Publication de données individuelles respectueuse de la vie privée : une démarche fondée sur le co-clustering

Tarek Benkhelif

UNIVERSITE BRETAGNE LOIRE MATHSTIC

UNIVERSITÉ DE NANTES

THESE DE DOCTORAT DE

L'UNIVERSITE DE NANTES

COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par

« Tarek BENKHELIF »

« Publication de données individuelles respectueuse de la vie privée »

« Une démarche fondée sur le co-clustering»

Thèse présentée et soutenue à Nantes, le 27 novembre 2018
Unité de recherche : Laboratoire des Sciences du Numérique de Nantes (LS2N – équipe Duke)
Thèse N° : 145081

Rapporteurs avant soutenance :

Benjamin Nguyen	Professeur des universités, INSA Centre Val de Loire
Christophe Rosenberger	Professeur des universités, École nationale supérieure d'ingénieurs de Caen)

Composition du Jury :

Président :	Maryline Laurent	Professeur des universités, Telecom SudParis
Examinateurs :	Pierre Gançarski	Professeur des universités, ICube
Dir. de thèse :	Marc Gelgon	Professeur des universités, Université de Nantes
Co-dir. de thèse :	Guillaume Raschia	Maître de conférences, Université de Nantes

Invité(s)

Françoise Fessant	Docteur Ingénieur R&D, Orange Labs
Matthieu Grall	Chef du service de l'expertise technologique, CNIL

Remerciements

Je remercie en premier lieu les personnes qui ont contribué à l'élaboration de mes travaux de thèse en particulier mes encadrants, Françoise Fessant, Guillaume Raschia et Marc Gelgon.

Je tiens à exprimer toute ma gratitude envers Françoise pour m'avoir offert l'opportunité de travailler sur le sujet passionnant que représente la protection de la vie privée, d'abord grâce au stage, puis à travers la thèse de trois ans. Je tiens à saluer la bonne humeur, la patience, le soutien et l'investissement dont elle a fait preuve tout au long de l'aventure. Je la remercie pour l'encadrement scientifique qu'elle m'a fournit : sa rigueur, son recul, son sens de l'organisation, et ses conseils m'ont permis d'améliorer la qualité de mes travaux tout en tenant les délais.

Je suis reconnaissant envers Guillaume et Marc. Je les remercie d'avoir co-encadré cette thèse, pour l'accueil qu'ils m'ont réservé lors de mes multiples visites à Nantes qui ont toujours été agréables et formatrices. Je les remercie pour leur implication dans mes travaux. Merci à vous deux d'avoir toujours su trouver du temps à m'accorder malgré vos emplois du temps bien chargés.

Je remercie l'équipe PROF d'Orange Labs à Lannion pour leur accueil pendant ces trois années de thèse. Merci à Fabrice, Vincent, Marc, Franck, Raphaël, Felipe, Bruno, Carine, Tanguy, Nicolas, Elias, Wissam, Barbara, Fabien, Anaïs, Sylvie, Philippe, Manu, Pierre et Oumaima pour leur bienveillance et leur disponibilité.

Je remercie les membres du jury de m'avoir fait l'honneur d'accepter d'étudier et d'évaluer mon travail. Merci aux rapporteurs (Pr. Benjamin Nguyen et Pr. Christophe Rosenberger) pour les rapports de qualité qu'ils ont rédigés, pour les points positifs soulevés et pour les conseils qui nous permettront d'améliorer notre travail à l'avenir. Merci également aux examinateurs (Pr. Maryline Laurent, Pr. Pierre Gançarski et Matthieu Grall) pour leur investissement et leur participation dans le jury.

Un grand merci à ma famille costarmoricaine Yassine, Wissam, Asmaa et Elias, qui ont fait de mon exil à Lannion une agréable expédition.

Je suis reconnaissant envers mes amis Okba, Islem, Hani, Redha, Rachid, Marwan,

Ghazi et Didine, merci pour les innombrables moments d'insouciance, de liberté et de bonheur que nous avons partagés ensemble.

Merci à ma grand-mère pour son amour et son soutien.

Je tiens à remercier tata Lala ma tante bien aimée, une personne d'une extrême bonté qui a toujours cru en moi. Tu as marqué mon enfance avec ta générosité, tes encouragements et ton amour.

Je remercie mes sœurs : Yasmine probablement ma première et éternelle meilleure amie et Lina la jeune femme qui restera à mes yeux ce petit bébé dont la venue au monde n'a cessé d'emplir mon cœur de joie jusqu'à ce jour. Merci, c'est un plaisir d'être votre grand frère.

Sarah, il te suffit souvent d'une phrase ou deux pour m'apaiser. Merci pour ton soutient inconditionnel, ta positivité et la joie que tu m'apportes au quotidien.

Aucun texte ne pourrait capturer fidèlement la gratitude que j'éprouve à l'égard de mes parents, sans doute les personnes les plus généreuses que je connaisse. Merci pour votre confiance, votre dévouement et les valeurs que vous m'avez inculquées. C'est un honneur d'être votre fils, je vous dois tout.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Sommaire

Les données jouent un rôle capital dans la société d'aujourd'hui, leur importance dans l'innovation, les sciences et la création de biens et de services est certaine. Alors que la digitalisation au cours de la dernière décennie a fortement accéléré l'augmentation de la quantité de données numériques, l'intérêt que manifestent les organismes publics et privés pour le stockage et l'exploitation massifs de ces données a lui aussi augmenté. En effet, la quantité de données collectées dans le monde d'ici 2020 devrait dépasser 44 milliards de gigaoctets [(ID14] et la Commission européenne estime que la valeur marchande des données européennes devrait atteindre 739 milliards d'euros [(ID17] la même année. Cette évolution fait surgir des préoccupations concernant le droit aux individus de contrôler l'utilisation et le stockage des informations les concernant. Cette thèse s'inscrit dans le registre de la publication de données respectueuse de la vie privée, dont le but est de publier des données utiles en vue de leur exploitation, tout en protégeant la vie privée des individus qui y sont représentés.

1 Protection de la vie privée

1.1 Protection de la vie privée : une définition unifiée ?

La protection de la vie privée est un concept qui est étudié en profondeur dans les disciplines philosophique, psychologique, sociologique, juridique et technique depuis plus de 100 ans [SDX11]. Les préoccupations relatives à la protection de la vie privée ont généralement été soulevées chaque fois que les progrès technologiques ont permis de modifier la façon dont les informations peuvent être recueillies, consultées ou utilisées. Il existe de nombreuses définitions et de nombreux aspects de la vie privée, mais la protection de l'information personnelle a fait l'objet d'une attention considérable au cours des dernières décennies.

De nombreuses études dans divers domaines ont amélioré notre compréhension de la protection de la vie privée et de la gestion de la vie privée à différents niveaux. Cependant, le tableau qui se dégage est fragmenté et généralement propre à une discipline, avec des concepts, des définitions et des relations qui sont incohérents et qui ne sont ni pleinement développés ni validés empiriquement [XDSH11]. Les définitions de la vie privée varient et dépendent du domaine, allant d'un "droit" en science juridique [WB90] à un "état d'accès limité ou d'isolement" en philosophie et en psychologie [Sch84] au "contrôle" dans les sciences sociales et les systèmes d'information [Cul93, Wes68]. Westin [Wes68] définit par exemple la vie privée comme "la revendication d'individus, de groupes ou d'institutions de déterminer par eux-mêmes quand, comment et dans quelle mesure l'information les concernant est communiquée à d'autres". Cependant, selon [Sol08], non seulement l'effort pour "trouver l'essence de la vie privée" a échoué, mais il n'y a aucun espoir de succès. Au lieu de poursuivre la recherche futile d'une théorie aussi grande et unifiée de la protection de la vie privée, Solove [Sol08] soutient que nous devrions aborder les questions de protection de la vie privée de la base vers le haut en cherchant à "résoudre certains problèmes". C'est précisément dans ce cadre-là que les communautés scientifiques des statistiques et informatique se sont intéressées à la problématique et se sont montrées particulièrement prolifiques, en proposant nombre de solutions au cours des dernières décennies.

1.2 Exploitation des données personnelles, bénéfice et risque

La collecte et l'utilisation des données à caractère personnel par les acteurs publics et privés présentent un avantage certain. D'une part pour les individus qui bénéficient directement de l'amélioration et du développement de nouveaux services. D'autre part pour les entreprises notamment pour : l'évaluation du risque client, l'accroissement du retour sur investissement, et l'amélioration de l'offre de services et de produits. A ce sujet Meglena Kuneva, la commissaire européenne à la protection des consommateurs dira : "Les données personnelles sont le nouveau pétrole de l'Internet et la nouvelle monnaie du monde numérique." [Kun09]. La métaphore est particulièrement intéressante. Elle couvre à la fois l'utilisation des données personnelles en tant que produit en soi mais aussi en tant que substance fondamentale pour un grand nombre d'activités économiques, et qui doit être considérée comme matière première ou "produit semi-fini".

Toutefois, le caractère personnel et sensible des données représente un enjeu pour les entreprises, qui sont soumises à une réglementation stricte en matière de collecte, de traitement et de partage des données. En plus des amendes et des sanctions légales prévues par les lois sur la protection des données, les atteintes à la vie privée peuvent également avoir un impact considérable sur la réputation des entreprises et sur les relations avec leurs partenaires, clients et employés, ce qui peut provoquer des pertes économiques considérables.

1.3 Quelques exemples d'atteinte à la vie privée

Les exemples suivants illustrent parfaitement les conséquences graves qu'induit un manquement à la protection de la vie privée.

1.3.1 Netflix

Netflix, le célèbre fournisseur de vidéo à la demande a été poursuivi en justice et a dû payer environ 9 millions de dollars pour avoir exposé les données de ses clients. Les faits remontent à 2010, quand l'entreprise organisa un concours [1] qui visait à améliorer

1. https://www.nytimes.com/2010/03/13/technology/13netflix.html

son système de recommandation de contenu. A cet effet, Netflix a publié 100 millions d'enregistrements qui répertorient les évaluations de 500.000 utilisateurs. Cependant, en reliant les recommandations disponibles sur la base de données du site IMDb avec le jeu de données 'dé-identifié', des chercheurs [NS06] ont montré qu'il était possible de ré-identifier les individus, révélant ainsi des informations sensibles.

1.3.2 Les taxis new-yorkais

En mars 2014, à la suite d'une demande d'accès à l'information, une copie complète des logs des déplacements et des tarifs des taxis de la ville de New York a été publiée. Les données contenaient plus de 173 millions d'enregistrements de déplacements individuels, chaque enregistrement de trajet comprenait le lieu et l'heure de prise en charge et de dépôt, et les frais de la course. Chaque voyage était en outre décrit par le numéro de licence du conducteur et le matricule du véhicule, qui avaient été anonymisés en les remplaçant par un identifiant crypté (leurs hashes MD5). De telles fonctions de hachage ne sont utiles dans l'anonymisation que s'il existe un très grand nombre de possibilités pour les entrées de hachage. Cependant, il n'y a qu'environ 22 millions de possibilités pour les numéros de licence et de matricule, qui peuvent être hachés en quelques minutes. Une fois les hashes connus :

— Il était trivial de chercher le bon numéro, qui pouvait ensuite être corrélé avec d'autres ensembles de données pour révéler l'identité des conducteurs, leurs déplacements complets et les salaires qu'ils gagnaient.

— Les paparazzi de New York capturent fréquemment des célébrités entrant ou sortant des taxis jaunes, et dans beaucoup de leurs photos est affiché le matricule unique du taxi comme illustré dans la Figure 1.1. En recherchant de tels clichés sur Google, puis en utilisant les horodatages des photos et les descriptions qui les accompagnent pour établir où elles ont été prises, il a été possible de déterminer les lieux de ramassage et de dépôt, le montant du tarif et le pourboire que plusieurs célébrités ont payé à leur chauffeur.

Dans cette affaire, la tentative d'anonymisation s'est effondrée en raison d'une mau-

vaise utilisation de la cryptographie [2].

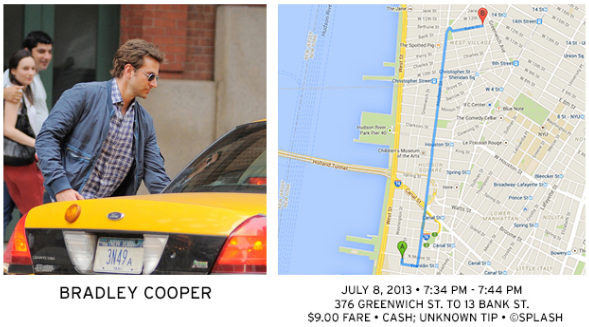


Figure 1.1 – Célébrités, paparazzi et faille de protection de la vie privée [Tro]

1.3.3 Cambridge Analytica

Les informations personnelles de plus de 87 millions d'utilisateurs ont été partagées de manière inappropriée avec le cabinet de conseil politique Cambridge Analytica [3]. La firme partenaire agrée de Facebook a pu proposer un questionnaire aux utilisateurs du réseau social. En réalité, l'application de Cambridge Analytica poursuivait d'autres fins : elle récupérait les données Facebook de toute personne répondant au questionnaire, mais, elle récupérait également au passage, les données Facebook des amis de la personne répondant au questionnaire. Cela a permis à Cambridge Analytica de récupérer les données de 87 millions d'utilisateurs alors que 270.000 personnes ont répondu au questionnaire. C'est-à-dire que pour 1 personne répondant au questionnaire, l'application a accédé en moyenne aux données de 321 autres personnes. En plus de l'identification (nom, prénom), la date de naissance, les données de localisation, les pages Facebook visitées/likées, Cambridge Analytica disposait par ailleurs d'autres bases de données qu'elle aurait croisées pour dresser un profil individuel de chaque personne. Ce profilage a notamment été utilisé dans le cadre des élections américaines par l'équipe de campagne de Donald Trump.

2. https://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn
3. https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html

2 Publication de données respectueuse de la vie privée

Les données n'existent essentiellement que par deux grands moyens : les réseaux qui les font circuler et les bases de données qui permettent d'y accéder. Dans une telle architecture, trois types de systèmes qui peuvent susciter des préoccupations en matière de protection de la vie privée on été identifiés [ZZ07]. La figure 1.2 illustre ces systèmes.

2.1 Protection de la vie privée et base de données

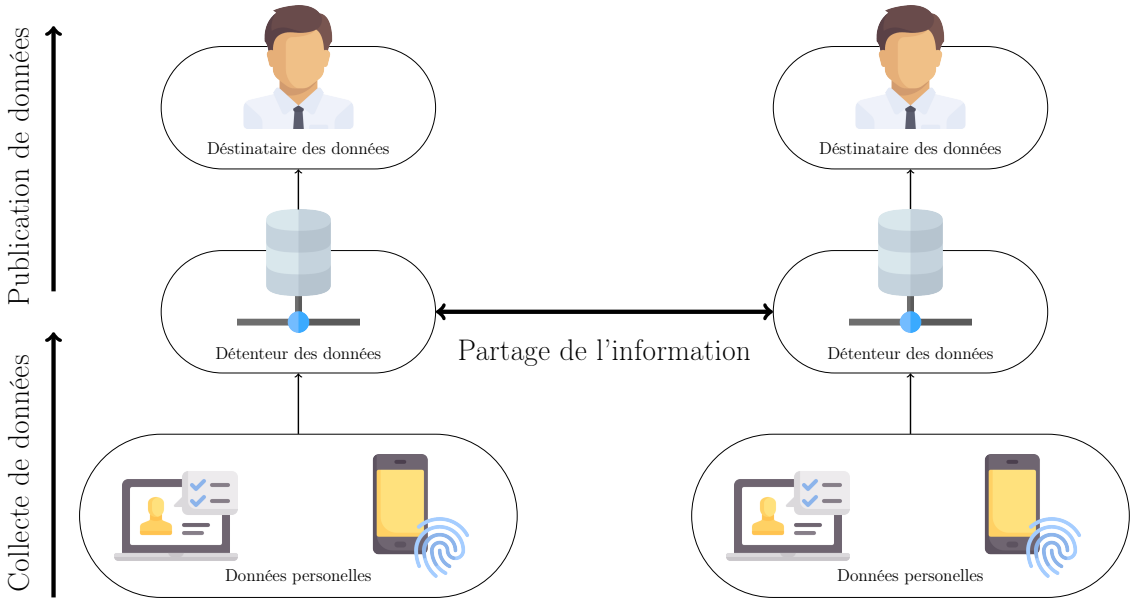


Figure 1.2 – L'architecture comporte généralement trois niveaux : les fournisseurs de données, le détenteur des données et le déstinataire des données. Le défi consiste à protéger l'information privée transmise entre les entités.

— **Partage de l'information**, ce système implique deux ou plusieurs parties qui ne se font pas confiance mutuellement. L'objectif est de garantir qu'aucune information privée au-delà du minimum nécessaire n'est divulguée lors de l'échange d'informations. Les techniques de calcul cryptographique sécurisé multipartite sont généralement utilisées pour ce type de systèmes [LP08].

— **Collecte de données**, système dans lequel un collecteur/analyste de données centralisé recueille et extrait les données de plusieurs fournisseurs de données répartis. Les techniques de perturbation aléatoire [AS00, HDC05] et cryptographiques [YZW05] sont habituellement appliquées pour protéger la vie privée dans

ces systèmes.

— **Publication de données**, dont l'objectif est de publier des données à des fins d'analyse sans compromettre l'anonymat des personnes représentées. Les mécanismes pour ce système visent à assainir les données originales ou les statistiques à publier en les masquant [Swe02b, MKGV07, LL07, Dwo06a].

Dans le cadre de cette thèse, nous nous intéressons au système de la *Publication de données*.

2.2 Motivation de la publication des données

Il y a une très forte demande économique et citoyenne pour l'ouverture des données que ce soit pour la recherche ou le marketing. Le secteur public en particulier, à travers ses instituts de statistique nationaux, de santé ou de transport, est soumis à des pressions pour mettre à disposition du public autant d'informations que possible au nom de la transparence. Les entreprises privées sont également concernées par la valorisation de leur données à travers l'échange ou la publication. Orange a ainsi récemment mis à disposition de la communauté scientifique différents jeux de communications mobiles collectées sur ses réseaux de Côte d'Ivoire ou du Sénégal, dans le cadre des challenges D4D (Data for Developement) dans un objectif de service à des projets de développement ou d'amélioration de politiques publiques [BEC+12].

3 Règlement général sur la protection des données

Le règlement n° 2016/679, dit règlement général sur la protection des données (RGPD) est une loi en application depuis mai 2018, qui vient remplacer la directive 95/46/CE sur la protection des données personnelles (datant de 1995). Elle s'applique chaque fois que des données à caractère personnel concernant des résidents européens sont traitées (y compris collectées, transformées, consultées ou effacées), au sein ou en dehors de l'Union. Les données personnelles représentent toute information se rapportant à une personne physique identifiée ou identifiable (ci-après dénommée «personne concernée»), est réputée

être une «personne physique identifiable» une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale.

Les personnes concernées ont des droits sur les données à caractère personnel, tels que le droit d'accès, l'effacement, l'opposition au traitement et la portabilité de leurs données dans un format structuré, couramment utilisé et lisible par une machine. Les responsables du traitement des données sont soumis à une série d'obligations, telles que la confidentialité, la notification en cas de violation des données et la réalisation d'évaluations des risques. De plus, ils ne doivent traiter les données que s'ils ont une base légale - comme le consentement - pour le faire, ou dans un but précis et limité, et pour une période de stockage limitée.

Les obligations issues du RGPD induisent un changement profond dans la manière de gérer les données personnelles et de les protéger, et l'un des principaux changements apportés à ce règlement par rapport aux précédents est la pénalité : les entreprises en infraction pourraient se voir infliger une amende allant jusqu'à 4% du chiffre d'affaires annuel global ou 20 millions d'euros.

3.1 RGPD et apprentissage automatique

La loi sur la protection des données régit déjà la collecte et l'utilisation des données dans le but de construire des modèles d'apprentissage automatique [VBE18] et, dans certaines conditions limitées, l'application des résultats des modèles aux personnes concernées. Par exemple, (i) les modèles ne peuvent être formés à partir de données personnelles sans un motif légal spécifique, tel que le consentement, le contrat ou l'intérêt légitime ; (ii) les personnes concernées devraient être informées de l'intention de construire un modèle et (iii) conservent généralement un droit d'objection ou de retrait du consentement ; (iv) dans les situations où les modèles informent une décision importante et uniquement automatisée, les personnes peuvent faire appel au responsable du traitement pour obtenir

des informations significatives sur la logique du traitement ou faire reconsidérer manuellement la décision prise. L'utilisation de l'apprentissage automatique pour transformer des données personnelles "normales" en données personnelles "de catégorie spéciale", telles que l'ethnie, l'opinion politique ou les données relatives à la santé, nécessite également l'établissement d'une base légale, qui sera généralement plus stricte que celle établie dans le cadre des données personnelles en général.

3.2 RGPD et données anonymes

L'anonymisation de données personnelles consiste à modifier le contenu ou la structure de ces données afin de rendre très difficile ou impossible la « ré-identification » des personnes.

En dépit de la restriction apportée par le RGPD sur l'exploitation des données personnelles en utilisant l'apprentissage automatique, une exception est faite pour les données rendues anonyme. En effet, la raison 26 du Règlement général sur la protection des données stipule que : *"...Les principes de protection des données ne devraient donc pas s'appliquer aux informations anonymes, à savoir les informations qui ne concernent pas une personne physique identifiée ou identifiable ou les données à caractère personnel rendues anonymes de telle sorte que la personne concernée n'est pas ou plus identifiable."*

C'est donc dans ce contexte, où les acteurs privés et gouvernementaux sont soumis à des obligations quant à l'exploitation de leur données personnelles, que la communauté de recherche s'intéresse à l'anonymisation de données.

4 Problématiques et solutions envisagées

Les contributions de cette thèse visent à apporter des solutions pour l'anonymisation de microdonnées en vue de leur publication, tout en assurant des garanties de protection. Le G29, qui regroupe les autorités de protection des données européennes, évalue la robustesse d'une technique d'anonymisation sur la base de trois critères :

— **L'individualisation** : est-il toujours possible d'isoler un individu dans les données

anonymisées ?

— **La corrélation** : est-il possible de relier entre eux des ensembles de données distincts concernant un même individu ?

— **L'inférence** : peut-on déduire de l'information sur un individu ?

Les deux premières contributions de la thèse visent à maitriser les risques de ré-identification et fournissent de ce fait, des solutions contre l'individualisation et la corrélation.

1. Contribution 1 : afin de dissocier les données et les utilisateurs réels, une solution consiste à créer un ensemble de données synthétiques à partir des données originales, c'est-à-dire un ensemble fictif de données ayant les mêmes caractéristiques qu'un ensemble de données réelles, de telle sorte qu'il n'est pas possible de ré-identifier les enregistrements réels initiaux. La solution consiste à utiliser une technique de co-clustering pour agréger les données puis à utiliser le modèle produit pour générer une table de données synthétiques du même format que les données initiales,
2. Contribution 2 : elle vise à anonymiser les données selon le formalisme du k-anonymat. Nous utilisons là encore le co-clustering, nous proposons de plus, une solution pour la génération automatique de hiérarchies de généralisation.
3. Contribution 3 : elle assure une protection plus stricte et apporte une solution contre l'individualisation, la corrélation et l'inférence. Elle vise à générer des données synthétiques en respectant le formalisme de la confidentialité différentielle, par le biais d'une technique de co-clustering.

5 Organisation du manuscrit

Complété par cette introduction et une conclusion, ce manuscrit est structuré en 4 chapitres :

Le chapitre 2 dresse un état de l'art des solutions existantes pour l'anonymisation de microdonnées. Il présente les principaux formalismes de protection de la littérature :

d'un côté les approches qui visent à fondre chaque individu dans un groupe d'individus (k-anonymat, l-diversité et t-proximité), de l'autre, la confidentialité différentielle qui est basée sur l'ajout d'un bruit perturbateur aux données originales. Nous présenterons dans ce chapitre quelque outils existants permettant d'assurer ces formalismes. Enfin, sont présentés les principaux critères d'évaluation quantitative de l'anonymisation en terme d'utilité et de risque.

Le chapitre 3 est articulé autour de la première contribution. Après avoir rappelé le principe du co-clustering et décrit plus particulièrement l'approche MODL sur laquelle se base notre co-clustering, on présente la méthodologie d'anonymisation proposée. L'utilité des données synthétiques est évaluée sur différentes tâches de fouille. On évalue également le niveau de protection offert par les données publiées.

Le chapitre 4 est consacré à la deuxième contribution. On présente d'abord quelques méthodes existantes pour assurer le k-anonymat : d'un coté les approches non-perturbatrices basées sur le principe de généralisation, d'un autre, les méthodes perturbatrices telles que la microagrégation. Une solution pour la génération de hiérarchies de généralisation est présentée. Ensuite, une approche de k-anonymat hybride basée sur le co-clustering est décrite. Enfin, cette dernière est évaluée et comparée aux algorithmes de la littératures en mesurant l'utilité des données anonymisées.

