil de génération de données d'assurés

Motivation

Les données d'assurance de personnes sont souvent considérées par leur propriétaire comme un secret concurrentiel. De plus, le retard par rapport aux meilleures pratiques de partage confidentiel empêche la dissémination efficace de ces données pour la recherche.

Plutôt que de tenter de renverser la vapeur sur une centaine d'années de pratique précautieuse, nous avons préféré développer un outil permettant de générer des données dites *plausibles*.

Par rapport à un générateur purement aléatoire, notre outil se démarque sur trois points :

1. Le format des données est automatiquement adapté à une étude de mortalité type, et peut être extensible si de nouveaux champs pertinents apparaissent.

2. Nous pouvons générer les décès sur une courbe de mortalité connue (par défaut la table de mortalité CIA9704[8] sur base âge atteint, mais d'autres tables peuvent être utilisées), ce qui évite d'avoir des résultats erratiques susceptibles de confondre l'utilisateur.

3. Les quasi-identificateurs peuvent également facilement suivre une distribution basée sur les données d'industrie où d'en vigueur à une compagnie.

Origine du nom et développement

Synthétique + Assure = Synthure. Se prononce comme le mot français *ceinture*.

Le code source de l'outil est librement disponible et le développement de trouve dans un dépôt public[1]. Le projet est écrit en Clojure[30] et est déposé sous la même licence

[1] Bitbucket : `https://bitbucket.org/jonesberg/synthure`

que le langage[2].

Fonctionnement et paramétrisation

Pour le moment, le logiciel est codé afin d'appuyer le développement de l'outil PEPS et la paramétrisation reste minimale. Les valeurs par défaut sont toutefois soigneusement sélectionnées et il est possible d'obtenir un ensemble de données convaincant avec celles-ci. Les prochaines versions tenteront de combler cette lacune par des options plus exhaustives.

La variable `db` disponible dans le fichier `core.clj` définit une table de hachage comportant les paramètres de sortie. Il est possible de changer la valeur de la clé `:subname` pour le nom de fichier souhaité. Une fois le fichier compilé dans le REPL[3], il suffit d'appeler la création de la base de données

```
(create-db db)
```

puis de générer les records

```
(insert-n-values! db n)
```

où `n` est le nombre de records générés.

Paramètres par défaut sur les dates

Les paramètres suivants sont configurables en ajustant le code source. Les prochaines versions veilleront à permettre une paramétrisation plus facile. Outre les dates, plusieurs autres paramètres sont possibles. Au besoin, référer à la paramétrisation sélectionnée pour *PEPS*.

[2]Disponible dans le dépôt de code et à l'adresse suivante : `https://www.eclipse.org/legal/epl-v10.html`

[3]Read-eval-print loop, un outil de développement interactif disponible pour le langage Clojure.

Date de début et de fin d'exposition

Nous considérons un horizon rétrospectif de 10 ans entre le début de l'exposition et la fin. De cette façon, nous enregistrons uniquement les événements entre le 1er janvier 2005 et le 1er janvier 2015. Les polices terminées ou payées avant cette date ne seront pas enregistrées dans la base de données et celles simulées avec des dates d'événement après la borne supérieure seront considérées comme *en force* (code « INF »).

Hypothèses de déchéance

Nous utilisons les tables de mortalité de l'Institut Canadien des Actuaires[8] sans amélioration. Ces tables sont les plus à jour dans ce qui est disponible publiquement. Nous discriminons selon le sexe et le statut fumeur, mais utilisons uniquement celles calculées sur l'âge atteint[4].

Les hypothèses de terminaison sont basées encore une fois sur une version simplifiée[5] des études de l'Institut Canadien des Actuaires. Dans le cas des produits temporaires, nous assumons que les produits ne sont pas renouvelables, c'est-à-dire que le taux de terminaison est de 100% lors de la fin du terme.

La distribution fractionnaire des déchéances est répartie uniformément sur la période, ce qui est l'hypothèse la plus fréquente et la plus aisée à comprendre. Les prochaines versions du programme permettront d'utiliser une distribution différente.

Hypothèses d'arrivée : naissance et prise d'assurance

Nous supposons ici simplement que les naissances sont uniformes entre aujourd'hui et il y a 120 ans[6]. Pour la prise d'assurance, nous assumons une prise uniforme entre les 18es anniversaires — âge de la majorité — et le 65es — âge normal de la retraite.

[4]Comme nous avons accès aux dates de naissance, il ne nous est pas nécessaire d'utiliser la méthode approximative de l'âge arrondi.

[5]nous n'effectuons ici aucune discriminaison sur le sexe et/ou le statut fumeur et faisons abstraction de la période de sélection.

[6]Les tables de mortalité récentes sont le plus souvent limitées à 120 ans.