Le chapitre 5 est consacré à la troisième contribution. Nous dressons un état de l'art de la publication de données différentiellement privée. Les solutions existantes partent d'un ensemble de données original et visent à publier soit, un ou plusieurs histogrammes, soit un ensemble de données synthétiques. L'algorithme DPCocGen, une stratégie de partitionnement en deux phases basée sur le co-clustering pour la génération de données synthétiques est proposé. Finalement, On montre par des expérimentations que la solution surpasse les approches existantes et on montre que les données synthétiques produites préservent suffisamment d'informations et peuvent être utilisées pour plusieurs tâches de fouille de données.

La Figure1.3 résume l'organisation de ce manuscrit et les contributions de la thèse.

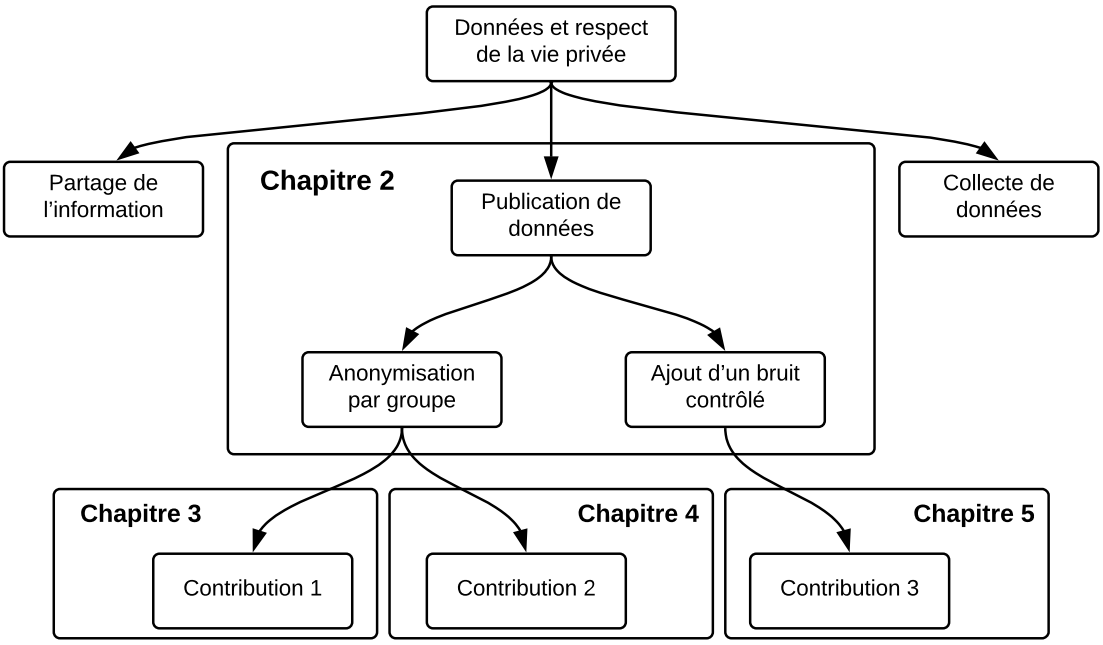


Figure 1.3 – Organisation de la thèse

Chapitre 2

État de l'art

Sommaire

1 Introduction

La taille et la diversité des informations numériques collectées par les gouvernements et les organisations publiques et privées sont en augmentation exponentielle. Ainsi, il existe une forte demande pour la publication de données que détiennent ces différents acteurs. Cependant, les données dans leur forme originale contiennent généralement des informations sensibles sur les personnes, et les dévoiler constitue une menace pour la vie privée des individus.

Nous nous intéressons à la publication de données personnelles. Un scénario classique de collecte et de publication de données est illustré dans la figure 2.1. Dans la *phase de collecte*, le *détenteur des données* recueille des informations auprès des *individus*. Lors de

la *phase de publication*, le *détenteur des données* fournit ces données à une tierce partie qu'on appellera *destinataire des données*.

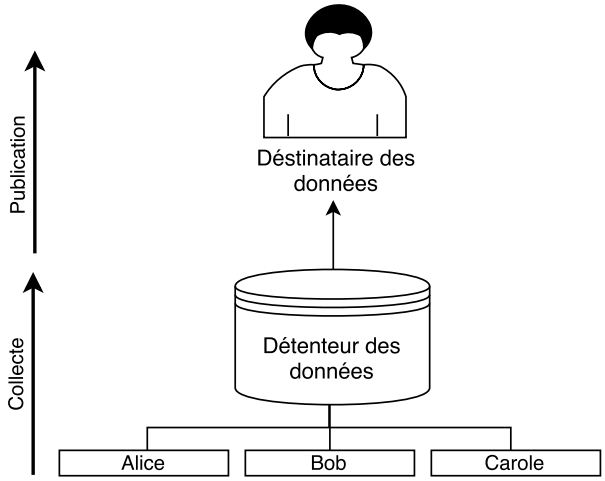


Figure 2.1 – Collecte et publication de données [FWCY10]

Il existe deux modèles de publication de données [Geh06]. Dans le modèle *malhonnête*, le détenteur des données n'est pas digne de confiance et peut tenter de divulguer des informations sensibles concernant les *individus*. Plusieurs approches : cryptographique [YZW05], communication anonyme [JJR02], et méthodes statistiques [War65] ont été proposées dans le but de collecter les informations personnelles de manière anonyme sans compromettre l'identité des *individus*. Dans le modèle *honnête*, le *détenteur des données* est digne de confiance et les *individus* sont disposés à lui fournir leurs informations personnelles. Cependant, la confiance n'est pas transitive au *destinataire des données*. Dans le cadre de cette thèse, nous nous intéressons exclusivement au modèle de publication de données *honnête* et nous traitons des problèmes de vie privée qui surviennent lors de la *phase de publication*.

Le domaine de recherche de la *Publication de Données Respectueuse de la Vie Privée* et celui du *Contrôle de la Divulgation Statistique* fournissent des mécanismes qui permettent la publication de données tout en préservant la vie privée des individus. Ces mécanismes visent à assainir les données originales ou les statistiques à publier en les masquant. Cependant, ce processus réduit non seulement les risques de divulgation, mais aussi l'utilité des données assainies, ainsi, le compris entre protection et perte d'information représente l'enjeu majeur dans de telles approches. Ce chapitre introduit les connaissances de base

nécessaires à la compréhension des contributions de cette thèse. Nous dressons un panorama des différentes méthodes d'anonymisation, de quelques formalismes de protection ainsi que des mesures d'utilité et de risque qui ont été proposées dans la littérature.

2 Protection des données

2.1 Types de données

Le type de données publiées détermine les potentielles menaces, ainsi que les formalismes de protection les plus adaptés. On distingue trois formats pour la publication de données statistiques [DFSSC16] :

- **Microdonnées.** Un ensemble d'enregistrements, où chaque enregistrement comporte de l'information sur un individu spécifique (personne physique ou morale).
- **Données tabulées.** Les données publiées ne font pas référence à une seule personne, mais à un groupe d'individus. Ce format est largement utilisé par les instituts de statistiques nationaux pour la publication de statistiques officielles, on retrouve des représentations telles que les tableaux de fréquences et d'amplitudes. Les menaces pour la vie privée des individus sont diminuées par rapport aux *Microdonnées* car seuls des agrégats sont publiés, mais l'analyse des données est limitée aux valeurs publiées.
- **Base de données dynamique.** Les *Microdonnées* et les *Données tabulées* offrent une vue statique de l'ensemble des données recueillies. Néanmoins, un destinataire des données spécifique peut ne pas être intéressé par l'intégralité des données, mais uniquement par un sous-ensemble d'entre elles. L'idée des bases de données dynamiques (ou configuration interactive), est que l'utilisateur est autorisé à soumettre des requêtes à la base de données, et les données sont seulement fournies pour les requêtes soumises.

Dans cette thèse, nous nous focalisons sur la publication de microdonnées. Les microdonnées sont le format le plus flexible, car elles ne restreignent pas les destinataires à une vue spécifique, et offrent à ces derniers la possibilité d'effectuer une analyse personnalisée

sur les données. Elles constituent cependant le type de données le plus vulnérable aux attaques.

2.2 Microdonnées

Un ensemble de microdonnées peut être modélisé comme une table où chaque ligne se réfère à un individu différent et chaque colonne contient des informations sur l'un des attributs collectés. Nous utilisons la notation T $(A_1, ..., A_n)$ pour désigner un ensemble de microdonnées avec des informations sur les attributs $A_1, ..., A_n$. Les attributs d'un ensemble de microdonnées peuvent être aussi bien numériques que catégoriels et, sont généralement classés comme suit :

- **Identifiant.** Tout attribut qui identifie uniquement un individu de manière explicite : Numéro de sécurité sociale, adresse ou nom...etc
- **Quasi-Identifiant(QI).** Ensemble d'attributs dont les valeurs prises conjointement peuvent identifier un individu : code postal, date de naissance et sexe...etc
- **Attributs sensibles.** Ces attributs contiennent une information sensible concernant les individus dans l'ensemble de données et leur divulgation pourrait conduire à des discriminations : salaire, opinion politique ou religieuse...etc
- **Attributs non-sensibles.** Tout attribut qui ne fait partie d'aucune des trois catégories précédentes.

La solution la plus banale pour protéger les micodonnées serait de supprimer les attributs de type identifiant avant de publier les tables. Cependant, cette méthode est insuffisante pour préserver la vie privée des individus concernés. Sweeney [Swe02b] l'illustre dans son étude en croisant deux bases de données, une base de données médicale dont on a supprimé l'attribut identifiant et une liste électorale avec des données publiques. Le croisement a été effectué sur la base du triplet de valeurs {code postal, date de naissance, sexe} et a permis de relier des données médicales à des individus parmi lesquels se trouvait le gouverneur de l'état du Massachusetts de l'époque. Sweeney estime qu'aux États-Unis 87% de la population peut être identifié de manière unique en se basant seulement sur le quasi-identifiant {code postal, date de naissance, sexe}. Cet exemple est illustré Fi-

gure 2.2.

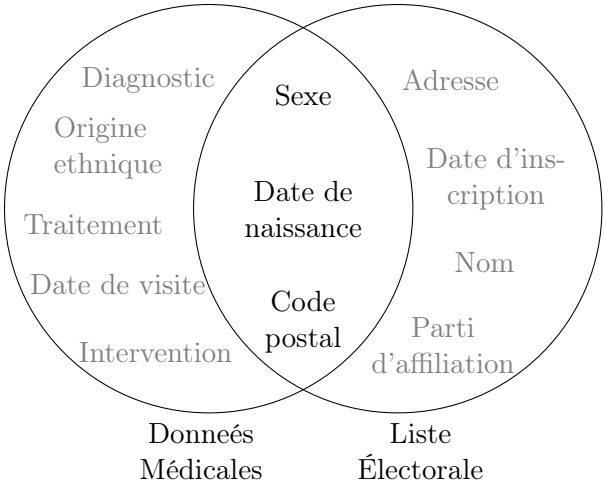


Figure 2.2 – Attaque par liaison et ré-identification des individus [Swe02b]

Dans l'exemple ci-dessus, l'individu est ré-identifié en liant son quasi-identifiant provenant d'une source de connaissance externe (une base de données publique) à l'enregistrement correspondant dans la table de données publiée. Une telle attaque est appelée *attaque de ré-identification*. Dès lors qu'un individu est réidentifié, ses attributs sensibles sont révélés. De ce fait, les méthodes d'anonymisation se basent sur le QI et cherchent à réduire le niveau de précision des informations publiées.

2.3 Paradigmes de protection de la vie privée

Né de l'intérêt des organismes de statistiques nationaux pour la diffusion des données qu'ils produisent, sous une forme qui limite les risques de divulgation, le *Contrôle de la Divulgation Statistique* (ou SDC) est un domaine auquel s'est initialement intéressée la communauté de recherche en statistiques [AW89] [Ski09], et qui traite principalement des données agrégées.

Plus récemment, la communauté de recherche en informatique, s'est elle aussi penchée sur la problématique en l'intitulant *Publication de Données Respectueuse de la Vie Privée* (ou PPDP) [FWCY10]. Celle-ci se focalise sur la publication de *microdonnées* et offre ainsi aux destinataires des données plus de liberté.

Même si les deux paradigmes partagent le même objectif, historiquement, les approches développées pour la protection de la vie privée dans les données, diffèrent selon

la communauté [Dre11]. En effet, la communauté statistique s'intéresse principalement à la validité statistique des données publiées, et les méthodes développées par celle-ci ne donnent pas de garanties quant à la protection, le niveau de protection est déterminé *a posteriori*, et est spécifique à chaque jeu de données. En revanche, les méthodes développées par la communauté informatique tentent de fournir des garanties formelles *a priori*, à travers la définition de formalismes de protection stricts.

De nos jours, les deux paradigmes tendent à se rapprocher [DM15], plusieures méthodes issues du *Contrôle de la Divulgation Statistique* peuvent être utilisées afin de satisfaire des formalismes de protection introduits par la *Publication de Données Respectueuse de la Vie Privée* assurant de ce fait, des garanties de *protection a priori*.

2.4 Risques de divulgation sur les microdonnées

Afin de protéger les microdonnées des menaces de divulgation, les détenteurs ne publient pas la table originale T mais une version modifiée T', T' est une version *assainie* ou *anonymisée* de T. Cependant, dès lors qu'elles sont publiées, ces données sont vulnérables à plusieurs menaces. Dans la communauté de la *Publication de Données Respectueuse de la Vie Privée* [FWCY10] les menaces suivantes sont considérées :

- **La divulgation d'identité**. Survient lorsque l'adversaire est en mesure d'associer un individu connu à un enregistrement dans la table publiée T'. Par exemple, l'attaque de ré-identification où l'adversaire peut déterminer l'appartenance de la victime à un groupe d'enregistrements dans T', et est capable en ayant recours à une connaissance externe d'identifier de manière non ambiguë, l'enregistrement de la victime dans le groupe.
- **La divulgation d'attribut**. Dans cette attaque, l'adversaire peut ne pas être en mesure d'identifier précisément l'enregistrement de sa victime. Il peut en revanche inférer la valeur de son attribut sensible dans la table publiée T', en se basant sur l'ensemble de valeurs sensibles associées au groupe auquel appartient l'enregistrement de la victime dans T'. Par exemple, si un hôpital publie des données qui montrent que toutes les patientes âgées de 56 à 60 ans ont un cancer, un adversaire

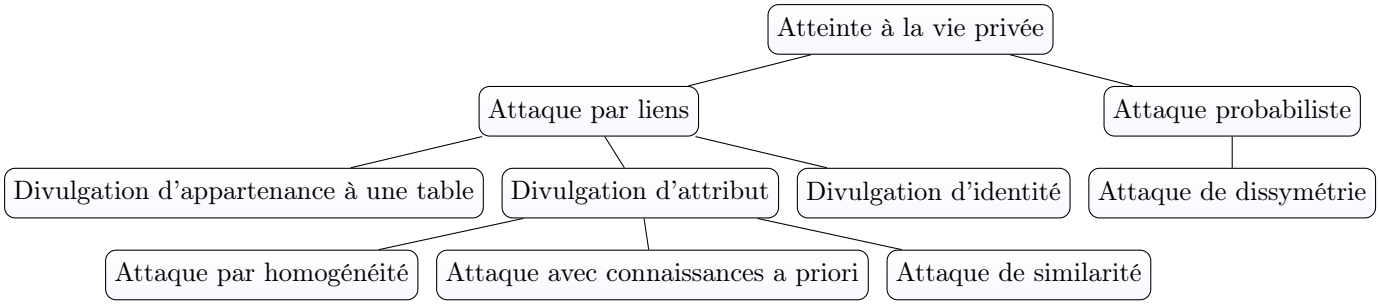


Figure 2.3 – Une typologie des attaques à la vie privée [Fre17]

déduit le diagnostic d'une femme âgée de 58 ans sans avoir à identifier l'individu en particulier.

- **La divulgation d'appartenance à une table**. Les deux scénarios d'attaques précédents supposent que l'adversaire sait que l'enregistrement de la victime appartient à la table publiée T'. Seulement, dans certains cas la présence (ou l'absence) de l'enregistrement de la victime dans la table T' révèle à elle seule l'information sensible de la victime. Cette attaque survient lorsqu'un adversaire est en mesure d'inférer de manière certaine la présence ou l'absence de l'enregistrement de la victime dans T'.
- **L'attaque probabiliste**. On dit qu'une telle attaque survient si l'adversaire modifie ses croyances probabilistes sur l'information sensible d'une victime après avoir accédé aux données publiées. Par exemple, un adversaire, en se basant sur ses connaissances a priori, et ce avant de consulter la table publiée, sait que sa victime est atteinte du VIH avec une probabilité de 0.1. Si en consultant la table publiée, l'adversaire peut déduire que la victime est atteinte de VIH avec un probabilité de 0.8, alors l'adversaire a réussi une attaque probabiliste, la différence entre les probabilités *a priori* (0.1) et *a posteriori* (0.8) est significative.

A ces familles d'attaques à la vie privée, s'ajoutent des attaques plus spécifiques que nous aborderons tout au long de la Section 4. La Figure 2.3 [Fre17] montre une typologie de l'ensemble des attaques traitées.

3 Méthodes pour l'anonymisation de microdonnées

le *Contrôle de la Divulgation Statistique* et la *Publication de Données Respectueuse de la Vie Privée* fournissent plusieures méthodes pour l'assainissement de microdonnées. Ces méthodes peuvent être classées en deux catégories [DFSSC16] les méthodes d'assainissement, et les méthodes de génération de données synthétiques.

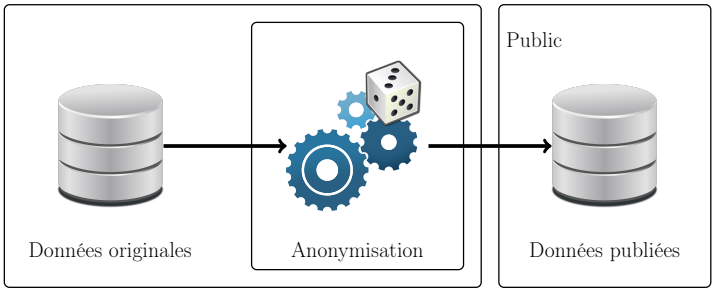


Figure 2.4 – Publication de données

3.1 Assainissement

Les méthodes d'assainissement génèrent les données anonymisées T' en modifiant les enregistrements originaux dans T. Les méthodes d'assainissement sont à leur tour divisées en deux catégories, en fonction de leur effet sur les données originales [DFSSC16] : les *méthodes non perturbatrices* et les *méthodes perturbatrices*.

— *Méthodes non perturbatrices.* Les méthodes d'assainissement non perturbatrices n'altèrent pas les données. Elles opèrent une suppression ou une réduction partielle des détails dans les données originales. On peut citer des méthodes telles que la généralisation [Swe02a] et la suppression locale [CFM+13].

— *Méthodes perturbatrices.* L'ensemble de microdonnées est altéré avant sa publication. Les méthodes utilisées doivent être telles que les statistiques calculées sur les données perturbées ne soient pas significativement différentes des statistiques obtenues à partir des données originales. On retrouve dans cette catégorie des méthodes telles que l'ajout de bruit [Bra02], la micro-agrégation [ODf01], la permutation de données [Rei82], et le re-échantillonnage [MSV12].

3.2 Génération de données synthétiques

Une alternative aux méthodes de protection décrites précédemment est la génération de données synthétiques où tout ou une partie des données originales est remplacée par des données synthétiques. Cette technique est maintenant communément adoptée par les instituts de statistiques qui publient des jeux de données synthétiques (enquêtes sur les revenus ou les ménages) sous forme d'agrégats, de résumés ou de tableaux croisés [BSA13], [KRR+11], plus rarement sous forme de données individuelles [LZHH13]. On distingue trois types de données synthétiques :

— *Entièrement synthétique* [Rub93], où chaque valeur d'attribut de chaque enregistrement a été synthétisée.

— *Partiellement synthétique* [Rei03], où seules les valeurs d'attributs avec un risque de divulgation élevé sont synthétisées.

— *Hybride* [MS08, DFGN10], où l'ensemble original de données est mélangé avec un ensemble de données entièrement synthétique.

Il existe différentes manières de générer les données synthétiques :

— soit en construisant un modèle générateur à partir des données originales ; on trouve dans cette catégorie des modèles de type SVM [Dre10], forêts aléatoires [CR10] ou encore imputations multiples [RD10],

— soit en se basant sur la connaissance d'un expert du domaine analysé qui propose des règles de modélisation et spécifie les contraintes entre attributs [JLR+06, ARPDMT16].

On pourra se reporter à [SM17] pour une synthèse détaillée de techniques de génération de données synthétiques. La qualité des modèles générateurs est primordiale et les caractéristiques des données synthétiques qui en sont issues doivent s'approcher du mieux possible des données réelles pour être utiles. On distingue les mesures d'utilité générales et les mesures spécifiques. Les mesures générales s'attachent à produire des résumés de comparaisons entre les distributions réelles et les distributions synthétiques. Ces mesures sont adaptées quand on ne connait pas l'usage qui sera fait des données synthétiques publiées. Les mesures spécifiques permettent de comparer des résultats d'analyses entre

données originales ou synthétiques ou d'évaluer les paramètres des modèles utilisés pour les analyses. Dans ce cas, les données synthétiques sont jugées utiles si les inférences à partir des données originales et synthétiques correspondent. On trouvera une discussion sur les métriques d'utilité spécifiques aux données synthétiques dans [SRN+18].

4 Formalismes de protection

De nombreux formalismes de protection ont été proposés dans la littérature de la *Publication de Données Respectueuse de la Vie Privée*. Nous distinguons deux grandes familles de formalismes : d'un côté les approches qui visent à fondre chaque individu dans un groupe d'individus [Swe02b, MKGV07, LL07] de l'autre celles qui sont basées sur l'ajout d'un bruit perturbateur aux données originales [Dwo06a]. Nous allons passer en revue quelques formalismes de protection populaires, nous invitons les lecteurs intéressés à consulter [FWCY10] et [dVFLS12] pour une analyse en profondeur.

4.1 *k*-Anonymat

L. Sweeney a proposé dans [Swe02b], un formalisme de protection qui permet de protéger les données des attaques de ré-identification. L'intuition du k-anonymat est de cacher un individu dans la foule et ainsi brouiller le lien entre les individus et leurs enregistrements respectifs. Une table satisfait le k-anonymat si chaque enregistrement de la table ne peut être distingué d'au moins $k-1$ autres enregistrements.

Définition 2.1 (k-Anonymat)**.** Une table T' satisfait le k-anonymat si pour chaque enregistrement $t \in T'$ il existe $k-1$ autres enregistrements qui partagent les mêmes valeurs pour tous les attributs du quasi-identifiant.

Le k-anonymat assure la garantie qu'un individu ne peut pas être réidentifié avec une probabilité supérieur à $\frac{1}{k}$.

La table 2.1a représente une table de données de patients que pourrait détenir un établissement hospitalier par exemple. La table 2.1b est un exemple d'une version 3-

anonyme de la table 2.1a, chaque enregistrement est identique en termes de valeurs du quasi-identifiant ⟨Code Postal, Age⟩ à au moins deux autres enregistrements.

Définition 2.2 (Classe d'équivalence)**.** La classe d'équivalence d'un enregistrement $t \in T$ est l'ensemble des enregistrements dans T qui partagent avec t les valeurs de tous les attributs du QI. En utilisant la notion de classe d'équivalence, la définition du k-anonymat peut être reformulée comme suit : un ensemble de données est k-anonyme si et seulement si la classe d'équivalence de chaque enregistrement contient au moins k enregistrements.

	Code Postal	Age	Diagnostic
1	23677	29	Hypertension
2	23602	22	Hypertension
3	23678	27	Hypertension
4	23905	43	Arythmie
5	23909	52	Hypertension
6	23906	47	Athérosclérose
7	23605	36	Hypertension
8	23673	36	Cancer
9	23607	32	Cancer

(a) Table originale des patients

	Code Postal	Age	Diagnostic
1	236**	2*	Hypertension
2	236**	2*	Hypertension
3	236**	2*	Hypertension
4	2390*	≥ 40	Arythmie
5	2390*	≥ 40	Hypertension
6	2390*	≥ 40	Athérosclérose
7	236**	3*	Hypertension
8	236**	3*	Cancer
9	236**	3*	Cancer

(b) Une version 3-anonyme de la table 2.1a

Tableau 2.1 – Tables patients

Le formalisme de protection k-anonymat, définit les conditions que les données publiées doivent respecter pour que le risque de divulgation soit maitrisé. Cependant, il ne spécifie pas la marche à suivre pour atteindre de telles conditions. On distingue principalement deux types d'approches pour assainir les données au sens du k-anonymat :

- **Généralisation et Suppression.** Utilisée dans la première méthode pour atteindre le k-anonymat [Sam01a] puis plus tard dans des approches comme [LDR05, LDR06a, BA05], cette technique est une combinaison de la généralisation d'attribut et de la suppression locale. La généralisation réduit la granularité de l'information contenue dans les attributs du QI, augmentant ainsi le nombre d'enregistrements partageant les mêmes valeurs pour ces attributs. Une hiérarchie de généralisation doit être définie pour chaque attribut, qui sert de guide au procédé de généralisation. La suppression consiste à retirer des enregistrements.

- **Micro-agrégation.** Dans ce type d'approches [ODf01] ou encore [LM05] on commence par regrouper les enregistrements dans les données avec des techniques de clustering, de telle sorte que chaque cluster soit peuplé d'au moins k enregistrements et que les enregistrements d'un même cluster soient aussi similaires que possible. Les enregistrements de chaque cluster sont ensuite remplacés par un individu représentatif du cluster, typiquement, le centroide du cluster.

4.2 ℓ-Diversité et t-proximité

Le k-anonymat protège une table contre la divulgation de l'identité de ses individus, mais ne protège pas suffisamment les attributs sensibles qui ne font pas partie du quasi-identifiant.[MKGV07, TV06, XT06b] mettent en évidence cette faiblesse dans le modèle du k-anonymat, avec deux types d'attaques : l'*attaque par homogénéité* et l'*attaque avec connaissance a priori* [MKGV07].

— *L'attaque par homogénéité.* Prenons la table 2.1b des patients 3-anonyme. L'attribut Diagnostic est sensible. Supposons qu'*Alice* sache que *Bob* est un homme de 22 ans, que le code postal de son domicile est 23678 et que l'enregistrement de *Bob* appartient à la table. D'après la table 2.1b, *Alice* peut conclure que *Bob* correspond à l'un des trois premiers enregistrements, et qu'il souffre d'hypertension avec certitude.

— *L'attaque avec connaissance a priori.* Supposons qu'en connaissant l'âge et le code postal de *Carole*, alors *Alice* peut conclure que *Carole* correspond à un enregistrement dans la dernière classe d'équivalence de la table 2.1b. En outre, supposons qu'*Alice* sache que *Carole* a un très faible risque de maladie cardiaque. Cette connaissance de base permet à *Alice* de conclure que *Carole* est probablement atteinte d'un cancer.

Pour remédier à ces limitations du k-anonymat, Machaanavajjhala et al. [MKGV07] ont introduit la ℓ-diversité comme un formalisme de protection plus fort.

Les auteurs ont d'abord formulé le formalisme de protection bayésien-optimal pour mettre en évidence l'impact des données publiées sur la croyance de l'adversaire. En pre-

mier lieu, *la croyance a priori* de l'adversaire est modélisée comme étant la distribution jointe f des valeurs sensibles et des valeurs du QI pour l'ensemble de la population Ω. Étant donnée cette distribution, la croyance a priori de l'adversaire concernant la vraisemblance de l'association d'un quasi-identifiant donné q à une valeur sensible s est la probabilité conditionnelle d'observer l'association entre q et s. Ensuite, ils calculent la croyance de l'adversaire *influencée* par les données publiées, appelée la *croyance a posteriori*. Enfin, ils considèrent qu'une *divulgation* se produit si on peut observer un changement significatif entre les croyances a priori et a posteriori de l'adversaire. Cette définition de la divulgation est appelée principe de *non-informativité*. Le principe tire ses origines dans les premières définitions de la divulgation statistique données par Dalenius [Dal77] et est à la racine de nombreux formalismes de protection [MKGV07, LLV10].

Définition 2.3 (Principe de non-informativité [MKGV07])**.** Les données anonymisées publiées doivent fournir à l'adversaire très peu d'informations additionnelles par rapport à sa connaissance initiale. En d'autres termes, les croyances *a priori* et *a posteriori* ne doivent pas être trop différentes.

Le formalisme de protection bayésien-optimal présente plusieurs inconvénients en raison notamment du fait que l'éditeur ne connaît pas nécessairement la distribution jointe exacte entre les valeurs du quasi-identifiant et les valeurs sensibles. Le détenteur des données ne sait pas forcément ce que l'adversaire sait. En outre, il peut y avoir plusieurs adversaires avec diverses connaissances de base. Machaanavajjhala et al. [MKGV07] jugent que la protection bayésienne-optimale n'est pas réalisable en pratique, et proposent donc comme alternative un modèle appelé ℓ-Diversité.

Définition 2.4 (ℓ-diversité [MKGV07])**.** Une classe d'équivalence est ℓ-diverse s'il y a au moins ℓ valeurs "bien-représentées" pour l'attribut sensible. Une table est dite ℓ-diverse si chaque classe d'équivalence est ℓ-diverse.

Machaanavajjhala et al. [MKGV07] ont fourni au terme "bien-représenté" plusieurs interprétations :

1. **ℓ-diversité distincte.** La compréhension la plus simple de "bien-représenté" serait d'assurer qu'il y ait au moins ℓ valeurs *distinctes* pour l'attribut sensible dans chaque classe d'équivalence. la ℓ-diversité distincte ne protège pas des attaques probabilistes. Une classe d'équivalence pourrait avoir une valeur qui apparait beaucoup plus fréquemment que les autres, ce qui permettrait à un adversaire de conclure qu'une victime dans la classe d'équivalence est très susceptible d'avoir cette valeur. Ceci a motivé le développement de deux notions plus fortes de ℓ-diversité.

2. **ℓ-diversité avec entropie.** L'entropie d'une classe d'équivalence E est définie comme suit :

$$\text{Entropie}(E) = -\sum_{s \in S} p(E,s) \log p(E,s) \tag{2.1}$$

   Où S est le domaine de l'attribut sensible, et $p(E,s)$ est la fraction d'enregistrements dans E dont la valeur de l'attribut sensible est s.

   Une table respecte la ℓ-diversité avec entropie si pour chaque classe d'équivalence E, Entropie$(E) \geq \log \ell$. La ℓ-diversité avec entropie est plus forte que la ℓ-diversité distincte, mais comme l'ont souligné les auteurs dans [MKGV07], la ℓ-diversité avec entropie peut s'avérer trop stricte, si quelques valeurs de l'attribut sensible sont très communes, car dans ce cas, l'entropie des classes d'équivalence a tendance à être petite.

3. **(c,ℓ)-Diversité récursive.** La (c,ℓ)-diversité récursive s'assure que les valeurs les plus fréquentes n'apparaissent pas trop fréquemment, et que les valeurs les moins fréquentes n'apparaissent pas trop rarement. Soit m le nombre de valeurs dans une classe d'équivalence, et r_i, $1 \leq i \leq m$ est le nombre de fois que la $i^{\text{ème}}$ valeur sensible la plus fréquente apparait dans une classe d'équivalence E. Alors E est dite (c,ℓ)-diverse récursive si $r_1 \leq c(r_l + r_l + \ldots + r_m) \leq m$. Une table est dite ($c$,$\ell$)-diverse récursive si chacune des classes qui la compose est (c,ℓ)-diverse récursive.

La première méthode proposée [MKGV07] pour atteindre la ℓ-diversité exploite la

généralisation et suppression. Cette approche est cependant sujette à la malédiction de la dimensionnalité introduite dans [Agg05]. D'autres procédés pour la création de tables ℓ-diverses sont décrits dans [XT06a].

4.2.1 *t*-proximité

Bien que le formalisme de protection de la ℓ-diversité représente une avancée au-delà du k-anonymat en protégeant les tables contre les attaques de divulgation d'attribut, il a cependant quelques faiblesses que mettent en évidence Li et al. dans [LL07] :

— *L'attaque de dissymétrie.* Cette attaque exploite l'éventuelle différence entre la distribution des fréquences des valeurs de l'attribut sensible dans une classe d'équivalence, comparé à la distribution des fréquences des valeurs de l'attribut sensible dans toute la population (ou l'ensemble des microdonnées publiées). Une telle différence dans les distributions met en évidence des changements dans la probabilité qu'un individu appartenant à une classe d'équivalence soit associé à une certaine valeur d'attribut sensible. Considérons l'exemple suivant, dans la table2.1b, *Alice* sait que son ami Bob a 32 ans et vit aux environs du 23600. Dans la classe d'équivalence qui possède le QI $\langle 236^{**},3^{*}\rangle$, deux des trois enregistrements ont la valeur *Cancer* pour l'attribut Diagnostic. *Alice* peut donc déduire que *Bob* est atteint d'un cancer avec une probabilité de 66.66%, en comparaison avec la probabilité dans toute la table qui est de 22.22%.

— *L'attaque de similarité.* Cette attaque se produit lorsque, dans une table ℓ-diverse, les valeurs de l'attribut sensible associées aux enregistrements dans une classe d'équivalence sont sémantiquement similaires, même si elles sont syntaxiquement différentes. Par exemple, en considérant la table2.1b, supposons qu'*Alice* sache que sa voisine *Carole* est âgée de 65 ans et vit comme elle au 23900. La classe d'équivalence caractérisée par le QI $\langle 2390^{*},\geq 40\rangle$ possède les valeurs *Arythmie, Hypertension* et *Athérosclérose* pour l'attribut sensible Diagnostic. Par conséquent, *Alice* peut découvrir que *Carole* est atteinte d'une maladie cardiovasculaire.

Afin de protéger les tables des faiblesses soulevées dans le formalisme de protection

de la ℓ-diversité, Li et al. [LL07] ont introduit le modèle de la t-proximité.

Ils proposent que la distribution de l'attribut sensible dans l'ensemble de données entier soit considérée comme une source auxiliaire de *connaissances a priori* de l'adversaire. Selon le formalisme de protection t-proximité, un adversaire qui connait la distribution globale de l'attribut sensible dans la version publiée de la table gagne peu d'information sur une classe d'équivalence en apprenant la distribution de l'attribut sensible dans la table.

La t-proximité exige que la différence entre la distribution de l'attribut sensible dans une classe d'équivalence et la répartition globale de cet attribut sensible ne soit pas supérieure à un seuil donné t.

Définition 2.5 (t-Proximité [LL07])**.** Une classe d'équivalence respecte la t-proximité si la distance entre la distribution d'un attribut sensible dans cette classe et la distribution de l'attribut dans tout l'ensemble de données n'excède pas un seuil t. Un ensemble de données satisfait la t-proximité si toutes les classes d'équivalence respectent la t-proximité.

Le formalisme de t-proximité exploite une notion de distance entre les distributions, mais n'en préconise aucune. Cependant, la distance *earth mover's distance* (EMD) [RTG00] est une distance largement adoptée pour la t-proximité. L'avantage d'EMD est qu'elle est capable de capturer la distance sémantique entre les valeurs. $EMD(P, Q)$ calcule le coût de la transformation d'une distribution P en une distribution Q en déplaçant la masse de probabilité.

Comme pour le k-anonymat, le moyen le plus commun pour atteindre la t-proximité est d'utiliser la généralisation et suppression. Les algorithmes pour le k-anonymat basés sur ces principes peuvent être adaptés afin d'assurer la t-proximité en ajoutant la contrainte de la t-proximité dans le processus de recherche de la généralisation minimale. Dans [LL07], l'algorithme Incognito et dans [LLV10], l'algorithme Mondrian sont adaptés à la t-proximité.

4.3 Confidentialité Différentielle

Dans cette section, nous introduisons le principe fondamental de la confidentialité différentielle. Nous abordons ensuite quelques mécanismes qui permettent de l'atteindre. Enfin, nous présentons les propriétés de composition de plusieurs mécanismes respectueux de la confidentialité différentielle.

4.3.1 Définition

La confidentialité différentielle est un formalisme de protection robuste qui offre des garanties formelles quant à la protection des données. Celui-ci impose que quelle que soit l'information qui peut être extraite à partir des données publiées, *l'impact de la présence ou de l'absence d'un seul individu soit limité*. La confidentialité différentielle garantit notamment, que si un adversaire connaît toutes les informations de tous les enregistrements dans l'ensemble de données D mis à part un enregistrement, le résultat d'un algorithme probabiliste respectueux de la confidentialité différentielle ne doit pas fournir à l'adversaire trop d'informations additionnelles concernant l'enregistrement restant.

La confidentialité différentielle a été initialement proposée dans une configuration *dynamique* [Dwo06a], où un tiers de confiance dispose d'une base de données à laquelle sont soumises les requêtes des utilisateurs. L'hypothèse est qu'il existe un algorithme d'anonymisation entre l'utilisateur qui soumet sa requête et le tiers de confiance qui y répond. Afin de préserver la vie privée des individus, la suppression ou l'ajout d'un enregistrement dans un ensemble de données n'affecte pas significativement le résultat de la requête (Figure 2.5).

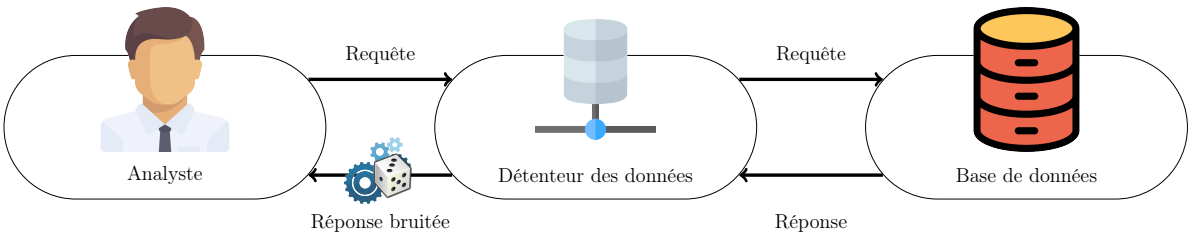


Figure 2.5 – L'architecture dynamique

Dans une configuration *statique* [Dwo10], le tiers de confiance prépare un agrégat ou

une base de données synthétique, destinés à être publiés pour répondre à différents types de requêtes (Figure 2.6).

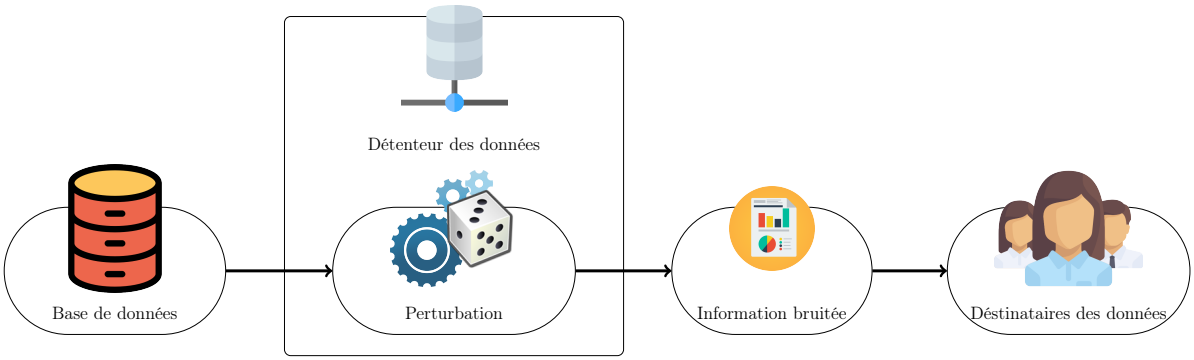


Figure 2.6 – L'architecture statique

Définition 2.6 (ε-Confidentialité différentielle [Dwo06a])**.** Un algorithme probabiliste $\mathcal{A}$ satisfait la ε-confidentialité différentielle si pour deux ensembles de données voisins D_1 et D_2, et pour tout résultat possible O de $\mathcal{A}$, $Pr[\mathcal{A}(D_1) = O] \leq e^{\varepsilon} \times Pr[\mathcal{A}(D_2) = O]$.

Deux notions de *voisinage* sont retrouvées dans la littérature : dans [Dwo06b], deux ensembles de données sont dits voisins, si un ensemble de données peut être obtenu à partir de l'autre en ajoutant ou supprimant un enregistrement. Dans [Nis08], deux ensembles de données sont dits voisins, si un ensemble de données peut être obtenu à partir de l'autre en modifiant un enregistrement. Nous nous baserons dans ce travail sur la première notion de voisinage.

Le paramètre ε représente le budget de confidentialité. Plus ε est petit, plus le ratio $\frac{Pr[\mathcal{A}(D_1)=O]}{Pr[\mathcal{A}(D_2)=O]}$ est proche de 1. Ce qui revient à dire que les distributions de probabilité du résultat de l'algorithme $\mathcal{A}$ sur les ensembles voisins D_1 et D_2 sont approximativement égales. ε représente une mesure relative. Dans [LC11], il est montré que pour la même valeur de ε, les garanties de protection de la vie privée offertes par la ε-confidentialité différentielle, varient en fonction du domaine d'attribut en question et de la requête prise en charge. En pratique, la question du choix de ε s'avère difficile et demeure aujourd'hui un défi.

4.3.2 Mécanismes pour assurer la confidentialité différentielle

Mécanisme Laplacien Le mécanisme le plus répandu pour assurer la confidentialité différentielle est le mécanisme Laplacien, ce dernier fonctionne en ajoutant un bruit aléatoire à la réponse à une requête. D'abord, la vraie valeur de $f(D)$ est calculée, où f est la fonction de requête et D l'ensemble de données, ensuite un *bruit* aléatoire est ajouté à $f(D)$ et la réponse $\mathcal{A}(D) = f(D) + bruit$ est finalement retournée. L'amplitude du bruit est choisie en fonction du plus grand changement que peut provoquer un enregistrement sur la sortie de la fonction requête (par exemple, cela correspondrait à 1 pour une requête de dénombrement à travers D_1 et D_2), cette quantité définie par C. Dwork est appelée *sensibilité* de la fonction.

Définition 2.7 (L_1-sensibilité globale)**.** La L_1-sensibilité globale de $f : D \to \mathbb{R}^d$ est

$$\Delta(f) = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1 \tag{2.2}$$

Pour deux ensembles de données voisins D_1 et D_2.

La fonction de densité de la distribution Laplacienne est définie comme suit

$$Lap(x|\mu, b) = \frac{1}{2b} \exp(\frac{-|x - \mu|}{b}) \tag{2.3}$$

Où le réel μ est appelé paramètre de position et $b > 0$ le paramètre d'échelle.

L'utilisation d'un bruit issu d'une distribution Laplacienne, $bruit = Lap(\Delta f/\varepsilon)$, avec le paramètre de position $= 0$, et le paramètre d'échelle $= \Delta f/\varepsilon$ garantit la ε-confidentialité différentielle [NRS07a].

Exemple : Sensibilité

Soient deux requêtes f_1 et f_2, telles que f_1 = 'Le nombre d'individus atteints d'hypertension' et f_2 = 'La somme des âges des individus de l'ensemble de données'. Supposons que $age \in [0, 130]$. Nous avons alors, $\Delta(f_1) = 1$ et $\Delta(f_2) = 130$.

Mécanisme Exponentiel Le mécanisme Laplacien a été mis au point pour les requêtes dont les résultats sont numériques. Pour les requêtes dont les résultats ne sont pas numériques [MT07], ont proposé le mécanisme exponentiel. Celui-ci est basé sur une fonction d'utilité qui évalue l'utilité de chaque résultat possible à une requête, puis sélectionne une sortie $t \in \mathcal{T}$ qui est proche de l'optimum (au sens de la fonction d'utilité), tout en respectant la confidentialité différentielle.

Définition 2.8 (Mécanisme Exponentiel [MT07])**.** Pour une base de données D, un ensemble de résultats possibles $\mathcal{T}$, un budget de confidentialité ε, une fonction d'utilité $u : D \times \mathcal{T} \to \mathbb{R}$ qui assigne une valeur réelle (score) à chaque sortie $t \in \mathcal{T}$, où une valeur importante représente une meilleure utilité. Un algorithme probabiliste qui choisit une sortie $t \in \mathcal{T}$ avec un probabilité $\exp(\frac{\varepsilon u(D,t)}{2\Delta u})$ garantit la ε-confidentialité différentielle.

$\Delta u = \max_{\forall t, D_1, D_2} |u(D_1, t) - u(D_2, t)|$ désigne la sensibilité de la fonction utilité, elle représente la variation maximum de u à travers deux ensembles de données voisins D_1 et D_1, quel que soit t.

Le mécanisme tire une sortie t, en construisant une distribution de probabilité à travers l'ensemble de sortie $\mathcal{T}$. La probabilité associée à chaque sortie t augmente proportionnellement à $\exp(\frac{\varepsilon u(D,t)}{2\Delta u})$, ainsi, la sortie avec un plus grand score a plus de probabilité d'être retournée.

Exemple

Soit la requête quelle-est la nationalité la plus commune dans l'ensemble de données D ? à partir de $\mathcal{T}$ =$\{$Chinoise, Indienne, Américaine, Grecque$\}$. La fonction $u(D, t)$ = nombre d'individus dans D qui possèdent la nationalité t, pourrait être envisagée comme fonction d'utilité. La sensibilité serait donc $\Delta u = 1$. Le mécanisme retourne une nationalité qui est partagée par K individus avec une probabilité $\exp(\frac{\varepsilon K}{2})$.

Mécanisme géométrique Le mécanisme géométrique [GRS12] a été proposé dans le cas d'une requête dont le résultat est entier. Au lieu d'ajouter un bruit réel comme le fait le mécanisme Laplacien, le mécanisme géométrique $\mathcal{A}(D) = f(D) + X$ ajoute un bruit

$X \in \mathbb{Z}$ au résultat de la requête $f(D)$. X est tiré à partir de la distribution géométrique :

$$Pr[X = x] = \frac{1+\alpha}{1-\alpha}\alpha^{|x|} \tag{2.4}$$

L'utilisation d'un bruit issu d'une distribution geométrique, où $\alpha = e^{-\varepsilon/\Delta f}, x \in \mathbb{Z}$, garantit la ε-confidentialité différentielle [GRS12].

4.3.3 Relaxation de la confidentialité différentielle

(ε,δ)-Confidentialité différentielle La solide garantie apportée par la ε-confidentialité différentielle se paye par l'ajout d'un bruit conséquent au résultat des requêtes et des analyses, [DKM$^+$06] proposent la (ε,δ)-confidentialité différentielle pour améliorer l'utilité des données en réduisant les exigences de la protection.

Définition 2.9 ((ε,δ)-Confidentialité différentielle [DKM$^+$06])**.** Un algorithme probabiliste $\mathcal{A}$ satisfait la ε-confidentialité différentielle si pour tous ensembles de données voisins D_1 et D_2, et pour tout résultat possible O de $\mathcal{A}$, $Pr[\mathcal{A}(D_1) = O] \leq e^{\varepsilon} \times Pr[\mathcal{A}(D_2) = O] + \delta$

.

Sensibilité lisse

La sensibilité globale [NRS07a] décrite précédemment, mesure la plus grande variabilité dans le résultat de la requête f entre les ensembles de données voisins, elle constitue une borne supérieure de cette variabilité, mais dans la plupart des cas, la variabilité de f entre un ensemble spécifique D et ses voisins est inférieure à la sensibilité globale. L'ampleur du bruit généré en utilisant la sensibilité globale est bien trop grande pour certaine requêtes. [NRS07b] propose la sensibilité locale.

Définition 2.10 (Sensibilité locale [NRS07a])**.** Soit une requête $f : D \to \mathbb{R}^d$, la sensibilité locale SL_f de f est

$$SL_f(D) = \max_{D'} |f(D) - f(D')|_1 \tag{2.5}$$

Pour tout D', tel que D' et D sont voisins.

Un mécanisme intuitif serait d'ajouter à la réponse $f(D)$ un bruit proportionnel à la sensibilité locale $SL_f(D)$. Cependant, un tel mécanisme peut provoquer des fuites d'information. Par exemple, supposons que D' et D sont voisins, si D possède une très faible sensibilité locale $SL_f(D)$, et D' une grande sensibilité locale $SL_f(D')$, la réponse retournée par le mécanisme sur l'ensemble de données D est très proche de $f(D)$. Cependant, la réponse retournée avec l'ensemble de données D' pourrait être loin de $f(D)$. Dans ce cas, un attaquant est en mesure de déduire si l'ensemble de données original est D ou D', en fonction de la distance entre $f(D)$ et la sortie. Pour surmonter ce problème, les ateurs de [NRS07a] proposent de lisser l'échelle du bruit à travers les ensembles de données voisins, et introduisent la sensibilité lisse.

Définition 2.11 (Sensibilité lisse [NRS07a])**.** Pour $\beta > 0$, la β-sensibilité lisse $SL_{f,\beta}$ de f est

$$SL_{f,\beta}(D) = \max_{D^*}(SL_f(D^*) \exp(-\beta d(D, D^*)) \tag{2.6}$$

Quel que soit l'ensemble de données D^*.
$d(D, D^*)$ désigne la distance entre les ensemble de données D et D^*

4.3.4 Composition

Pour une séquence de mécanismes respectueux de la confidentialité différentielle, la composition des mécanismes garantit la confidentialité de la manière suivante :

Définition 2.12 (Composition séquentielle [McS09])**.** Pour une séquence de n mécanismes M_1,...,M_n où chaque M_i respecte la ε_i-confidentialité différentielle, la séquence des M_i mécanismes assure la $(\sum_{i=1}^{n} \varepsilon_i)$-confidentialité différentielle.

Définition 2.13 (Composition parallèle [McS09])**.** Si $D_{1 \leqslant i \leqslant n}$ sont des ensembles disjoints de la base de données originale et que M_i sont des mécanismes qui assurent la ε_i-confidentialité différentielle pour chaque $D_{1 \leqslant i \leqslant n}$, alors la séquence de M_i assure la $(\max(\varepsilon_i))$-confidentialité différentielle.

4.4 Lien entre k-anonymat et confidentialité différentielle

Le k-anonymat et la confidentialité différentielle adoptent des approches de limitation de la divulgation qui sont fondamentalement différentes. Il existe toutefois dans la littérature, quelques travaux qui lient les deux formalismes.

Des approches permettent d'assurer la confidentialité différentielle à partir d'un ensemble de données k-anonyme. Par exemple, les auteurs dans [LQS], démontrent qu'ajouter une étape d'échantillonnage aléatoire avant le k-anonymat conduit à un mécanisme différentiellement privé. Soria Comas et al. [SCDFSM14] montrent que la quantité de bruit nécessaire pour respecter la ε-confidentialité différentielle peut être considérablement réduite si la requête est exécutée sur une version k-anonyme de l'ensemble de données, obtenue par microagrégation de tous les attributs (au lieu de l'exécuter sur les données brutes).

4.5 Synthèse

Les formalismes de protection présentés dans cette section répondent à différents types de menaces, assurant ainsi différents niveaux de protection. Cependant, cette disparité se traduit aussi par une variation de l'utilité des données selon le formalise appliqué. Le compromis entre protection et perte d'information est un paramètre clef dans le choix d'un formalise de protection. Le Tableau 2.2, extrait de [FWCY10], illustre la protection qu'offrent quelques formalismes de protection de la littérature par rapport à chaque type de menaces.

5 Critères d'évaluation quantitative de l'anonymisation

5.1 Mesurer l'utilité des données

Un détenteur de données vise à publier des données qui en plus d'être protégées, sont utiles. Afin de fournir un niveau suffisant de protection des données, les algorithmes d'ano-

Formalisme de protection	**Type de Menace**			
	Divulgation d'identité	Divulgation d'attribut	Divulgation d'appartenance	Attaque Probabiliste
k-Anonymat [Swe02b]	✓			
k-Anonymat MultiR [NCN09]	✓			
ℓ-Diversité [MKGV07]	✓	✓		
Renforcement de la confiance [WFY07]		✓		
(α-k)-Anonymat [WLFW06]	✓	✓		
(X, Y)-Anonymat [WF06]	✓	✓		
(k-e)-Anonymat [ZKSY07]		✓		
(ϵ-m)-Anonymat [LTX08]		✓		
Confidentialité Personnalisée [XT06b]		✓		
t-Proximité [LL07]		✓		✓
δ-Présence [MNC07]			✓	
(c-t)-Isolation [CDM$^+$05]	✓			✓
(ϵ)-Confidentialité Différentielle [Dwo06a]			✓	✓
(d,γ)-Confidentialité [RSH07]			✓	✓
Confidentialité Distribuée [BLR08]			✓	✓

Tableau 2.2 – Formalismes de protection de la vie privée [FWCY10]

nymisation altèrent les données de telle sorte qu'aucun individu ne puisse être identifié de manière unique. Par exemple, un ensemble de données peut être trivialement généralisé à une seule classe d'équivalence en supprimant tous les attributs du quasi-identifiant. Cette approche assure un maximum de protection, mais les données qui en résultent sont inutiles. Puisque les données assainies doivent permettre d'effectuer des tâches de recherche et d'analyse, il est primordial d'assurer un bon compromis entre protection de la vie privée et utilité.

La pertinence d'une mesure d'utilité dépend fortement de l'usage des données. Des mesures qui conviennent à la publication de microdonnées peuvent ne pas être appropriées pour évaluer l'utilité des données dans un environnement de base de données interactive. Par exemple, dans une publication de microdonnées, nous pouvons évaluer dans quelle mesure la corrélation entre les attributs ou les distributions marginales sont conservées, mais ces mesures d'utilité ne sont pas appropriées pour les bases de données interactives.

Dans [WROK09], les auteurs évaluent plusieurs mesures d'utilité adaptées à la publi-

cation de microdonnées anonymisées. [SRN+16] considère les mesures d'utilité spécifiques aux données synthétiquement générées. En général, quantifier l'utilité revient à exploiter une ou plusieures mesures quantitatives pour évaluer la perte d'information entre l'ensemble de données original T et l'ensemble de données perturbé T'. Nous présentons ci-dessous quelques-unes de ces mesures.

5.1.1 Critères d'utilité pour les données généralisées

La métrique de classification [Iye02] CM est définie comme la somme des pénalités pour chaque enregistrement, normalisée par le nombre total d'enregistrements N.

$$CM = \frac{1}{N} \sum_{i=1}^{N} pénalité(r_i) \tag{2.7}$$

Un enregistrement r est pénalisé s'il est supprimé ou si son label de classe $classe(r)$ ne représente pas le label majoritaire de sa classe d'équivalence $majorité(CE(r))$.

$$pénalité(r) = \begin{cases} 1 \text{ si } r \text{ est supprimé} \\ 1 \text{ si } classe(r) \neq majorité(CE(r)) \\ 0 \text{ sinon} \end{cases} \tag{2.8}$$

Perte d'information [Iye02] Cette métrique capture la pénalité engendrée par la généralisation d'un attribut spécifique, en quantifiant la fraction des valeurs du domaine qui ont été généralisées et en tenant compte du niveau de la généralisation. Soit l'enregistrement r qui possède une valeur d'attribut P, et M_p le nombre de nœuds feuilles dans le sous-arbre du nœud P ; soit M est le nombre total de nœuds feuilles dans l'arbre de généralisation ; La perte d'information pour un enregistrement et pour un attribut spécifique est donnée par :

$$InfoLoss(r) = \frac{M_p - 1}{M - 1} \tag{2.9}$$

La perte globale d'information d'une table anonyme T' de taille N peut être calculée comme suit :

$$\frac{1}{N}\sum_{i=1}^{N} InfoLoss(r_i) \tag{2.10}$$

Métrique de complétude des données Cette métrique est utilisée dans le cas où la suppression est appliquée, et mesure la perte de données concernant certains individus. Elle mesure la complétude de la table anonyme relativement à la table originale par calcul du taux d'enregistrements supprimés.

5.1.2 Critères d'utilité génériques

La divergence de Kullback-Leibler [KL51] La divergence de Kullback-Leibler est une mesure non-symétrique largement utilisée dans la communauté des statistiques. L'ensemble de données original T est transformé en distribution de probabilité P ; l'ensemble de données assainies T' est lui transformé en une distribution Q. La divergence KL entre deux vecteurs de probabilités $P = (p_1, ..., p_k)$ et $Q = (q_1, ..., q_k)$ est donnée par la formule suivante :

$$KL(P \parallel Q) = \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{q_i}\right) \tag{2.11}$$

Le mesure KL s'interprète comme le nombre moyen de bits nécessaire pour coder des données tirées suivant la loi P, en utilisant un code optimal pour la loi Q. Bien que cette divergence soit positive ou nulle, elle ne possède pas de borne supérieure et peut donc potentiellement prendre des valeurs infinies.

La distance de Hellinger La distance de Hellinger est une alternative à la divergence de Kullback-Leibler. L'ensemble de données original T est transformé en une distribution de probabilité P. L'ensemble de données assainies T' est lui aussi transformé en une distribution Q. la La distance de Hellinger entre deux vecteurs de probabilités $P = (p_1, ..., p_k)$ et $Q = (q_1, ..., q_k)$ est donnée par la formule suivante :

$$D_{Hellinger}(P, Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}(\sqrt{p_i} - \sqrt{q_i})^2}. \tag{2.12}$$

Les motivations pour l'utilisation de la distance de Hellinger [Ant16] pour la mesure de la perte d'information sont les suivantes :

— il s'agit d'une métrique (au sens mathématique) ;

— elle possède un impact plus important lorsque les probabilités sont faibles ;

— elle peut être calculée lorsque les probabilités sont nulles.

Requêtes aléatoires de dénombrement Les requêtes aléatoires (ou requêtes linéaires) sont souvent utilisées comme mesure d'utilité des données. L'idée est de générer des requêtes de dénombrement avec des prédicats de requêtes aléatoires sur l'ensemble des attributs :

Select COUNT(*) From D Where $X_1 \in I_1$ and $X_2 \in I_2$ and ... and $X_d \in I_d$.

Pour chaque attribut X_i, I_i est un intervalle aléatoire généré à partir du domaine de X_i. L'erreur relative est utilisée pour mesurer l'exactitude d'une requête q, où $A_{original}(q)$ indique la vraie réponse de q sur les données originales T et $A_{anonyme}(q)$ est le compte calculé lorsque les données anonymisées T' sont utilisées.

$$ErreurRelative(q) = \frac{|A_{anonyme}(q) - A_{original}(q)|}{A_{original}(q)} \tag{2.13}$$

Performance en classification Afin de mesurer l'utilité des données assainies, de nombreux travaux [WFY05, Iye02, LDR06b] considèrent des modèles de classification. Une approche générique est synthétisée ci-dessous :

1. Diviser l'ensemble de données original T en deux ensembles : $T_{apprentissage}$ et T_{test},
2. A partir de $T_{apprentissage}$, exploiter le mécanisme d'anonymisation dont les données sont à évaluer pour produire un ensemble de données $T'_{apprentissage}$,
3. Sélectionner une variable cible, puis construire deux modèles de classification, avec un paramétrage identique, le premier $M_{original}$ en utilisant $T_{apprentissage}$ pour l'apprentissage, le deuxième $M_{anonyme}$ en utilisant $T'_{apprentissage}$,

4. Évaluer la qualité des modèles construits $M_{original}$ et $M_{anonyme}$ sur l'ensemble de test T_{test} en calculant différentes mesures de performance en classification telles que la *précision* ou l'*aire sous la courbe de ROC*.

5.2 Évaluation des risques

L'évaluation des risques concerne principalement les approches proposées dans le *Contrôle de la Divulgation Statistique*, car celles-ci se focalisent sur l'utilité des données. Le niveau de protection est déterminé *a posteriori*, et diffère d'un jeu de données à l'autre. Lorsque ces méthodes sont appliquées, les détenteurs de données doivent évaluer quantitativement le risque de divulgation, afin de vérifier s'il est inférieur à un seuil défini, auquel cas il est considéré comme acceptable. Dans cette optique, diverses mesures d'estimation du risque de divulgation ont été proposées [DL89, TFBJ04]. Leur validité dépend fortement des scénarios d'application envisagés, mais il existe néanmoins un consensus sur le fait que le risque de divulgation ne peut être réduit à zéro (à moins de supprimer toute l'information). Ainsi, en général, un seuil devrait être déterminé pour décider si un ensemble de données doit être publié ou non. Il existe principalement deux approches pour évaluer le risque de divulgation [KB08] : l'estimation de la *rareté* dans l'échantillon ou dans la population, et l'estimation de la *probabilité de ré-identifier* un enregistrement masqué à l'aide de connaissances externes.

Dans un scénario typique d'attaque de ré-identification, un attaquant a des connaissances sur les attributs quasi-identifiants. En prenant l'exemple d'une base de données médicale, l'attaquant peut connaître quelques attributs (âge, sexe, état matrimonial) à partir d'un registre public externe (données de recensement) ou d'une source privée d'information (par exemple, connaître l'âge et l'adresse de son voisin). L'attaquant essaie ensuite de faire correspondre ces quasi-identifiants avec les enregistrements modifiés dans la base de données publiée. La première approche s'intéresse à la fréquence d'apparition des combinaisons de quasi-identifiants. La deuxième approche consiste à estimer la probabilité qu'un enregistrement dans la base de données soit ré-identifié.

Mesure de la rareté [SE02] Dans les approches de la *SDC*, les données assainies sont souvent obtenues par échantillonnage. De ce fait, nombre de métriques d'évaluation de risque sont dépendantes de l'échantillon. Soit l'échantillon s de taille n sélectionné à partir d'une population finie U (s $\subset$ U) de N individus. La variable catégorielle formée par le croisement de touts les quasi-identifiants est désignée par X, avec des valeurs notées $1, \ldots, J$. Chacune de ces valeurs correspond à une combinaison possible de valeurs des quasi-identifiants. Par exemple, si les quasi-identifiants sont $\langle$âge, sexe, état matrimonial$\rangle$ une valeur possible de X pourrait être $\langle$49 ans, femme, divorcé$\rangle$.

Soit X_i la valeur de X pour l'individu i de la population. Soient les *fréquences dans la population* pour les différentes valeurs de X notées :

$F_j = \sum_{i \in U} I(X_i = j) \qquad , \qquad j = 1, \ldots, J$

où $I(.)$ est la fonction indicatrice : $I(A) = 1$ si A est vrai et $I(A) = 0$ sinon. Toutes les catégories dont le compte est nul sont exclues, de sorte que $F_j \geq 1$ pour $j = 1, \ldots, J$. Soient les *fréquences de fréquences dans la population* notées :

$N_r = \sum_{j=1}^{J} I(F_j = r) \qquad , \qquad r = 1, 2, \ldots$

Parallèlement, les quantités pour l'échantillon f_j et n_r sont définies de manière analogue à F_j et N_r, respectivement. Ainsi, les *fréquences dans l'échantillon* sont définies par :

$f_j = \sum_{i \in s} I(X_i = j) \qquad , \qquad j = 1, \ldots, J$

et les *fréquences de fréquences dans l'échantillon* notées :

$n_r = \sum_{j=1}^{J} I(F_j = r) \qquad , \qquad r = 1, 2, \ldots$

$$Pr(PU) = N_1/N = \sum_{j=1}^{J} I(f_j = 1)/N \tag{2.14}$$

$Pr(PU)$ représente la proportion d'individus dans la population qui sont uniques.

$Pr(SU)$ représente la proportion d'individus dans l'échantillon qui sont uniques.

Une métrique d'évaluation basée sur la rareté, proposée dans [SE02] est :

$$Pr(PU|SU) = \sum_{j} I(f_j = 1, F_j = 1) / \sum_{j} I(f_j = 1) \tag{2.15}$$

Il s'agit de la probabilité conditionnelle qui détermine si, pour un individu tiré au hasard de la population, l'individu est unique dans la population, sachant qu'il est unique dans l'échantillon.

Probabilité de ré-identification [FS69] Un processus de couplage d'enregistrements tente de classer des paires dans un espace produit $A \times B$ à partir de deux fichiers A et B dans M, l'ensemble des liens vrais, et U, l'ensemble des non-liens vrais. Fellegi et Sunter [FS69] ont considéré les ratios R de probabilités de la forme :

$$R = \frac{Pr(\gamma \in \Gamma|M)}{Pr(\gamma \in \Gamma|U)} \tag{2.16}$$

où γ est un modèle d'accord arbitraire dans un espace de comparaison Γ. Par exemple, Γ pourrait consister en plusieurs modèles représentant un accord simple ou non sur le nom, le prénom et l'âge. Par ailleurs, chaque $\gamma \in \Gamma$ pourrait également tenir compte de la fréquence relative avec laquelle des noms de famille spécifiques, tels que Scheuren ou Winkler, apparaissent ou sont traités avec différents types de comparaisons de données quantitatives. Les champs comparés (nom, prénom, âge) sont appelés variables d'appariement.

La règle de décision, où T_μ et T_λ dénotent les seuils de décision, est donnée par :

1. Si $R > T_\mu$, désigner la paire comme lien.
2. Si $T_\lambda \leq R \leq T_\mu$, désigner une paire comme lien possible.
3. Si $R < T_\lambda$, désigner la paire comme une paire sans lien.

6 Outils d'anonymisation

Il existe plusieurs outils « open source » pour l'assainissement des données, nous en présentons trois parmi les plus connus.

6.1 μ-argus

Le logiciel μ-Argus [HvdWR+03] (Anti-Re-identification General Utility System) est la plus ancienne des trois initiatives. Il a été développé par l'Agence nationale des statistiques des Pays-Bas dans le cadre du projet européen CASC (Computational Aspects of Statistical Confidentiality). Issu de la SDC, μ-Argus ne vise pas à atteindre un formalisme de protection en particulier. Il implémente plutôt des techniques d'assainissement perturbatrices (microagrégation, permutation) et non perturbatrices (recodage global, suppression locale) que l'utilisateur a la possibilité d'appliquer aux données. L'utilisation de μ-Argus exige que les éléments suivants soient spécifiés :

— les données et leur structure,

— les variables clés,

— un tableau des risques basé sur les combinaisons observées des variables clés, qui contient les risques individuels. Comme le risque individuel est le même pour tous les individus ayant la même combinaison de variables clés, le tableau des risques peut être construit sur des combinaisons,

— un seuil qui permet de classer les enregistrements selon qu'ils soient sûrs ou non sûrs, respectivement.

Après avoir identifié les enregistrements à risques, l'utilisateur choisit une technique, en spécifiant les valeurs de ses paramètres en entrée, et demande au système à ce qu'elle soit exécutée sur les données.

6.2 CAT

L'outil CAT (Cornell Anonymization Toolkit) est conçu pour anonymiser de manière interactive les données publiées afin de limiter les risques de divulgation en considérant plusieurs formalismes de protection. Il a été développé par le département d'informatique de l'Université Cornell [XWG09]. La boîte à outils contient les fonctions suivantes :

— Anonymisation des données à l'aide de la généralisation.

— Évaluation des risques de divulgation de chaque enregistrement dans des données rendues anonymes en se fondant sur des hypothèses spécifiées par l'utilisateur au

sujet des connaissances de base de l'adversaire. De plus, la répartition des risques de divulgation de tous les enregistrements de l'ensemble de données peut être illustrée dans un histogramme.

— Évaluation de l'utilité, en comparant les tableaux de contingence et les graphiques de densité entre les données d'origine et les données anonymisées. Les deux mesures fournissent aux utilisateurs un moyen intuitif d'apprendre la distorsion statistique induite par l'anonymisation de l'ensemble de données.

— Manipulation des enregistrements sensibles, en appliquant un traitement spécial aux enregistrements qui présentent des risques de divulgation beaucoup plus élevés que la plupart des autres enregistrements. De tels enregistrements pourraient être les valeurs aberrantes de l'ensemble de données, et peuvent potentiellement, dégrader gravement la qualité de l'anonymisation. Les utilisateurs peuvent alors éliminer ces enregistrements sensibles qui présentent des risques élevés de divulgation.

— Visualisation et interaction. L'ensemble des résultats des fonctions ci-dessus peut être affiché dans l'interface utilisateur, et les utilisateurs peuvent appliquer le processus itératif ci-dessus et observer les résultats jusqu'à ce qu'ils obtiennent une anonymisation satisfaisante en termes de confidentialité des données et d'utilité.

L'ensemble de données à anonymiser est conservé en mémoire principale, et toutes les fonctions ci-dessus sont exécutées sur ces données chargées en mémoire principale, elles ne sont sorties sur disque que lorsque les utilisateurs sont satisfaits du résultat de l'anonymisation.

6.3 ARX

ARX[1] est un logiciel « open source » pour l'anonymisation des données. Comme CAT, il implémente plusieurs formalismes de protection. Il est divisé en trois volets, qui modélisent différents aspects du processus d'anonymisation. Le processus est décrit Figure 2.7.

— Dans le volet « configuration », les données peuvent être importées, des règles de

1. `https://arx.deidentifier.org/anonymization-tool/`

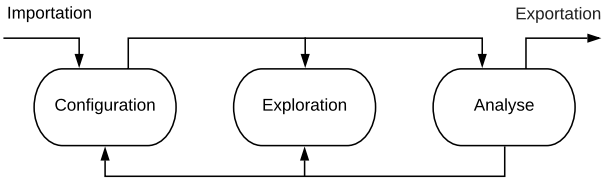


Figure 2.7 – Processus d'anonymisation d'ARX

transformation peuvent être créées, et la confidentialité ainsi que des modèles de qualité des données peuvent être sélectionnés et paramétrés.

— Au cours du processus d'anonymisation, ARX caractérise un espace de transformations potentielles de l'ensemble de données d'entrée. Pour chaque solution candidate, on détermine si les seuils de risque sont atteints et la qualité des données est quantifiée selon le critère donné. Cette perspective permet aux utilisateurs de parcourir le résultat de ce processus et de sélectionner des transformations intéressantes pour une analyse plus approfondie.

— Le volet « analyse » comprend deux éléments :

- l'évaluation de l'utilité : les données d'entrée et les données transformées sont affichées côte à côte. De plus, des statistiques descriptives peuvent être calculées
- l'évaluation du risque : diverses mesures reflétant les risques d'entrave à la vie privée sont présentées. Les métriques mises en œuvre dans ARX comprennent les risques de réidentification ainsi que des mesures de rareté. De plus, ce volet donne accès à une méthode de détection des attributs qui doivent être modifiés selon la méthode Safe Harbor de la loi américaine Health Insurance Portability and Accountability Act (identifiants HIPAA).

7 Conclusion

Dans un contexte où la protection des données personnelles suscite une attention croissante, les chercheurs tentent de répondre aux besoins actuels en matière d'anonymisation. Pour faire face au défi de publier des données utiles tout en protégeant la vie privée des individus qui les constituent, un grand nombre de techniques d'assainissement, de forma-

lismes de protection et d'outils d'anonymisation ont été proposés par les communautés de la *Publication de Données Respectueuse de la Vie Privée*, et de *Contrôle de la Divulgation Statistique*.

Ce chapitre nous a permis d'avoir une vue d'ensemble sur l'anonymisation des micro-données. Nous avons constaté l'existence de plusieurs formalismes de protection, chaque formalisme est motivé par la nécessité de contrer un ou plusieurs type d'attaques. Notons que pour atteindre chaque formalisme de protection, il existe une multitude d'algorithmes dans la littératures. Par souci de clarté, nous avons fait le choix de ne pas les présenter dans ce chapitre. En revanche, pour chaque formalisme de protection traité dans les chapitres qui suivent, nous en décrivons quelques-uns dans un état de l'art spécifique au chapitre.

la Figure 2.8 représente une typologie des solutions de publication de données respectueuse de la vie privée, qui ont été abordées dans ce chapitre et positionne les contributions de la thèse que l'on décrit dans les chapitres qui suivent. Nous nous intéressons à la publication de données respectueuse de la vie privée et nous nous focalisons sur les micro-données. Nous avons identifié deux paradigmes de protection proposés dans la littérature : le *Contrôle de la Divulgation Statistique* (SDC) et la *Publication de Données Respectueuse de la Vie Privée* (PPDP). La première contribution lie les deux paradigmes ; elle se base sur la génération de données synthétiques, technique qui provient de le SDC avec l'anonymisation par groupe, une famille de techniques d'anonymisation qui provient de la PPDP. La deuxième contribution est basée sur un approche perturbatrice qui vise à atteindre le formalisme de protection du k-anonymat, qui a été proposé dans la PPDP. Enfin, la troisième contribution s'inscrit dans la famille des techniques basées sur l'ajout de bruit contrôlé, elle vise à assurer la ε-confidentialité différentielle qui provient également de la PPDP.

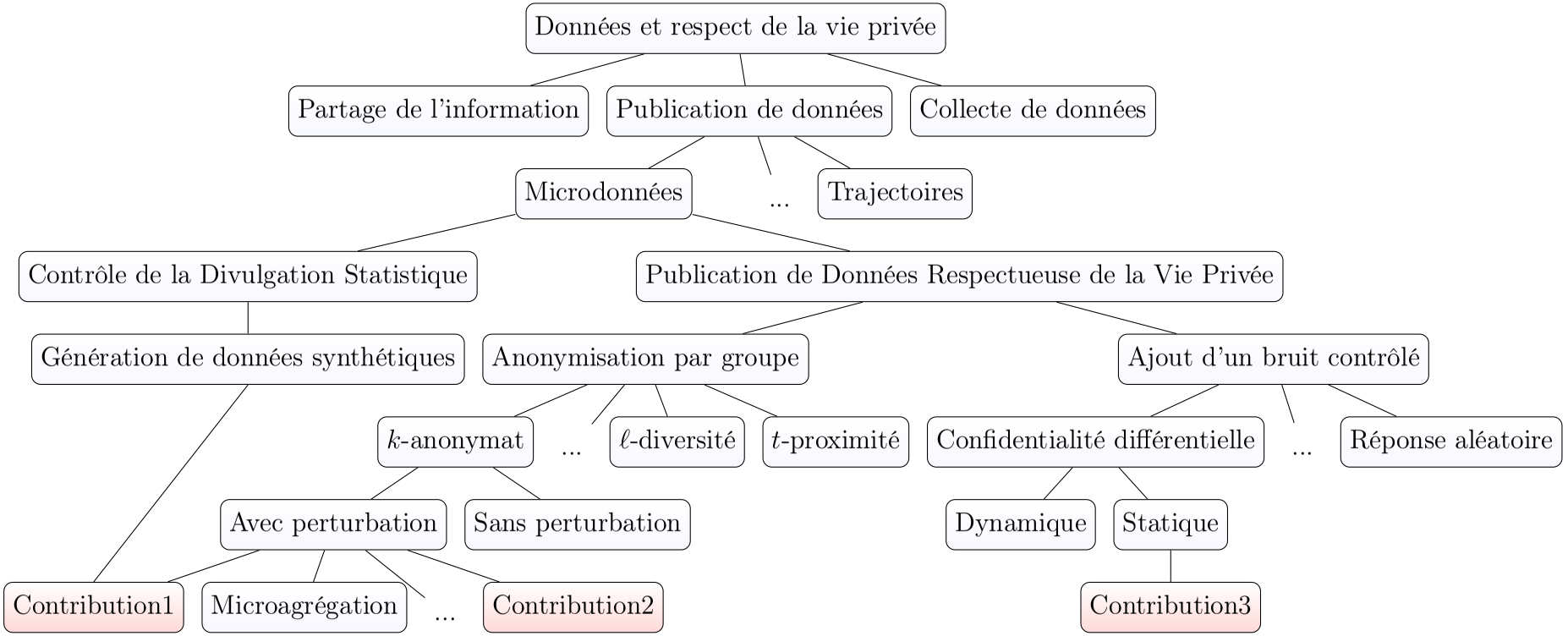


Figure 2.8 – Positionnement des travaux de la thèse

Chapitre 3

Génération de données synthétiques à l'aide du co-clustering pour la protection de la vie privée

Sommaire

1 Introduction

On s'intéresse à une solution d'anonymisation pour pallier le risque de ré-identification. Le processus d'anonymisation auquel nous nous intéressons doit être agnostique à l'usage que l'on fait des données. Ces dernières doivent être dans le même format que les données de départ tout en conservant suffisamment d'information pour permettre à l'utilisateur

d'effectuer des analyses de natures diverses.

Dans ce chapitre, on présente une solution qui vise à se prémunir du risque de ré-identification sur des microdonnées destinées à la publication. Pour construire la solution d'anonymisation, on s'appuie sur deux types de techniques : le co-clustering qui permet de structurer et de résumer l'information contenue dans une table de données et la génération de données synthétiques qui consiste à remplacer les individus originaux de la table par des individus fictifs. On décrit dans ce chapitre les outils et techniques sur lesquels on s'appuie pour construire la solution d'anonymisation. Enfin, La solution est évaluée en termes d'utilité et de protection sur deux jeux de données.

2 Le Co-clustering MODL

Dans cette section, nous présentons la technique clé du co-clustering sur laquelle repose notre méthodologie d'anonymisation.

2.1 Le co-clustering, généralités

La classification croisée (ou co-clustering) est une technique non supervisée qui a pour objectif de réaliser une classification (simultanée) des lignes et des colonnes d'un tableau de données [Har75]. La classification croisée est une extension de la classification simple (clustering) qui permet d'extraire la structure sous-jacente dans les données sous forme de groupes de lignes et groupes de colonnes. L'avantage de cette technique, par rapport à la classification simple, réside dans l'étude simultanée (jointe) des lignes et des colonnes qui permet d'extraire un maximum d'informations sur leurs dépendances. Cette technique est adaptée, par exemple, dans des contextes comme l'analyse des paniers de consommation où l'objectif est d'identifier les sous-ensembles de clients ayant tendance à acheter les mêmes de produits, plutôt que de grouper simplement les clients (ou les produits) en fonction des modèles d'achat/vente.

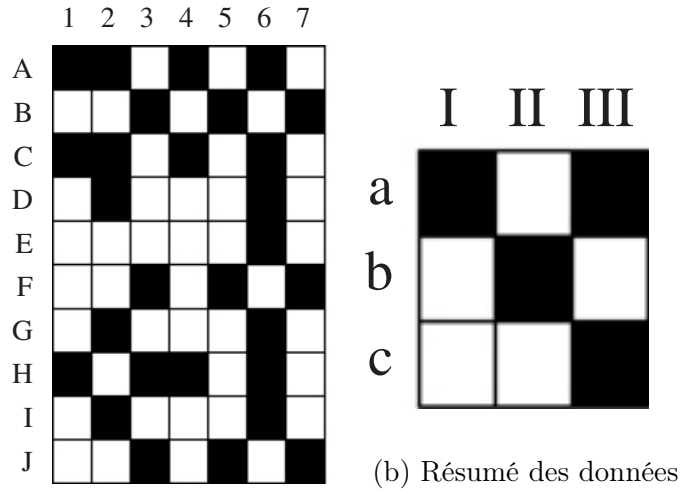
Les données étudiées dans les problèmes de co-clustering sont de même nature que les données traitées par les approches de clustering : elles sont composées d'un ensemble

d'observations m sans étiquette, décrites par plusieurs variables explicatives, dénotées $\{X_1, X_2, ..., X_d\}$. Ces variables peuvent être continues ou catégorielles, prenant alors un nombre fini de valeurs différentes. Les valeurs prises par les variables descriptives sont partitionnées afin d'obtenir de nouvelles variables $\{X_1^M, X_2^M, ..., X_d^M\}$ appelées variables-partitions. Les valeurs de ces nouvelles variables sont les clusters obtenus par les partitions des valeurs des variables $\{X_1, X_2, ..., X_d\}$. Chacune des variables X_i^M a $\{k_1, k_2, ..., k_d\}$ valeurs qui sont des groupes de valeurs si la variable est nominale et des intervalles si la variable est continue.

Dans la littérature, différentes approches de co-clustering ont été développées. Ces méthodes diffèrent principalement par le type des données étudiées (continues, binaires, catégorielles), les hypothèses considérées, la forme souhaitée pour les résultats (classification stricte ou floue, hiérarchie, etc.). Parmi les familles de méthodes, on peut citer les méthodes de reconstruction de matrices où le problème est présenté sous forme d'approximation matricielle [CC00] et les méthodes de modèles à blocs latents basées sur les modèles de mélange où les blocs (les classes des lignes et les classes des colonnes) sont définis par des variables latentes à estimer. Les approches à blocs latents traitent l'ensemble des lignes et l'ensemble des colonnes d'un tableau de données (les individus et des variables explicatives) simultanément en cherchant à organiser la matrice de données en blocs homogènes [GN13].

L'utilité du co-clustering réside dans sa capacité à créer des groupes facilement interprétables et dans sa capacité de réduction d'une grande table de données en une matrice significativement plus petite et ayant la même structure que les données originales.

La Figure 3.1 montre l'exemple d'un tableau de données binaires représentant $n = 10$ individus et $d = 7$ variables binaires (3.1a) [GN08] et le tableau binaire résumant le résultat d'une classification croisée en $3 \times 3 = 9$ blocs binaires (3.1b), résumé qui permet de visualiser plus simplement les principales associations.



(a) Ensemble de données binaires

(b) Résumé des données

Figure 3.1 – Tableau de données binaires et matrice de co-clustering

2.2 L'algorithme MODL

L'approche de co-clustering MODL (Minimum Optimized Description Length) proposée par Boullé [Bou07] est basée sur la famille des modèles en grille de données qui cherchent à partitionner chaque variable en intervalles dans le cas numérique et à grouper les valeurs dans le cas catégoriel. Le produit cartésien des partitions univariées forme une partition multivariée de l'espace de représentation dans un ensemble de cellules (la grille de données).

Les paramètres qui définissent complètement le modèle de co-clustering sont : le nombre de groupes (ou intervalles) pour chaque variable à expliquer, la partition de chaque variable à expliquer en groupes de valeurs (ou intervalles), la distribution des observations sur les cellules de la grille de données ainsi définie, la distribution des observations de chaque groupe sur les valeurs du groupe, pour chaque variable à expliquer.

Le meilleur modèle de co-clustering $\mathcal{M}$ est défini comme étant le plus probable connaissant les données et est obtenu par une approche bayésienne dite Maximum A Posteriori (MAP). Le critère à optimiser est constitué d'un terme de prior sur le modèle $P(\mathcal{M})$ qui traduit la longueur de codage du modèle et d'un terme de vraisemblance connaissant les paramètres du modèle $P(\mathcal{M}|D)$ qui traduit la longueur de codage des données

connaissant les données.

Le critère établit un compromis entre la précision et la robustesse du modèle en grille. Les modèles les plus proches des données seront préférés. Les modèles complexes (beaucoup de clusters pour les variables catégorielles et/ou beaucoup d'intervalles pour la variable numérique) seront pénalisés. Le modèle s'interprète selon un principe MDL (Minimum Description Length).

L'optimisation est réalisée à l'aide d'une heuristique gloutonne ascendante démarrant de la grille au grain le plus fin et réalisant à chaque étape la fusion de clusters qui fait décroitre le plus le critère. Une post-optimisation améliore cette heuristique en effectuant des permutations au sein des clusters.

La construction du critère, ainsi que l'algorithme d'optimisation et les propriétés asymptotiques de l'approche sont détaillés dans [Bou11] pour le cas d'un co-clustering à deux dimensions catégorielles et dans [Bou12] pour le cas de données mixtes (numériques et catégorielles).

2.3 Le co-clustering comme estimateur de densité jointe et approximateur universel

Boullé [Bou12] a démontré que le critère MODL dans le cas d'un co-clustering de deux variables catégorielles converge asymptotiquement vers la vraie loi de probabilité jointe des variables et se comporte comme un approximateur universel.

2.4 Simplification du co-clustering

La partition obtenue par un co-clustering MODL peut être très fine et contenir un très grand nombre de clusters. Le nombre de clusters dans le cas d'une variable catégorielle est compris entre 1 et n, le nombre de valeurs prises par la variable. Ainsi, plus le nombre de modalités de la variable est important, plus le nombre potentiel de clusters est élevé. Asymptotiquement, on peut d'ailleurs obtenir autant de clusters que de modalités sans qu'il s'agisse de sur-apprentissage. C'est pour cette raison qu'une simplification de la grille a été proposée [BGR12] de manière à la rendre plus lisible. La simplification se fait

hiérarchiquement par fusions successives des clusters, en maîtrisant la dégradation du modèle. L'outil construit une hiérarchie des parties de chaque dimension (i.e., clusters ou intervalles adjacents) en utilisant une stratégie agglomérative ascendante, en partant de la grille optimale résultant de la procédure d'optimisation, jusqu'au modèle nul, i.e., la grille (unicellulaire) où aucune dimension n'est partitionnée. Les hiérarchies sont construites en fusionnant les parties de manière à minimiser la perte d'information de la grille avant fusion. Les fusions sont faites séquentiellement plutôt qu'indépendamment sur chaque partition. À chaque étape, la meilleure fusion parmi toutes les fusions possibles sur les deux partitions est effectuée. Ainsi, le co-clustering reste informatif à chaque niveau de la hiérarchie.

2.5 Les logiciels Khiops et Khiops CoViz

Pour mener à bien nos expérimentations, nous avons utilisé le logiciel Khiops qui implémente l'approche MODL et Khiops CoViz qui est la brique logicielle de visualisation de Khiops [1]. La visualisation d'un résultat de co-clustering de variables prend la forme d'une grille (ou matrice) dont les dimensions sont partitionnées : les variables catégorielles sont partitionnées en clusters et les variables numériques en intervalles. Différentes informations de visualisation sont disponibles comme la fréquence des cellules, ainsi que différents critères qui fournissent des informations supplémentaires sur les interactions entre variables comme la contribution à l'Information Mutuelle des cellules de la grille (co-clusters). L'outil permet plusieurs variantes de visualisations à différentes échelles de la grille. La grille optimale et les hiérarchies correspondantes constituent les principales structures de l'outil de visualisation. L'utilisateur peut ainsi choisir la granularité de la grille nécessaire à son analyse tout en contrôlant soit le nombre de parties soit le taux d'information (i.e., le pourcentage d'information gardé dans le modèle.)

1. `www.khiops.com`

3 Mise en œuvre du co-clustering sur des données multidimensionnelles

La mise en œuvre du co-clustering sur des données multidimensionnelles est un processus en plusieurs étapes. On présente ci-dessous le déroulé des différentes étapes.

3.1 Préparation des données

L'approche MODL est adaptée à l'étude de données complexes où les observations sont décrites par d variables. On peut donc traiter des données en exploitant toutes les variables. Le problème est que le nombre d'observations nécessaires pour peupler une grille de co-clustering croît exponentiellement avec le nombre de dimensions. Dans certains problèmes, les données sont décrites par de nombreuses variables mais le faible nombre d'observations ne permet pas à MODL d'inférer une structure [Gui13].

Pour traiter des données multidimensionnelles à l'aide du co-clustering on utilise une méthodologie en deux étapes i) lors de la première étape, toutes les variables numériques sont discrétisées en intervalles selon un nombre d'intervalles choisi par l'analyste, par discrétisation en fréquences égales par exemple, les variables catégorielles sont conservées telles quelles. ii) la deuxième étape consiste à utiliser une méthode de co-clustering entre individus et variables discrétisées, conduisant à des regroupements d'individus d'une part, et de parties de variables d'autre part. On se ramène ainsi au cas de deux variables catégorielles. Cette méthodologie a été expérimentée et évaluée dans [BBR17].

La base de données Iris Pour des raisons didactiques, on déroule la méthodologie sur la base Iris [2]. Iris contient 150 individus, décrits par 5 attributs, un catégoriel *Class* avec 3 valeurs ⟨Setosa, Versicolor, Versicolor⟩, et quatre continus, *Sepal length, Sepal width, Petal length, Petal width.*

On commence par discrétiser les variables numériques en $p = 3$ partiles. Le résultat de la discrétisation des variables pour la base Iris est illustré dans le Tableau 3.1.

2. `https://archive.ics.uci.edu/ml/datasets/iris`

id	Sepal length	Sepal width	Petal length	Petal width	Class
1	5,1	3,5	1,4	0,2	Setosa
2	6	3	4,8	1,8	Virginica
3	7,2	3,2	6	1,8	Virginica
...	...	...	...	...	...
150	6	2,2	4	1	Versicolor

(a) Table originale

id	Sepal length	Sepal width	Petal length	Petal width	Class
1	$]4.299, 5.4]$	$]3.2, 4.4]$	$]0.999, 2.633]$	$]0.099, 0.867]$	Setosa
2	$]5.4, 6.3]$	$]2.9, 3.2]$	$]2.633, 4.9]$	$]1.6, 2.5]$	Virginica
3	$]6.3, 7.9]$	$]2.9, 3.2]$	$]4.9, 6.9]$	$]1.6, 2.5]$	Virginica
...	...	...	...	...	...
150	$]5.4, 6.3]$	$]1.999, 2.9]$	$]2.633, 4.9]$	$]0.867, 1.6]$	Versicolor

(b) Table discrétisée

Tableau 3.1 – Table originale a), table discrétisée correspondante b)

id	Partie de variable
1	$SepalLength]4.299, 5.4]$
1	$SepalWidth]3.2, 4.4]$
1	$PetalLength]0.999, 2.633]$
1	$PetalWidth]0.099, 0.867]$
1	$Class\{setosa\}$
2	$SepalLength]5.4, 6.3]$
...	...

Tableau 3.2 – Nouvelle représentation en deux variables catégorielles

3.2 Transformation des données en deux variables

On transforme maintenant le jeu de donnée discrétisé en deux variables, *id* et *Partie de variable*, en créant pour chaque individu initial un enregistrement par variable, mémorisant le lien entre l'individu et sa partie de variable. L'ensemble des n individus initiaux représentés par d variables est ainsi transformé en un nouveau jeu de données de taille $N = n \times d$ ayant deux variables catégorielles, la première comportant $V = N$ valeurs et la seconde autant de valeurs que de modalités différentes.

Dans la base Iris par exemple, cette étape résulte en deux colonnes de $N = 750$ lignes (les observations). Le Tableau 3.2 montre les six premières instances : la variable *id* comportant $n = 150$ valeurs (les individus) et la variable *Partie de variable* comportant $4 \times 3 + 3 = 15$ valeurs différentes (les discrétisations des 4 variables continues en 3 intervalles ainsi que 3 valeurs pour la variable *Class*).

3.3 Co-clustering

Les données étant maintenant représentées sous forme de deux variables catégorielles, la méthode MODL est appliquée pour rechercher un modèle d'estimation de densité jointe entre ces deux variables. Dans le co-clustering résultant, les individus de la base initiale (valeurs de la variable *id*) sont regroupés s'ils sont distribués de façon similaire sur les groupes de parties de variables (valeurs de la variable *Partie de variable*), et réciproquement.

Pour la base Iris, le co-clustering le plus fin produit une grille à 3 clusters d'individus et 7 clusters de parties de variables. Le résultat du co-clustering est présenté Figure 3.3. Le Tableau 3.3 détaille la composition des clusters.

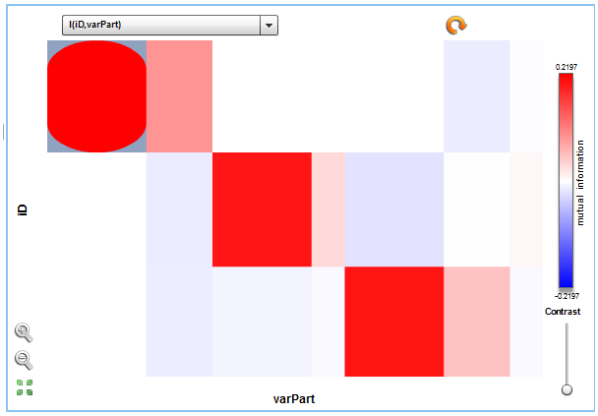


Figure 3.2 – Capture Khiops CoViz

On dispose d'une interface de visualisation (Figure 3.2) qui permet d'analyser les relations entre les individus et les parties de variables. Une information disponible est l'information mutuelle. Plus la contribution à l'information mutuelle est élevée plus la cellule est rouge. Plus elle est négative, plus elle est bleue. Lorsque les cellules sont blanches, c'est que la contribution à l'information mutuelle est proche de zéro, indiquant respectivement un excès d'interactions, un déficit ou aucune interaction particulière comparée à ce qui est attendu en cas d'indépendance des variables partitionnées.

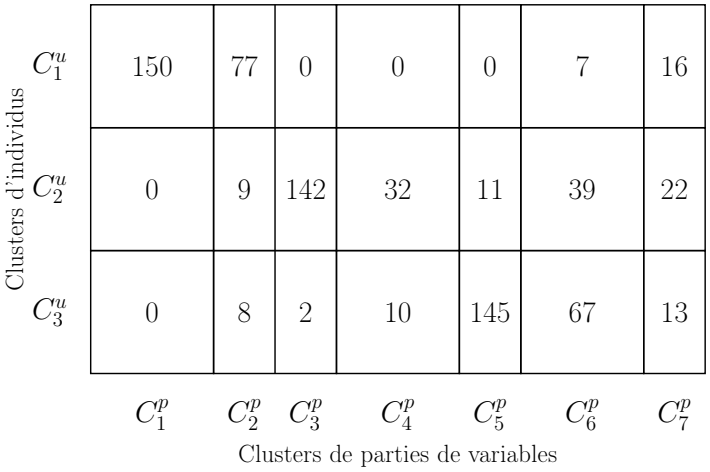


Figure 3.3 – Grille de co-clustering optimisée

Cluster	Composition	δ_j^p
C_1^p	$Class\{setosa\}$	50
	$PetalLength]0.999, 2.633]$	50
	$PetalWidth]0.099, 0.867]$	50
C_2^p	$SepalLength]4.299, 5.4]$	52
	$SepalWidth]3.2, 4.4]$	42
C_3^p	$Class\{virginica\}$	50
	$PetalWidth]1.6, 2.5]$	48
	$PetalLength]4.9, 6.9]$	46
C_4^p	$SepalLength]6.3, 7.9]$	42
C_5^p	$Class\{versicolor\}$	50
	$PetalWidth]0.867, 1.6]$	52
	$PetalLength]2.633, 4.9]$	54
C_6^p	$SepalWidth]1.999, 2.9]$	57
	$SepalLength]5.4, 6.3]$	56
C_7^p	$SepalWidth]2.9, 3.2]$	51

(a) Composition des clusters de parties de variables

Cluster	$\lvert C_i^u \rvert$
C_1^u	50
C_2^u	51
C_3^u	49

(b) Cardinalité des clusters d'individus

Tableau 3.3 – Composition des clusters

3.4 Simplification du co-clustering

Simplifier le co-clustering peut aider à la discussion en termes de structuration de l'information. La Figure 3.4 et le Tableau 3.4 présentent ainsi une version simplifiée du co-clustering en 2×3 cellules conservant 64% de l'information.

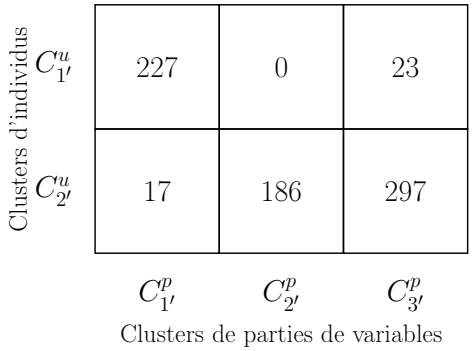


Figure 3.4 – Grille de co-clustering simplifié

Cluster	Composition	δ^p_j
$C^p_{1'}$	$PetalWidth]0.099, 0.867]$	50
	$PetalLength]0.999, 2.633]$	50
	$Class\{setosa\}$	50
	$SepalLength]4.299, 5.4]$	52
	$SepalWidth]3.2, 4.4]$	42
$C^p_{2'}$	$PetalLength]4.9, 6.9]$	46
	$PetalWidth]1.6, 2.5]$	48
	Class{virginica}	50
	$SepalLength]6.3, 7.9]$	42
$C^p_{3'}$	$SepalWidth]1.999, 2.9]$	57
	$PetalWidth]0.867, 1.6]$	52
	$PetalLength]2.633, 4.9]$	54
	$Class\{versicolor\}$	50
	$SepalLength]5.4, 6.3]$	56
	$SepalWidth]2.9, 3.2]$	51

(a) Composition des clusters de parties de variables

Cluster	$\lvert C^u_i \rvert$
$C^u_{1'}$	50
$C^u_{2'}$	100

(b) Cardinalité des clusters d'individus

Tableau 3.4 – Composition des clusters : co-clustering simplifié

4 Anonymisation et génération de données synthétiques

Dans cette section, nous détaillons comment la technique du co-clustering, décrite dans la Section 2, nous permet de construire un modèle inspiré du k-anonymat à partir d'un tableau multidimensionnel de n individus et d attributs qui peuvent être à la fois numériques et catégoriels, et comment ce modèle est exploité pour générer un tableau d'individus synthétiques.

La méthodologie d'anonymisation que l'on propose s'appuie sur les différentes pro-

priétés de la grille de co-clustering décrites précédemment :

— la grille simplifiée,

— les propriétés d'estimation de distribution jointe.

L'ensemble du processus peut être organisé en quatre étapes (Figure 3.5) : la préparation des données, le co-clustering, la simplification du co-clustering et la génération de données synthétiques. Notons que notre solution n'est pas adaptée aux situations où le nombre d'instances disponibles est trop faible pour produire un modèle de co-clustering.

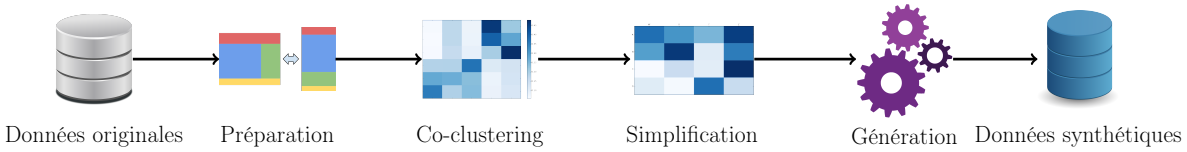


Figure 3.5 – Description des différentes étapes de la solution d'anonymisation

Les étapes de préparation des données, et de co-clustering sont effectuées comme décrites dans la Section 3.

4.1 Étape de peuplement des clusters d'individus

A ce stade, nous partons de la grille de co-clustering la plus fine. La grille peut être agrégée dans un objectif différent de celui d'une analyse exploratoire facilitée. L'objectif qui nous intéresse ici est le peuplement des clusters. On agrège la grille jusqu'à ce qu'il y ait suffisamment d'individus par cluster (au moins k individus).

À cette étape du processus, on a une représentation de clusters d'individus k-anonymes (en ce sens qu'un individu est fondu dans un groupe de k et n'est plus décrit que par une densité de probabilité sur les clusters de parties de variables).

4.2 Étape de génération de données synthétiques

Le but de cette dernière étape est de générer des enregistrements basés sur le modèle de co-clustering, qui peut être considéré comme un estimateur de la densité jointe entre les individus et les parties de variables [Bou12].

Les paramètres que nous obtenons à partir du co-clustering sont : les cardinalités des clusters *id* et *Partie de variable* (nombre d'observations), la composition des clusters

de *Partie de variable*, le compte de chaque co-cluster (le nombre d'observations dans la cellule de chaque cluster *id* $\times$ cluster de *Partie de variable*).

Nous dénotons ces paramètres par :

— $N_{g_u}^{(u)}$: nombre d'observations du cluster d'id g_u,

— $N_{g_p}^{(p)}$: nombre d'observations du cluster de partie de variable g_p,

— n_{v_k} : nombre d'observations de la partie de variable v_k de la variable X_k,

— $N_{g_u g_p}$: nombre d'observations du co-cluster (g_u, g_p) id $\times$ partie de variable,

Afin d'illustrer cette étape, nous utilisons l'exemple de co-clustering Figure 3.3 ainsi que le le Tableau 3.3. Nous générons autant d'individus qu'il y en a dans l'ensemble de données original. Pour cela, nous parcourons un à un les clusters d'individus et nous générons autant d'individus que la population du cluster.

Soit le cluster courant $P(C_3^u)$ qui est peuplé de 49 individus. Pour chacun des individus, nous générons les valeurs pour chacune des cinq variables. Nous détaillons ci-dessous le cas de la variable catégorielle $Class$. La variable $Class$ est repartie sur trois clusters de parties de variables $\{C_1^p, C_3^p, C_5^p\}$. On affecte une modalité à la variable au prorata de cette répartition :

$$P(setosa|C_3^u) = \frac{n_{setosa}}{N_{C_1^p}^{(p)}} \frac{N_{C_3^u C_1^p}}{N_{C_3^u}^u} = \frac{50}{150} \frac{0}{245} \tag{3.1}$$

$$P(virginica|C_3^u) = \frac{n_{virginica}}{N_{C_3^p}^{(p)}} \frac{N_{C_3^u C_3^p}}{N_{C_3^u}^u} = \frac{50}{144} \frac{2}{245} \tag{3.2}$$

$$P(versicolor|C_3^u) = \frac{n_{versicolor}}{N_{C_5^p}^{(p)}} \frac{N_{C_3^u C_5^p}}{N_{C_3^u}^u} = \frac{50}{156} \frac{145}{245} \tag{3.3}$$

Nous normalisons par :

$$P(Class|C_3^u) = P(setosa|C_3^u) + P(virginica|C_3^u) + P(versicolor|C_3^u) \tag{3.4}$$

Et obtenons :

$$P(setosa|C_3^u, Class) = 0, \tag{3.5}$$

$$P(virginica|C_3^u, Class) \approx 0.015, \tag{3.6}$$

$$P(versicolor|C_3^u, Class) \approx 0.985. \tag{3.7}$$

De manière générale, la probabilité s'exprime de la manière suivante :

$$P(X_i|C^u) = \sum_{j=1}^{J} P(X_i = v_j|C^u) \tag{3.8}$$

Où X_i dénote la variable, C^u le cluster id, v_j la valeur j et J le nombre de modalités que prend X_i.

Il est à noter que ces probabilités sont obtenues en utilisant uniquement les paramètres inférés du modèle de co-clustering et qu'aucun accès aux données originales n'est effectué pendant la synthèse.

Pour les individus de C_3^u la valeur $versicolor$ pour la variable $Class$ a ainsi, beaucoup plus de chance d'être affectée que les deux autres modalités.

4.3 Résumé de la méthode

A partir d'un ensemble originale de données T, nous passons par une première phase de préparation. Nous apprenons ensuite un modèle de co-clustering $\mathcal{M}$ à partir des données. Nous simplifions le modèle pour que chaque cluster d'individus soit peuplé d'au moins k individus. Nous générons à partir de cette représentation un ensemble de données fictives T' destiné à la publication.

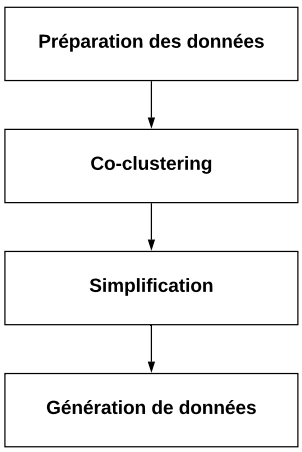


Figure 3.6 – Résumé de la méthode

4.4 Génération de données synthétiques et risque de ré-identification

L'alternative aux méthodes d'anonymisation est la génération de données synthétiques, qui semble avoir l'avantage "philosophique" de contourner le problème de la ré-identification [DFMBMSS06] : puisque les enregistrements publiés sont inventés et ne dérivent d'aucun enregistrement original, certains auteurs affirment qu'aucun individu ayant fourni des données originales ne peut se plaindre d'avoir été ré-identifié. D'autres auteurs [Win04, Rei05] font remarquer qu'à y regarder de plus près, même des données synthétiques pourraient contenir certains enregistrements permettant de ré-identifier des informations confidentielles. En bref, des données synthétiques sur-ajustées aux données originales pourraient mener à la divulgation tout comme les données originales.

Cet avantage des données synthétiques est cependant surtout théorique [DFSSC16] car le modèle doit être construit à partir de l'analyse des données originales. Ainsi, proposer un modèle qui capture de manière appropriée toutes les propriétés de la population n'est, en général, pas faisable : il peut y avoir des dépendances entre des variables difficiles à modéliser ou même à observer dans les données originales. Étant donné que seules les propriétés incluses dans le modèle seront présentes dans les données synthétiques, il est important d'inclure toutes les propriétés des données que nous voulons préserver. Pour réduire la dépendance à l'égard des modèles, des alternatives aux données entièrement synthétiques ont été proposées : données partiellement synthétiques et données hybrides. Toutefois, l'utilisation de ces approches alternatives pour réduire la dépendance à l'égard du modèle a un coût en termes de risque de divulgation.

En ce qui concerne le risque de divulgation, la production de données entièrement synthétiques est considérée comme une approche très sûre. Étant donné que les données synthétiques sont générées uniquement sur la base du modèle ajusté, l'analyse du risque de divulgation des données synthétiques peut être réduite à l'analyse du risque de divulgation de l'information sur les données originales que le modèle incorpore. Comme cette information est habituellement réduite à certaines propriétés statistiques des données originales, le risque de divulgation est sous contrôle. En particulier dans les données

entièrement synthétiques, elles semblent contourner le problème de la ré-identification : puisque les documents publiés sont inventés et ne dérivent d'aucun document original, on pourrait conclure que personne ne peut se plaindre d'avoir été ré-identifié. En y regardant de plus près, cet avantage est moins clair. Si, par hasard, un enregistrement synthétique publié correspond aux attributs non confidentiels d'un citoyen (âge, état civil, lieu de résidence, etc.) et aux attributs confidentiels (salaire, crédit, etc.), la ré-identification à l'aide des attributs non confidentiels est facile et ce citoyen peut avoir l'impression que ses attributs confidentiels ont été indûment révélés. Dans ce cas, il est peu probable que le citoyen soit satisfait ou même qu'il comprenne l'explication selon laquelle le document a été produit synthétiquement.

Cette divergence d'avis sur la question justifie notre recherche d'une solution efficace qui vise à assurer le k-anonymat (Chapitre 4), garantissant de ce fait une protection sûre face aux attaques de ré-identification.

5 Évaluation

On évalue la méthodologie d'anonymisation avec deux jeux de données : la base Adult, banc d'essai traditionnel en matière de fouille de données, et un jeu de données réelles issues de la facturation chez Orange.

5.1 Adult

Pour cette première série d'expérimentations, nous utilisons la base de données Adult [3] qui contient 48842 observations décrites par 14 variables numériques et catégorielles, parmi lesquelles on retient les variables sex, age, race, marital status, education, native country, workclass, occupation, et class. Les variables continues sont discrétisées en déciles, les variables catégorielles ne sont pas modifiées. On réserve 20% des observations (choisies de manière aléatoire) pour les évaluations et on utilise les 80% restants pour la mise en oeuvre du co-clustering et la génération des données synthétiques.

3. `https://archive.ics.uci.edu/ml/`

5.1.1 Variation de k

La grille la plus fine du co-clustering est obtenue pour 32 clusters d'individus et 55 clusters de modalités. A ce niveau, le cluster d'individus le moins peuplé compte 461 individus, le cluster le plus peuplé 1152. On peut remonter dans la hiérarchie du co-clustering jusqu'à ce que tous les clusters soient peuplés du nombre k d'individus désiré ; ainsi pour obtenir $k = 1000$ il faut remonter jusqu'à 15 clusters d'individus. On obtient une représentation plus grossière des clusters d'individus et des clusters de modalités. La Figure 3.7 donne les populations min et max des clusters d'individus obtenus à un niveau donné de la hiérarchie (avec en abscisse le nombre de clusters d'individus et en ordonnée les populations correspondantes).

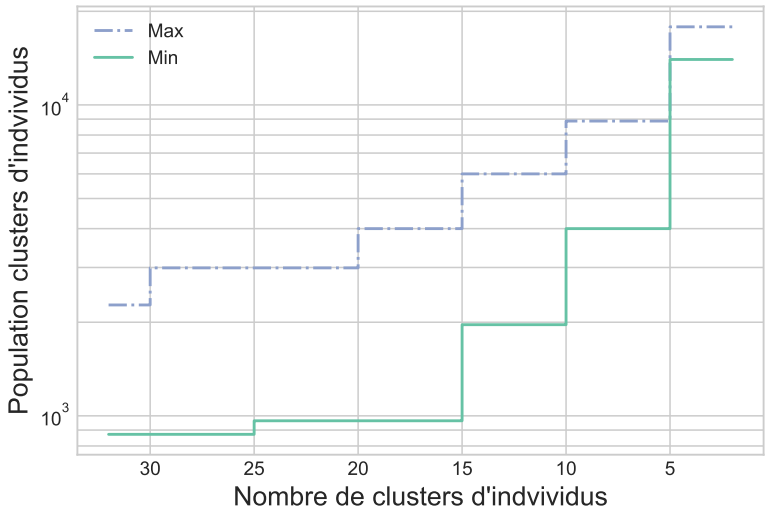


Figure 3.7 – Variation de la population des clusters d'individus

Ci-après, k dénote la population minimale de cluster d'individus à un niveau de hiérarchie donnée.

On génère pour chaque k plusieurs jeux de données synthétiques selon la méthode détaillée Section 4.2.

5.1.2 Évaluation de la qualité

On propose d'évaluer la qualité des données synthétiques produites en montrant que ces données ne peuvent pas être distinguées des données originales. Pour cela, on applique

le protocole ci-dessous :

- construire un nouvel ensemble de données T'' issu de l'union des données originales T et des données synthétiques T',
- ajouter une nouvelle variable catégorielle appelée Origine qui prend la valeur *Rel* ou *Synthtique* selon l'ensemble de provenance de l'individu,
- apprendre un modèle de classification sur D avec comme variable cible *Origine*,
- évaluer la performance de classification. Une mauvaise performance de classification signifiant une bonne qualité des données synthétiques.

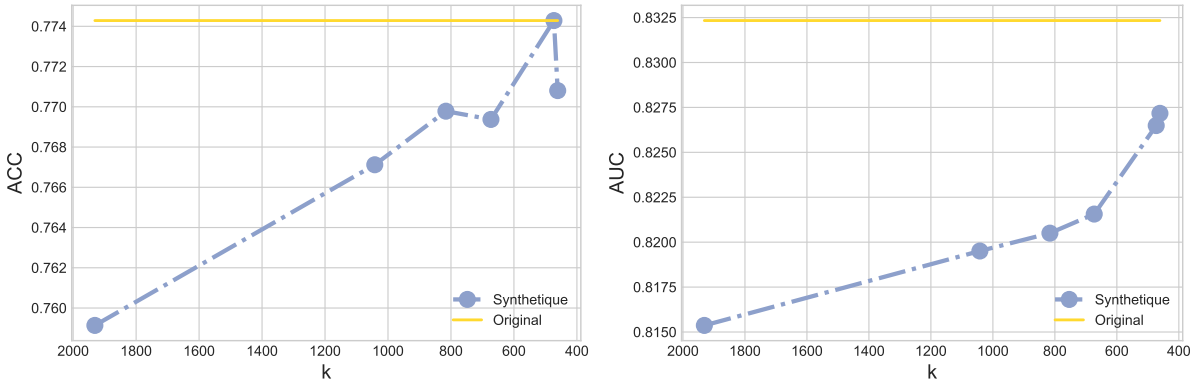
Ici, le modèle de classification utilisé est un classifieur Bayésien naïf avec sélection de variables et moyennage de modèles [Bou08]. Il est implémenté dans la suite logicielle Khiops [4]. Toutes les analyses supervisées qui seront réalisées par la suite le seront à partir de l'outil Khiops. Le jeu de données synthétiques qui est utilisé est celui produit au niveau le plus fin du co-clustering (pour 33 clusters d'individus).

On obtient une performance de classification sur T'' de 0.5 et une AUC de 0.5. Les deux jeux de données synthétique et réel ne peuvent pas être distingués. Le résultat est vérifié pour l'ensemble des jeux synthétiques générés

5.1.3 Évaluation de l'utilité

Deux classifieurs sont appris : le premier avec l'ensemble des données synthétiques générées à un niveau donné du co-clustering, le second avec les données réelles. Puis les deux modèles sont déployés successivement sur les données de test réelles qui avaient été mises de coté précédemment. On évalue ainsi les performances des classifieurs sur les «vrais individus». Les critères qui sont évalués sont le taux de bonne classification (ACC) et l'aire sous la courbe de ROC (AUC). On présente Figure 3.8 les résultats expérimentaux obtenus avec la variable cible Sex. 8 variables explicatives ont été retenues pour l'apprentissage des modèles age, race, marital status, education, native country, workclass, occupation, et class. On donne en abscisse les différentes valeurs de k expérimentées, et en ordonnée l'ACC Figure 3.8a et l'AUC Figure 3.8b obtenus sur l'ensemble de test. On

4. `www.khiops.com`



(a) ACC (b) AUC

Figure 3.8 – Performances prédictives pour l'attribut Sex

Configuration	Original		$k = 461$		$k = 816$		$k = 1042$		$k = 8892$	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
sex	0.774	0.832	0.771	0.827	0.770	0.821	0.767	0.820	0.667	0.500
age	0.323	0.722	0.308	0.707	0.300	0.690	0.295	0.689	0.276	0.645
race	0.871	0.735	0.863	0.718	0.863	0.706	0.863	0.706	0.859	0.476
marital status	0.676	0.849	0.672	0.843	0.672	0.839	0.671	0.837	0.593	0.729
education	0.406	0.719	0.397	0.706	0.393	0.696	0.376	0.686	0.335	0.500
native country	0.920	0.736	0.917	0.714	0.917	0.713	0.917	0.713	0.917	0.500
workclass	0.736	0.741	0.736	0.728	0.734	0.716	0.740	0.722	0.751	0.500
occupation	0.318	0.757	0.303	0.746	0.291	0.734	0.288	0.727	0.179	0.500
class	0.823	0.879	0.820	0.877	0.813	0.874	0.811	0.874	0.762	0.777

Tableau 3.5 – Performances de classification mesurées sur l'ensemble de test avec les différentes configurations selon la variable cible choisie

indique également les performances obtenues quand ce sont les données réelles qui sont utilisées pour l'apprentissage du modèle (droite jaune). La Figure 3.7 fait le lien entre le nombre de clusters d'individus retenus pour construire un ensemble synthétique, à un niveau de la hiérarchie du co-clustering, et le k correspondant.

On observe que les performances de classification obtenues sur les données de test réelles, à partir des modèles appris sur les ensembles synthétiques sont proches de celles que l'on obtient quand ce sont les données réelles qui sont utilisées pour l'apprentissage du modèle. Les performances se dégradent quand le niveau d'agrégation du co-clustering devient élevé.

Différentes variables cibles ont été évaluées de la même manière, les résultats de cette analyse sont rapportés dans le Tableau 3.5. On observe un comportement de classification similaire à celui présenté ci-dessus.

5.1.4 Évaluation de la protection

Pour évaluer le niveau de protection contre la ré-identification d'un ensemble de données synthétiques. On se place dans le contexte où un attaquant qui dispose des données synthétiques, et qui connaît également toute la base initiale, sauf un individu, et on évalue la capacité de retrouver cet individu à partir des informations dont dispose l'attaquant. On se demande ici dans quelle mesure l'individu est caché à l'attaquant ? Le protocole expérimental est le suivant :

— sélectionner une ligne correspondant à un individu dans le fichier original ; on fait l'hypothèse que l'observation sélectionnée est inconnue de l'attaquant,

— l'attaquant peut faire la correspondance entre les 2 bases, réelle et synthétique, en fonction de la connaissance dont il dispose,

— évaluer le niveau d'incertitude de l'attaquant en comptant à combien d'individus synthétiques il peut attribuer l'observation réelle inconnue.

Nous procédons ainsi pour tous les individus du fichier réel.

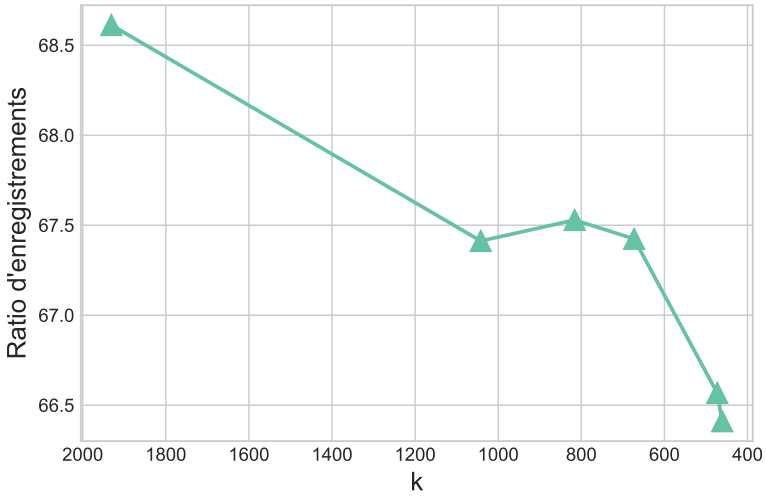


Figure 3.9 – Ratio moyen d'enregistrements qui peuvent être attribués à un enregistrement manquant

Nous présentons les résultats de cette dernière expérimentation Figure 3.9. On donne en abscisse la valeur de k, et en ordonnée le ratio moyen d'enregistrements qui peuvent être attribués à un enregistrement manquant. Ainsi, Pour un ensemble synthétique, généré à partir du niveau le plus fin du co-clustering, en moyenne, une observation réelle inconnue

peut être attribuée à 66,4% des enregistrements synthétiques.

5.2 Base de données de facturation

On expérimente également la méthodologie d'anonymisation avec un ensemble de données réel issues du système de facturation d'Orange. Il s'agit de facturations qui ont donné lieu à réclamation et échange avec le service client qui a éventuellement eu à faire une proposition commerciale. Le jeu données est constitué de 19200 enregistrements. Il est composé de 16 attributs catégoriels et 4 attributs continus.

Prétraitement et mise en forme de la table Les variables sont prétraitées de la même manière que pour l'expérimentation précédente : les quatre variables numériques sont discrétisées en déciles, les variables catégorielles ne sont pas modifiées, puis on re-décrit l'ensemble des variables en deux variables catégorielles. Le coclustering est appliqué ensuite sur les données ainsi préparées.

80% de la base est utilisée pour la construction du co-clustering et la génération des données synthétiques. On conserve 20% des observations originales pour l'évaluation.

Le coclustering au niveau le plus fin est composé de 123 clusters d'individus et 91 clusters de parties de variables.

On trace ci-dessous l'évolution des populations dans les clusters d'individus en fonction de l'agrégation du co-clustering

La procédure de génération synthétique est appliquée à différents niveaux d'agrégation du co-clustering pour 122, 80, 60, 40 et 20 clusters d'individus. Pour chaque niveau, on génère 5 fichiers synthétiques différents.

On cherche à évaluer l'utilité des données synthétiques générées sur un problème de classification supervisée. La variable cible choisie ici est la variable « décision » avec 3 modalités : {compromis, favorable, refus}. Les 19 autres variables sont utilisées comme variables explicatives.

La Figure 3.11 montre l'AUC mesurée en fonction de la valeur de k. On observe un comportement de classification similaire à celui mesuré sur la base Adult. On en conclut

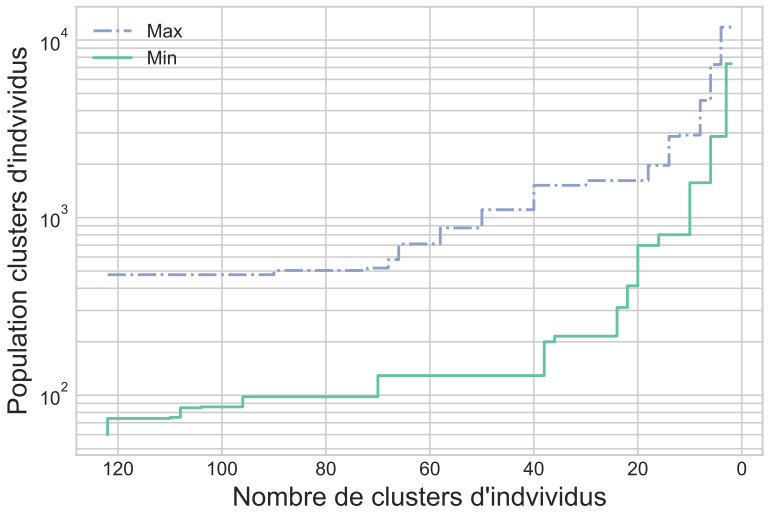


Figure 3.10 – Variation de la population des clusters d'individus, jeux de données de facturation

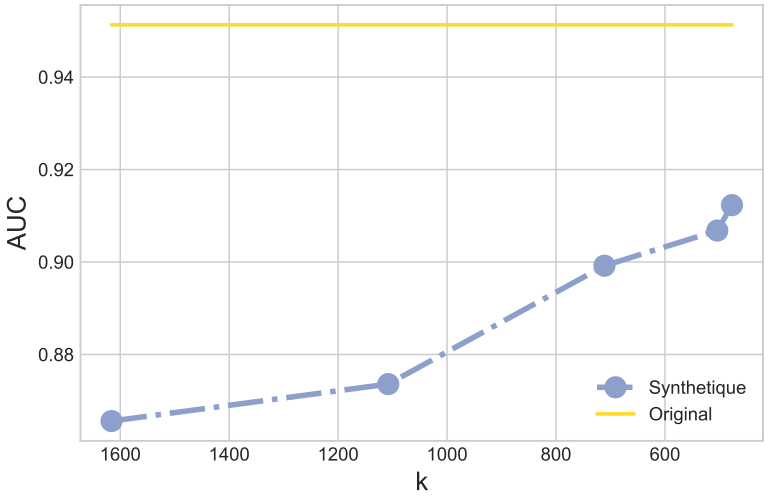


Figure 3.11 – Performances prédictives pour l'attribut décision

que les données synthétiques générées conservent bien les propriétés des données réelles.

6 Conclusion

Dans ce chapitre, nous avons proposé une méthodologie pour anonymiser des microdonnées en vue de leur publication. Dans une première phase, un co-clustering des données partitionne conjointement les individus et les variables descriptives et permet de constituer des groupes d'individus. Obtenir le co-clustering optimal ne nécessite aucun

paramétrage utilisateur. Le co-clustering est également un modèle générateur sur lequel on s'appuie pour construire des individus synthétiques du même format que les individus initiaux. On a montré que les données synthétiques conservent les propriétés des données originales et qu'il est donc possible d'envisager leur utilisation pour la fouille. Enfin, nous avons montré expérimentalement qu'il est difficile de réidentifier un individu.

Chapitre 4

Atteindre le k-anonymat avec le co-clustering

Sommaire

1 Introduction

Dans le chapitre précédent, nous avons décrit notre algorithme d'anonymisation basé sur le co-clustering et la génération de données synthétiques. L'objectif était de fournir une solution pour contrôler le risque de ré-identification dans les données à publier. Même si une telle approche permet de rendre la réidentification difficile pour un attaquant, elle ne fournit pas de garantie formelle sur le niveau de protection apporté. Nous nous tournons donc vers un formalisme de protection qui est le k-anonymat. Dans ce chapitre, l'objectif est de présenter notre nouvelle approche basée sur le co-clustering destinée à assurer le

k-anonymat.

Le k-anonymat est une technique bien maitrisée et l'état de l'art est riche en algorithmes permettant de l'atteindre. Cependant, de tels algorithmes :

— dans le cas non-perturbateur, sont soumis à la *malédiction de la dimensionnalité* et deviennent inefficaces quand le nombre d'attributs QI augmente [Agg05].

— dans le cas perturbateur, peuvent potentiellement noyer l'information dans le bruit qu'ils ajoutent aux données [SDW06].

L'approche que nous proposons doit nous permettre : 1) d'anonymiser un ensemble de données avec potentiellement un grand nombre d'attributs QI, 2) permettre de proposer des échelles d'anonymisation très importantes, en termes par exemple de valeur du cœfficient k, 3) enfin, elle doit nécessiter un minimum d'intervention de la part de l'utilisateur.

D'abord, nous dressons un petit état de l'art des différentes stratégies ainsi que plusieurs algorithmes de la littérature visant à assurer le k-anonymat. Nous décrivons ensuite notre outil de génération de hiérarchie de généralisation automatique. Enfin, nous présentons une approche non-paramétrique inédite basée sur le co-clustering et inspirée des approches de généralisation.

2 Méthodes pour assurer le k-anonymat

Nous distinguons dans la littérature deux principales familles d'approches pour assurer le k-anonymat, les méthodes perturbatrices et non-perturbatrices

2.1 Méthodes non-perturbatrices.

Les méthodes non perturbatrices ne faussent pas les données. Elles s'appuient plutôt sur les principes de *généralisation et de suppression.* Dans ce qui suit, nous allons présenter ces principales techniques ainsi que deux algorithmes exploitant ces techniques dans le but d'assurer le k-anonymat.

2.1.1 Les techniques

La généralisation [Sam01b]. Pour une variable catégorielle V_i, la généralisation (ou recodage global [HDFF+12]) combine plusieurs catégories pour former de nouvelles catégories plus générales, ce qui donne un nouveau V_i' avec $|D(V_i')| < |D(V_i)|$, où $|\cdot|$ est l'opérateur de cardinalité et D le domaine de V_i. Pour une variable continue, la généralisation consiste à remplacer V_i par une autre variable V_i' qui est une version discrétisée de V_i. En d'autres termes, une plage potentiellement infinie $D(V_i)$ est mappée sur une plage finie $D(V_i')$.

La généralisation est un mécanisme largement utilisé pour assurer le k-anonymat. Elle procède à la transformation des valeurs des quasi-identifiants des enregistrements appartenant à une classe d'équivalence en une valeur moins spécifique, de telle sorte que les enregistrements d'une même classe d'équivalence soient indistinguables des autres enregistrement à travers la valeur de leur QI. Pour cela, la généralisation s'appuie sur des hiérarchies de généralisation prédéfinies ; une hiérarchie contenant au moins deux niveaux. Les nœuds feuilles correspondent aux valeurs originelles et constituent le niveau le plus bas. Un lien entre une valeur de niveau $n-1$ et une valeur de niveau n décrit la possibilité de généralisation de la valeur de niveau $n-1$ par la valeur de niveau n. La racine de la hiérarchie représente la valeur la plus générale.

Des exemples de hiérarchies de généralisation pour un attribut catégoriel Figure 4.1 et continu Figure 4.2 sont présentés ci-dessous :

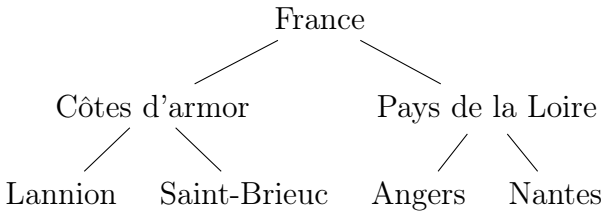


Figure 4.1 – Hiérarchie de généralisation pour l'attribut ville

Le recodage supérieur et inférieur. Le codage supérieur et inférieur est un cas particulier de recodage global qui peut être utilisé sur des variables qui peuvent être classées, c'est-à-dire des variables continues ou catégorielles ordinales. L'idée est que les

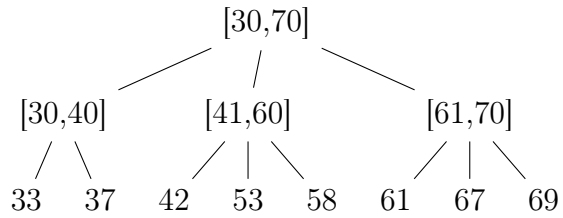


Figure 4.2 – Hiérarchie de généralisation pour l'attribut âge

valeurs supérieures (celles qui dépassent une valeur «limite supérieure») sont regroupées pour former une nouvelle catégorie. Il en va de même pour les valeurs inférieures à un certain seuil [Tem08].

La suppression. La suppression peut être appliquée à des enregistrements entiers (suppression globale). Elle équivaut dans ce cas à échantillonner l'ensemble de données protégé à partir de l'ensemble de données original. La suppression peut être appliquée aux valeurs de quelques attributs dans quelques enregistrements (suppression locale).

Si une combinaison de valeurs de QI est partagée par trop peu d'enregistrements, la suppression globale peut être utilisée pour effacer les enregistrements qui partagent cette combinaison. La suppression locale quant à elle remplace certaines valeurs de QI par des valeurs manquantes, dans le but d'augmenter le nombre d'enregistrements qui s'accordent sur la combinaison de valeurs de QI. La suppression globale est très souvent utilisée en association avec la généralisation, avec l'objectif de satisfaire le k-anonymat sans trop généraliser les données.

Nous allons maintenant présenter trois algorithmes de la littérature qui permettent d'assurer le k-anonymat, un algorithme de recherche exhaustif et deux algorithmes plus récents.

2.1.2 Les algorithmes

L'algorithme MinGen [Swe02a]. Le travail de Sweeney [Swe02a] a introduit l'algorithme de généralisation minimale (MinGen). L'idée principale de cet algorithme est de transformer l'ensemble de données d'origine avec (1) une *généralisation minimale* et (2) une *distorsion minimale*. Si l'on considère les attributs âge (Figure 4.2) et ville (Fi-

gure 4.1) comme étant les QI de la table de données originale T, qui possèdent deux simples hiérarchies à trois niveaux. On peut obtenir à partir de ces deux variables 9 $(3*3)$ combinaisons de niveaux de généralisation différents (y compris l'ensemble de données d'origine). Par exemple, l'ensemble de données ⟨[30,40], Lannion⟩ est moins généralisé que l'ensemble de données ⟨[30,70], Lannion⟩. Si les deux ensembles de données satisfont le k-anonymat, alors l'ensemble de données ⟨[30,40], Lannion⟩ est considéré comme un ensemble de données k-généralisé minimal.

Dans la notion de k-généralisation minimale, il peut y avoir plusieurs ensembles de données k-généralisés minimaux. Pour différencier les multiples ensembles de données k-généralisés minimaux, Sweeney propose de mesurer la distorsion dans la cellule induite par la généralisation à l'aide d'une mesure de précision. La valeur de précision est de 1 si l'ensemble de données généralisé conserve la même granularité que l'ensemble de données d'origine, et de 0 si l'ensemble de données généralisé ne fournit pas de lignes distinctes. Si deux ensembles de données satisfont tous les deux le principe de k-généralisation minimale, on choisit l'ensemble de données qui donne une plus grande précision. En utilisant ces deux notions, *généralisation minimale* et *distorsion minimale*, l'algorithme MinGen est décrit comme suit :

1. Construire les hiérarchies de généralisation pour les attributs QI.
2. Générer tous les ensembles de données généralisées possibles en fonction des hiérarchies de généralisation.
3. Sélectionner des ensembles de données k-généralisés minimaux.
4. Mesurer la précision des ensembles de données sélectionnés.
5. Choisir l'ensemble de données généralisées qui donne le maximum de précision.

L'algorithme évolue de façon exponentielle par rapport au nombre de variables et l'approche devient impraticable, même pour des ensembles de données de taille modeste.

L'algorithme de Samarati [Sam01b]. L'algorithme de Samarati est fondé sur un treillis qui représente les combinaisons possibles des niveaux de généralisation de tous les attributs du quasi-identifiant. Chaque nœud, dans le treillis, contient une liste décrivant

le niveau de généralisation de chaque attribut du QI Samarati part de l'hypothèse que les meilleures solutions sont celles qui impliquent un minimum de généralisations. Par conséquent, son algorithme est destiné à rechercher dans le treillis et à identifier le niveau le plus bas sur lequel se trouvent une ou plusieurs solutions (c'est-à-dire les généralisations qui satisfont le k-anonymat en considérant les suppressions, sans toutefois dépasser le nombre de suppressions autorisé).

1. Soit l'ensemble de données T à protéger (on ne prend en considération ici, que les attributs quasi-identifiants).
2. Considérer la hauteur moyenne dans la zone de recherche (la zone de recherche est initialement l'ensemble du treillis).
3. Vérifier si à cette hauteur, il y a au moins un nœud qui satisfait le k-anonymat avec une suppression minimale (le seuil de suppression est spécifié préalablement),
   (a) Si ce n'est pas le minimum, spécifier la moitié supérieure comme nouvelle zone de recherche.
   (b) Si c'est le minimum, spécifier la moitié inférieure comme nouvelle zone de recherche.
4. Si la zone de recherche comprend plus d'un niveau dans le treillis, répéter l'étape 2. Sinon, retourner une solution résidant à ce niveau.

En d'autres termes, l'approche de Samarati profite du fait que si, à n'importe quel niveau du treillis, une solution peut être trouvée, alors tous les niveaux supérieurs doivent contenir des solutions et donc seuls les niveaux inférieurs doivent être vérifiés. Par conséquent, l'algorithme passe par le treillis avec une recherche binaire, coupant toujours l'espace de recherche en deux, en descendant si une solution est trouvée à un niveau, ou en montant si ce n'est pas le cas. Finalement, l'algorithme trouve la solution(s) avec la hauteur la plus basse, donc avec le moins de généralisations.

Ensuite, comme le suggère Samarati, la meilleure solution à ce niveau (c'est-à-dire avec le moins de perte d'information) par rapport à une préférence donnée (c'est-à-dire la mesure de la perte d'information) est choisie.

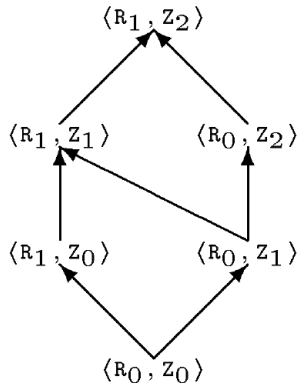


Figure 4.3 – Treillis de généralisation des deux attributs Sexe et Code postal [Sam01b]

L'algorithme Flash. L'algorithme Flash, proposé par Kohlmayer et al. [KPE+12], permet de construire un treillis de généralisation (selon le même principe que l'algorithme de Samarati). Le principal avantage de Flash par rapport à ses concurrents réside dans sa stabilité en terme de temps d'exécution. Pour cela, il met en œuvre une stratégie gloutonne. Il itère sur tous les niveaux du treillis et construit des chemins en effectuant des recherches gloutonnes, en profondeur d'abord, en direction du nœud sommet, jusqu'à ce que le nœud sommet ou un nœud qui n'a pas de successeur non traité soit atteint. Chaque chemin est ensuite parcouru de manière binaire : lorsqu'un nœud est déterminé comme anonyme, l'algorithme marque de manière prédictive ses successeurs et traverse la moitié inférieure du chemin. Si un nœud n'est pas anonyme, il marque de façon prédictive ses prédécesseurs et traverse la moitié supérieure du chemin. Chaque nœud non anonyme est ajouté à une file d'attente prioritaire. Après vérification d'un chemin, de nouveaux chemins sont construits à partir des nœuds de cette file d'attente. Lorsque la file d'attente est vide, l'algorithme poursuit avec la boucle extérieure. L'algorithme s'arrête lorsque la boucle extérieure se termine.

L'algorithme Flash est implémenté dans la boite à outils ARX.

2.1.3 Limites des approches non-perturbatrices

La satisfaction du k-anonymat avec une modification minimale des données en utilisant la généralisation (recodage) et la suppression se révèle être NP-difficile [MW04]. En effet, même la façon de combiner de manière optimale la généralisation et la sup-

pression locale est une question ouverte. Sauf si elles sont soigneusement combinées, ces deux méthodes non perturbatrices peuvent entraîner une perte substantielle de l'utilité des données. De plus, l'utilisation de la généralisation pour assurer le k-anonymat pose plusieurs problèmes pratiques. L'un d'eux est le coût de calcul pour trouver le recodage optimal. Ceci est en partie lié au nombre exponentiel de généralisations pouvant être définies pour chaque attribut.

2.2 Méthodes perturbatrices

L'ensemble de micro-données est déformé avant la publication. De cette façon, des combinaisons uniques de signatures dans l'ensemble de données d'origine peuvent disparaître et de nouvelles combinaisons uniques peuvent apparaître dans l'ensemble de données perturbées ; une telle confusion est bénéfique pour préserver les propriétés statistiques des données. La méthode de perturbation utilisée doit être telle que les statistiques calculées sur les données perturbées ne diffèrent pas de manière significative des statistiques qui seraient obtenues avec les données originales. La micro-agrégation et l'ajout de bruit [Dwo06a] sont des exemples de bruits perturbateurs.

2.2.1 Micro-agrégation

La microagrégation est une approche perturbatrice qui vise à satisfaire le k-anonymat. Définie à l'origine pour les données continues [DN93], elle a ensuite été étendue pour les données catégorielles [Tor04, DFT05]. Quel que soit le type de données, la microagrégation peut être définie en deux étapes principales [DFT05] :

1. **Partitionnement** : l'ensemble des enregistrements originaux est divisé en plusieurs clusters de manière à ce que les enregistrements d'un même cluster soient *similaires* et que le nombre d'enregistrements de chaque groupe soit supérieur ou égal à k.
2. **Agrégation** : un opérateur d'agrégation (par exemple, la moyenne pour les données continues ou la médiane pour les données catégorielles) est calculé pour chaque cluster et est utilisé pour remplacer les enregistrements originaux. En d'autres

termes, chaque enregistrement d'un cluster est remplacé par le prototype du cluster.

La Figure 4.4 illustre les deux étapes de la microagrégation pour un vecteur de données à une dimension. Dans un premier temps, les 3 instances $\{6, 5, 4\}$ sont groupées ensembles, elles sont en suite remplacées dans le vecteur résultant par leur moyenne qui a été choisie comme opérateur d'agrégation.

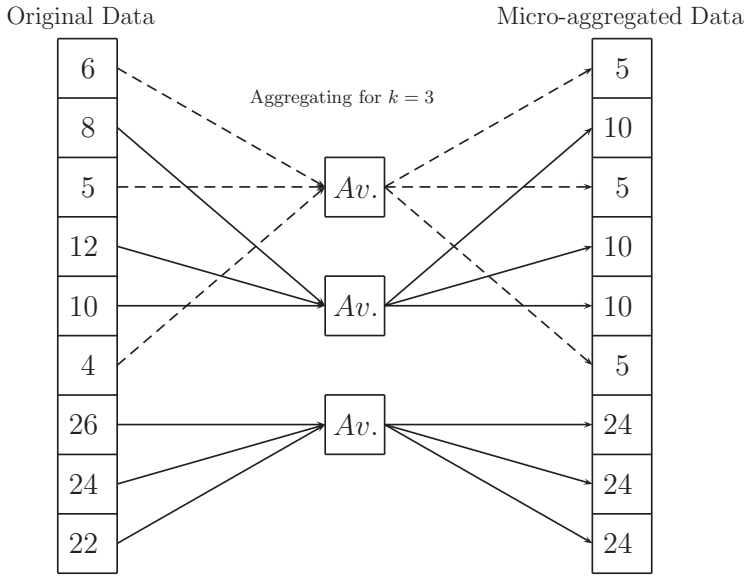


Figure 4.4 – Exemple de k-agrégation avec k = 3 [SSDF08]

Plusieurs taxonomies sont possibles pour classer les algorithmes de microagrégation dans la littérature [DFSSC16] : (i) par rapport à la tailles des groupes, fixe [DN93] ou variable [DFMS02], (ii) algorithme exact (seulement pour le cas univarié, [HM03]) ou la microagrégation heuristique, (iii) catégorielle [DFT05] ou continue [DFMBMSS06]. De plus, selon qu'ils traitent un ou plusieurs attributs à la fois. Ils peuvent être classées en deux catégories : univariée et multivariée. La microagrégation univariée applique la microagrégation à chaque variable indépendamment. La microagrégation multivariée construit des groupes en tenant compte de toutes les variables (ou d'un sous-ensemble) simultanément. Il existe des algorithmes optimaux en temps polynomial pour la microagrégation univariée, mais la microagrégation multivariée optimale a été démontrée comme étant NP-difficile. Pour cette raison, des méthodes heuristiques ont été développées.

La microagrégation assure le k-anonymat seulement lorsque la microagrégation multivariée est appliquée en traitant toutes les variables QI de l'ensemble de données simultanément. Dans le cas contraire, cela n'est pas garanti. La microagrégation univariée est considérée quand on veut obtenir une perte d'information plus faible.

2.3 Comparaison

De nombreux algorithmes on été proposés pour atteindre le k-anonymat, ils peuvent être comparés selon plusieurs axes. Un algorithme peut être : 1) perturbateur, ou non-perturbateur, 2) guidé par une métrique ou non, 3) optimal ou approché, par rapport à la qualité de la table k-anonyme produite ; optimal décrit la solution qui résulte en une perte minimale d'information 4) ascendant ou descendant (concerne les algorithmes basés sur la généralisation). Le tableau 4.1 dresse une comparaison de quelques algorithmes de la littérature.

Algorithme	Perturbateur	Guidé par une métrique	Optimal	Ascendant
MinGen [Swe02a]	✗	✗	✓	✓
Samarati [Sam01b]	✗	✗	✓	✓
Incognito [LDR05]	✗	✗	✓	✓
Flash [KPE+12]	✗	✗	✓	✓
Datafly [Swe97]	✗	✓	✗	✓
μ-argus [HvdWR+03]	✗	✓	✗	✓
MDAV [DFMBMSS06]	✓	✓	✗	N/A

Tableau 4.1 – Comparaison des algorithmes de k-anonymat

3 Génération automatique de hiérarchies de généralisation

Les algorithmes de k-anonymat basés sur la généralisation s'appuient sur les hiérarchies de généralisation des attributs quasi-identifiants. La construction de telles hiérarchies nécessite souvent une expertise humaine et rend les solutions à base de généralisation

difficiles à automatiser. De plus, les hiérarchies conditionnent fortement l'utilité que l'on obtient dans les données protégées.

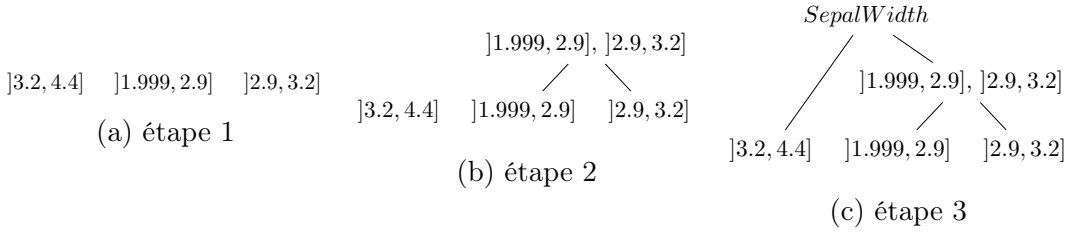
Afin de s'affranchir de la phase de construction de hiérarchies de généralisation qui peut être fastidieuse si le nombre d'attributs QI considérés est important, nous proposons une solution de génération automatique de hiérarchies de généralisation basée sur le co-clustering.

L'idée consiste à créer de nouvelles valeurs généralisées en fusionnant les parties de variable qui se retrouvent dans un même cluster. Nous agrégeons le co-clustering en partant du niveau le plus fin jusqu'au niveau le plus agrégé et créons une nouvelle valeur généralisée à chaque fois que des parties de variable se retrouvent dans un même cluster. On procède de la manière suivante :

1. Initialiser l'arbre de généralisation où toutes les modalités d'un attribut sont des nœuds feuilles.
2. A partir du niveau de co-clustering courant, si deux nœuds ou plus du niveau actuel de l'arbre se retrouvent dans le même cluster de parties de variable, créer un nouveau niveau ainsi qu'un nouveau nœud qui représente la généralisation des nœuds rassemblés dans le cluster. Nous considérons initialement le niveau de co-clustering le plus fin et le niveau 0 de l'arbre.
3. Tant que toutes les modalités ne sont pas rassemblées au sein d'un même nœud racine, utiliser la fonction d'agrégation du co-clustering et retourner à l'étape 2 en considérant le niveau de co-clustering actuel.

Exemple Nous illustrons le fonctionnement de notre solution avec l'exemple de la base Iris. Nous partons du niveau de co-clustering le plus fin (Tableau 3.3a). Nous nous intéressons ici à la variable *SepalWidth* discrétisée en 3 partiles $\{SepalWidth]3.2, 4.4], SepalWidth]1.999, 2.9], SepalWidth]2.9, 3.2]\}$ qui se trouvent respectivement dans les clusters $\{C_2^p, C_6^p, C_7^p\}$, A cette étape, toutes les parties de variable sont des nœuds feuilles (Figure 4.5a). Nous agrégeons ensuite le co-clustering jusqu'à ce que des nœuds du niveau précédent soient rassemblés dans le même clus-

ter de partie de variable. Au niveau suivant (Tableau 3.4a), les parties de variable $\{SepalWidth]1.999, 2.9], SepalWidth]2.9, 3.2]\}$ se retrouvent au sein d'un même cluster $C_{3'}^{p}$ (Figure 4.5b). Nous procédons ainsi jusqu'à satisfaire la condition que toutes les modalités se retrouvent dans un seul et même cluster (Figure 4.5c)



(a) étape 1 (b) étape 2 (c) étape 3

Figure 4.5 – Génération d'une hiérarchie de généralisation pour "SepalWidth"

On procède de la même manière pour la base Adult à partir du co-clustering obtenu au niveau le plus fin. On compare la hiérarchie obtenue pour l'attribut *work-class* (Figure 4.7) avec la hiérarchie fournie par [Iye02](Figures 4.6).

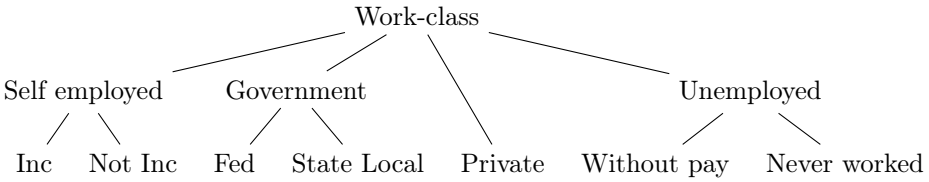


Figure 4.6 – Arbre de la hiérarchie de généralisation pour l'attribut work-class [Iye02]

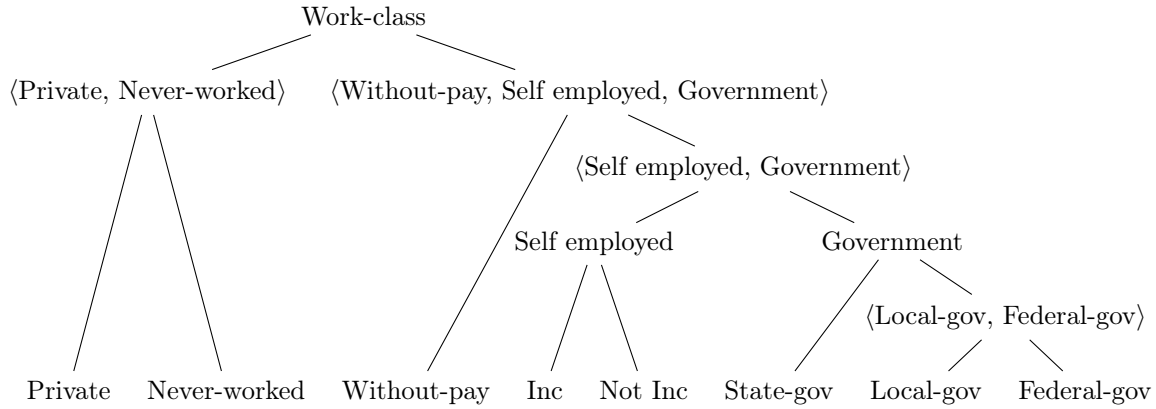


Figure 4.7 – Arbre de la hiérarchie de généralisation pour l'attribut work-class obtenu à l'aide de l'agrégation du co-clustering

Bien que notre solution ne s'appuie que sur la distribution des données, on peut constater qu'elle est en mesure de capturer quelques propriétés sémantiques telles que les catégories *Self employed* et *Government*.

4 Une approche de k-anonymat hybride basée sur le co-clustering

Dans la Section 2, nous avons vu qu'il existe deux types d'algorithmes proposés dans la littérature pour atteindre le k-anonymat : les algorithmes *perturbateurs* et *non-perturbateurs*. Nous avons aussi vu que la satisfaction du k-anonymat avec une modification minimale des données en utilisant la généralisation et la microagrégation multivariée optimale, on été démontrées comme étant NP-difficiles. Nous proposons dans cette section une approche hybride inédite basée sur le co-clustering.

Dans cette approche, on veut exploiter le co-clustering avec des clusters d'individus peuplés d'aux moins k individus, pour produire k individus possédant la même signature (croisement de modalités pour chaque dimension). Les premières étapes de notre solution sont similaires à celles décrites dans la Section 4 du Chapitre 3 : préparation, co-clustering et simplification. Nous partons donc du co-clustering simplifié dans lequel chaque cluster d'individus C_i^u contient aux moins k individus.

Principe de la solution. L'idée de cette approche est de transformer chaque cluster d'individus en une classe d'équivalence. Les valeurs partagées par chaque enregistrement de la classe sont calculées pour chaque variable. Ainsi, pour chaque cluster d'individus, on propose de procéder à des fusions successives de clusters de modalités jusqu'à ce que chaque variable de la classe d'équivalence possède au moins une partie de variable dans l'ensemble fusionné.

Décrire une classe d'équivalence. Pour chaque cluster C_i^u, on ordonne les clusters de modalités de variable C_j^p par ordre décroissant de nombre d'observations (comptes de cellules $C_i^u \times C_j^p$ de la grille). Par la suite, en partant du cluster C_j^p le plus lié au cluster C_i^u on fusionne les clusters C_j^p jusqu'à satisfaire la condition que « chaque dimension (variable) des données est représentée dans le cluster fusionné (par une ou plusieurs modalités) ». A ce stade, toutes les variable possèdent une ou plusieurs modalités. On génère alors autant d'enregistrements que d'individus présents dans C_i^u, et pour chaque

Algorithme 1 : algorithme : génération de données k-anonyme

Require: Matrice de co-clustering

1: **for** i in $1, 2, 3, \ldots I$ **do**
2: Ordonner les clusters de modalités de variable C_j^p du plus lié au moins lié.
3: $Modalites_{C_i^u}$ = le cluster de modalités C_j^p le plus lié au cluster d'individus C_i^u
4: **while** ¬ (chaque variable est représentée dans $Modalites_{C_i^u}$) **do**
5: Fusionner avec $Modalites_{C_i^u}$ le prochain cluster de modalités C_j^p le plus lié à C_i^u, en ignorant les clusters de modalités précédemment fusionnés
6: **end while**
7: **for** d in $1, 2, 3, \ldots D$ **do**
8: **if** Pour l'attribut d il existe une unique modalité dans $Modalites_{C_i^u}$ **then**
9: Assigner la modalité à la classe d'équivalence représentée par C_i^u
10: **else**
11: On calcule dans ce cas une valeur généralisée avec les différentes modalités présentes dans $Modalites_{C_i^u}$ pour l'attribut d et on l'assigne à la classe d'équivalence représentée par C_i^u
12: **end if**
13: **end for**
14: **end for**

variable, différents cas de figure peuvent se produire.

1. Soit les individus appartenant à C_i^u possèdent une modalité unique m, m est donc assignée à l'ensemble des enregistrement de la classe d'équivalence,
2. Soit les individus appartenant à C_i^u possèdent un ensemble de modalités, dans ce cas, on calcule une valeur généralisée qui est assignée aux enregistrement de la classe d'équivalence.

Calcul de valeurs généralisées. Pour le calcul de valeurs généralisées, nous proposons deux variantes :

— Soit utiliser une hiérarchie de généralisation pour trouver la valeur dans l'arbre qui rassemble les différentes valeurs groupées,
— Soit rassembler les valeurs groupées et constituer une nouvelle valeur généralisée.

Une approche hybride. L'approche est dite hybride dans en ce sens que l'affectation de valeurs pour une variable et pour une classe d'équivalence données, est :

— **Non perturbatrice** : si, suite aux fusions successives de clusters de parties de variables, toutes les valeurs présentes pour la classe d'équivalence dans l'ensemble de

données original sont affectées au cluster et prises en compte dans la généralisation,

— **Perturbatrice** : si, suite aux fusions successives de clusters de parties de variables, seul un sous-ensemble des valeurs présentes pour la classe d'équivalence dans l'ensemble de données original est affecté au cluster et est pris en compte dans la généralisation.

Exemple. Reprenons l'exemple du co-clustering de la base Iris Figure 3.3 et Tableau 3.3a. Soit le cluster d'individus C_3^u qui contient 49 individus, comme présenté plus haut (Algorithme 1). La première étape consiste à trier les clusters de modalités de variables $\{C_1^p : 0, C_2^p : 8, C_3^p : 2, C_4^p : 10, C_5^p : 145, C_6^p : 67, C_7^p : 13\}$. Les clusters du plus lié à C_3^u au moins lié sont : $\{C_5^p : 145, C_6^p : 67, C_7^p : 13, C_4^p : 10, C_2^p : 8, C_3^p : 2, C_1^p : 0\}$. Les premières parties de variables assignées à la classe d'équivalence que représente C_3^u sont les parties de variable que contient le cluster C_5^p notamment les parties de variables $\langle Class\{versicolor\}, PetalWidth]0.867, 1.6], PetalLength]2.633, 4.9]\rangle$. A ce stade, la condition d'arrêt n'est pas vérifiée. Il reste des variables qui ne sont pas représentés dans C_5^p, notamment les variables *Sepal length* et *Sepal width*. A l'étape suivante, le second cluster de parties de variable le plus lié à C_3^u est considéré. C_6^p et ses modalités sont assignées à C_3^u, au terme de cette seconde étape la condition d'arrêt est vérifiée. Ainsi, dans la table protégée produite, les individus de C_3^u, constituent une classe d'équivalence de 49 enregistrements dont les valeurs d'attributs sont $\langle Class\{versicolor\}, PetalWidth]0.8, 1.65], PetalLength]2.4, 4.85], SepalWidth]1.999, 2.9]$, $SepalLength]5.4, 6.3]\rangle$.

La complexité algorithmique de la méthode sur laquelle nous nous appuyons est en $O(m\sqrt{m}\log m)$ [Gui13], où m désigne le nombre d'observations. Cette complexité est calculée dans le pire des cas, c'est-à-dire lorsque la matrice des données est pleine. En pratique, l'algorithme est capable d'exploiter l'aspect creux habituellement observé dans ce type de données. Cela permet à notre solution de traiter des ensemble de données avec un grand nombre d'attributs.

5 Évaluation

Cette section vise à évaluer les approches de k-anonymat proposées et qualifier la qualité des données obtenues en utilisant ces approches. Nous utilisons les notations suivante pour faire références aux différentes approches :

- COC_HG fait référence à l'approche basée sur le co-clustering qui vise à assurer le k-anonymat en utilisant les hiérarchies de généralisation pour le calcul de valeurs généralisées.
- COC fait référence à l'approche basée sur le co-clustering qui vise à assurer le k-anonymat sans utiliser les hiérarchie de généralisation pour le calcul de valeurs généralisées.
- ARX fait référence à l'algorithme de k-anonymat *Flash* implémenté dans ARX.
- $MICRO$ fait référence à l'algorithme de k-anonymat par microagrégation.
- $SYNTH$ fait référence à l'algorithme de protection contre le risque de ré-identification basé sur co-clustering et la génération de données synthétique présenté dans le Chapitre 3.

On expérimente avec la base de données Adult [1] qui contient 48842 observations décrites par 14 variables numériques et catégorielles, parmi lesquelles on retient les variables sex, age, race, marital status, education, native country, workclass, occupation et class. Notre approche est capable de prendre en charge une nombre beaucoup plus important d'attributs QI, cependant, l'outil ARX ne permet pas de gérer un nombre supérieur à 10 QI. Nous nous restreignons donc à cette sélection de variables.

La grille la plus fine du co-clustering est obtenue pour 32 clusters d'individus et 50 clusters de modalités. A ce niveau, le cluster d'individus le moins peuplé compte 872 individus.

Nous utilisons les hiérarchies de généralisation construites à l'aide de notre outil (Section 4) pour les approches COC_HG et ARX.

Nous menons trois expérimentations. La première vise à déterminer laquelle des deux variantes proposées COC et COC_HG conserve au mieux l'utilité des données. La

1. https ://archive.ics.uci.edu/ml/

deuxième expérimentation vise à confronter l'approche proposée face aux méthodes de k-anonymat de la littérature ARX et $MICRO$. Enfin, la dernière expérimentation a pour but de qualifier la perte d'information imposée par un formalisme de protection stricte tel que le k-anonymat en comparant les données produite par COC_HG et $SYNTH$.

5.1 Intérêt de l'exploitation des hiérarchies de généralisation

Dans cette première expérimentation, nous utilisons l'algorithme décrit dans la section précédente pour assainir les données. COC_HG exploite les hiérarchies de généralisation, COC quant à lui calcule des valeurs généralisées sans exploiter les hiérarchies de généralisation.

Afin de mesurer l'utilité des données, nous utilisons la distance de Hellinger, comme présentée dans la Section 5.1.1 du Chapitre 2. Nous calculons la distance entre les distributions de probabilités des données originales, ainsi que celles obtenues en utilisant COC_HG et COC. Certaines valeurs étant généralisées, nous faisons donc l'hypothèse de l'uniformité au sein des valeurs généralisées pour estimer la distribution de chaque attribut. La Figure 4.8 représente la moyenne des distances de Hellinger mesurées sur chaque attribut en ordonnées, en fonction de la taille de k représentée en abscisse.

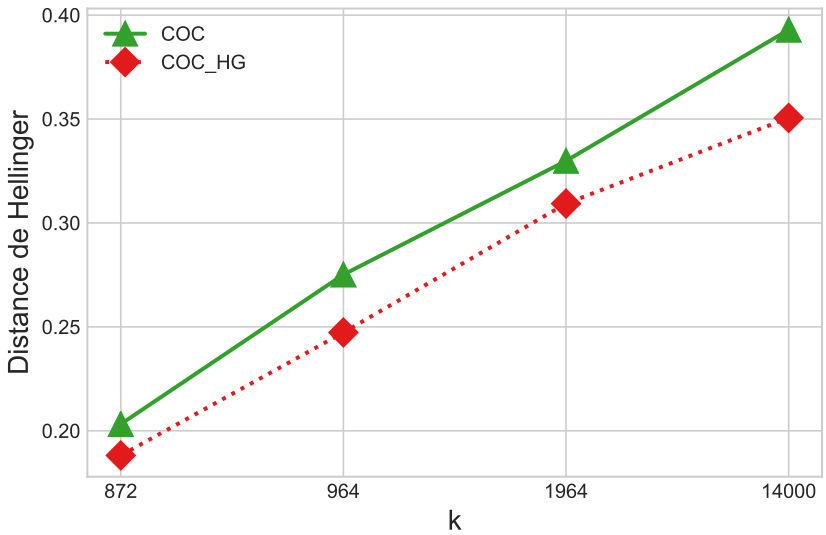


Figure 4.8 – Moyenne des distributions univariées COC vs COC_HG

On observe que les distributions de données originales sont de moins en moins conser-

vées à mesure que la valeur de k croît, et que pour le jeu de données testé la variante COC_HG qui tire partie des hiérarchies de généralisation conserve mieux les distributions, indépendamment de la valeur de k. Nous en déduisons que COC induit une plus grande perturbation sur le jeu de données testé, perturbation qui est atténuée dans COC_HG qui permet le calcul de valeurs généralisées moins distordues.

5.2 Contribution vs algorithmes de la littérature

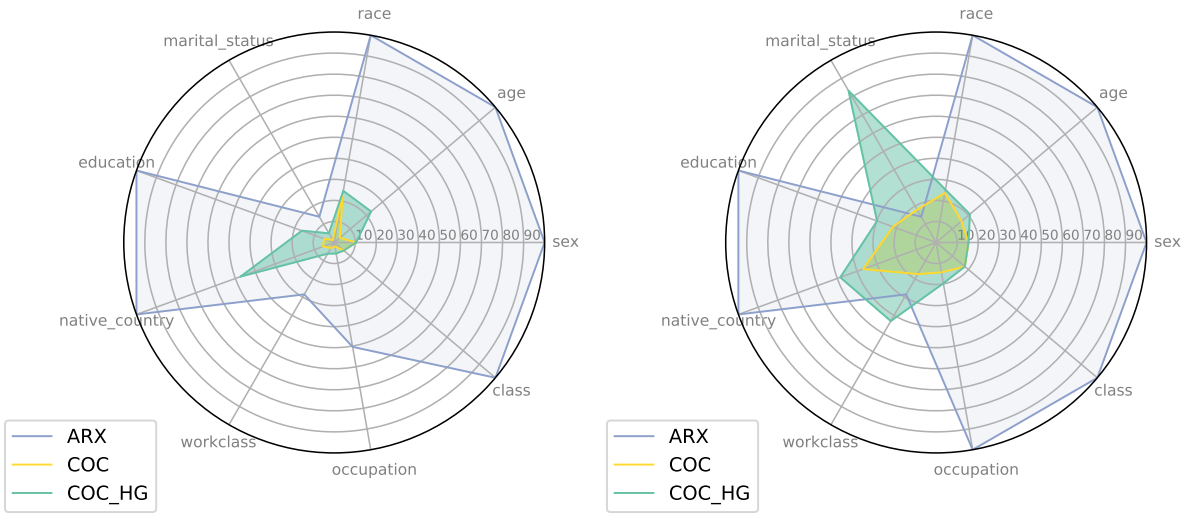
Dans cette expérimentation, nous comparons notre approche de k-anonymat à deux solutions de la littérature. La première est l'algorithme implémenté dans ARX. Nous avons utilisé le logiciel disponible sur le site de l'outil d'anonymisation [2]. Le deuxième est la microagrégation. Nous n'avons pas trouvé d'implémentation disponible pour atteindre le k-anonymat via la microagrégation dans le cas de données discrètes. Nous avons donc mis en œuvre l'algorithme correspondant en partant du co-clustering à un niveau d'agrégation donné, puis en créant, pour chaque cluster d'individus peuplé d'au moins k-individus, une classe d'équivalence à laquelle est assignée pour chaque variable, la partie de variable .la plus fréquente du cluster d'individus.

Dans l'implémentation de l'algorithme ARX, la valeur de k est limitée à 1000. Nous produisons donc des données assainies avec cette méthode uniquement pour les configurations $k = 872$ et $k = 964$. Ce sont les deux seules configurations obtenues à travers le co-clustering dont le k est < 1000.

Métrique de classification. Nous utilisons la métrique de classification (Section 5.1.1) du Chapitre 2 qui mesure la perte d'information pour les méthodes basées sur la généralisation. Nous ne considérons donc pas la microagrégation dans cette expérimentation. Les résultats de cette évaluation sont synthétisés par le diagramme en radar Figure 4.9 pour les configurations $k = 872$ (4.9a) et $k = 964$ (4.9b).

La métrique de classification est une mesure à minimiser. Plus la surface du polygone est étendue, plus la valeur de la métrique de classification est importante, indiquant de

2. https ://arx.deidentifier.org/downloads/



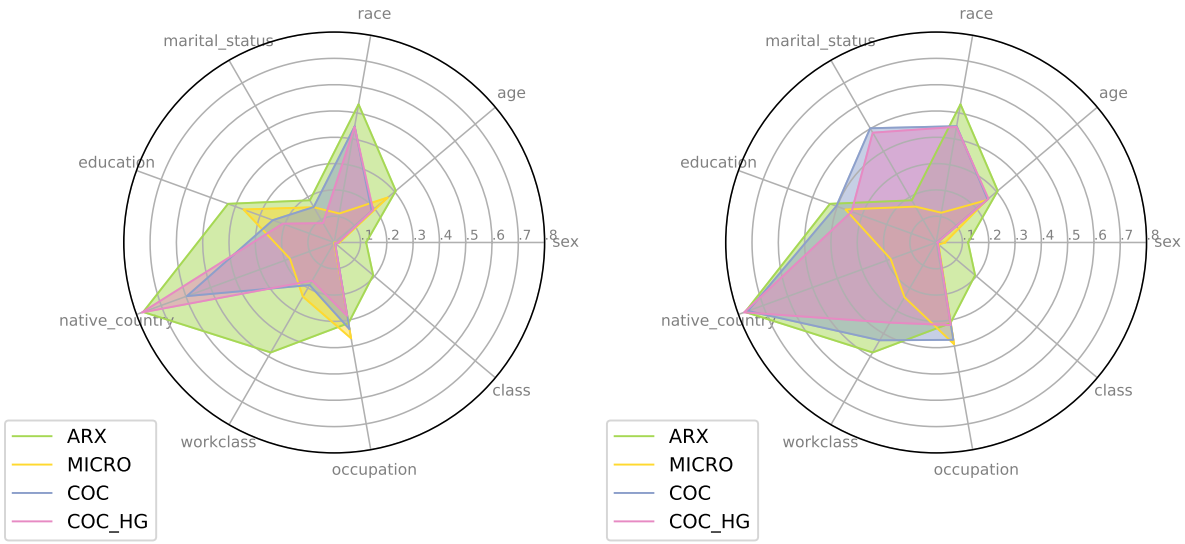
(a) $k = 872$ (b) $k = 964$

Figure 4.9 – Métrique de classification

ce fait une généralisation proportionnellement importante. Dans les deux cas, on observe que ARX produit les données les plus généralisées, étant une méthode exacte et non perturbatrice ce résultat est attendu. Pour les deux configurations testées $k = 872$ et $k = 964$, COC qui n'est pas contrainte d'utiliser les hiérarchies de généralisation, produit logiquement des données moins généralisées que celle produite par COC_HG.

Distribution univariée. Nous confrontons les différentes approches en terme de conservation des distributions de probabilités univariées. Nous faisons l'hypothèse de l'uniformité pour calculer des vecteurs de distributions de probabilité pour chaque attribut. La distance de Hellinger entre la distribution originale et celle des données perturbées est rapportées dans les diagrammes en radar, Figure 4.10 pour les configurations $k = 872$ (4.10a) et $k = 964$ (4.10b).

La distance de Hellinger est elle aussi à minimiser. Plus la surface du polygone est étendue, plus la distance entre les distributions des données assainies et les données originales est importante. On observe pour les deux configurations testées que les données obtenues avec ARX sont les plus éloignées des données originales. Quand $k = 872$, COC et COC_HG présentent les meilleurs résultats pour tous les attributs mise à part $race$ et $native\_country$ où $MICRO$ les surpasse. Pour $k = 964$, $MICRO$ présente de meilleurs résultats que COC et COC_HG notamment pour les attributs $marital\_status$



(a) $k = 872$ (b) $k = 964$

Figure 4.10 – Distances de Hellinger pour chaque attribut

et *native_country* où ces approches produisent une distribution très éloignée de la distribution originale. Pour expliquer ces résultats, nous effectuons un examen de la composition des clusters de parties de variables aux niveaux $k = 872$ et $k = 964$. Il en ressort que les parties de variables des attributs *race* et *native_country* se retrouvent rassemblées dans quelques clusters de parties de variables ce qui rend difficile de les répartir de manière pertinente selon le cluster d'individus considéré. Dans ce cas, la distribution s'apparente à une distribution uniforme où toutes les parties de variable sont réparties de manière équiprobable.

Distribution multivariée. Nous confrontons les différentes approches en terme de conservation des distributions de probabilités jointes. Nous faisons l'hypothèse de l'uniformité pour calculer des vecteurs de distributions de probabilités jointes allant de deux à six attributs. La distance de Hellinger entre les distributions originales et celles des données assainies est présentée Figure 4.11 pour le cas $k = 872$ et Figure 4.12 pour le cas $k = 964$.

On observe que les distributions jointes obtenues avec COC_HG sont les plus proches des distributions originales pour $k = 872$ et $k = 964$. Nous remarquons aussi que quand $k = 964$, $MICRO$ fournit les résultats les moins intéressants même si lors de l'expéri-

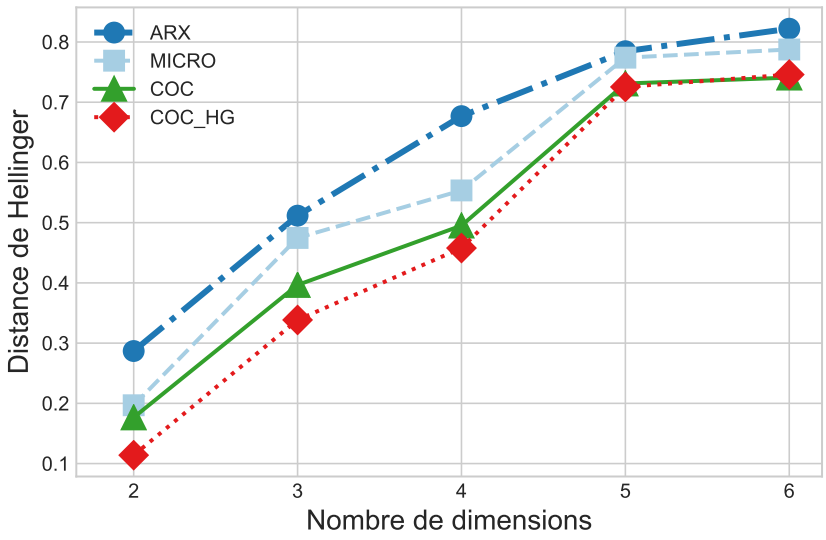


Figure 4.11 – Distribution multivariée $k = 872$

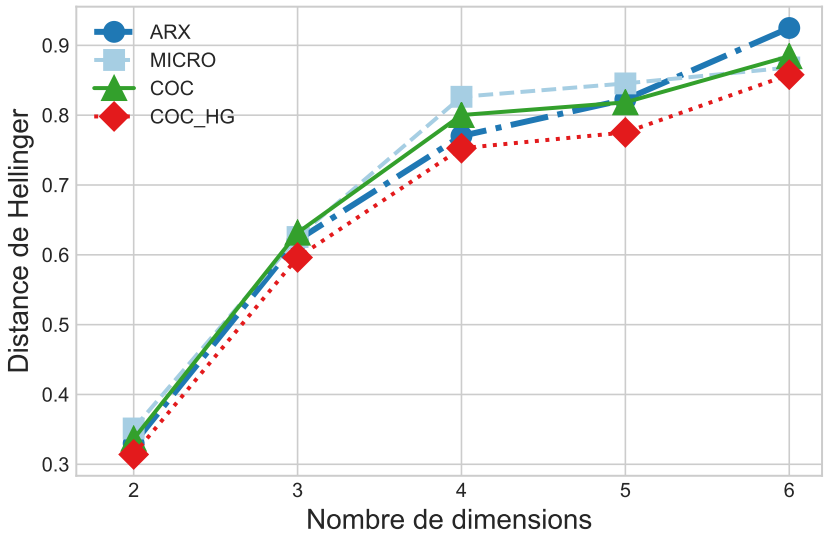


Figure 4.12 – Distribution multivariée $k = 964$

mentation précédente, et pour cette valeur de k, les distributions univariées des données produites par $MICRO$ étaient les plus proches des données originales. Nous expliquons cela par le fait de choisir la valeur la plus fréquente au sein d'un cluster, qui conserve bien la distribution globale mais ne suffit pas à capturer les corrélations entre les attributs.

5.3 k-anonymat vs génération de données synthétiques

Dans cette dernière expérimentation, le but est de mesurer la perte d'information qu'induit le respect d'un formalisme de protection tel que le k-anonymat. Nous compa-

rons les données produites avec l'approche présentée au Chapitre 3 et celles produites avec les approches du k-anonymat. Nous utilisons le même co-clustering et nous produisons des données en choisissant les même valeurs de k pour les approches COC_HG, COC et $SYNTH$. Nous faisons l'hypothèse de l'uniformité pour calculer des vecteurs de distributions de probabilités pour les attributs des données de COC_HG et COC. La distance de Hellinger moyenne entre la distribution originale et celle des données perturbées pour tous les attributs est rapportée Figure 4.13. Les distances entre les distributions jointes dans les données originales et en utilisant COC_HG et COC et $SYNTH$ pour $k = 872$ sont illustrées Figure 4.14

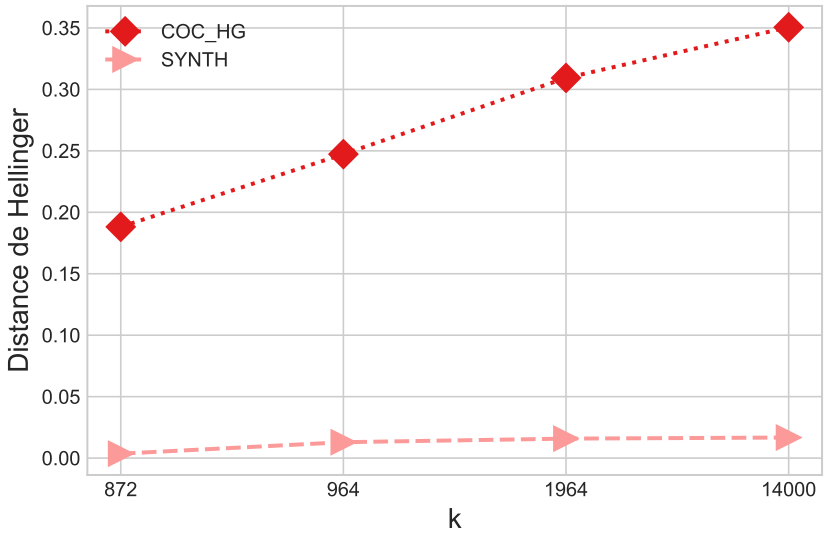


Figure 4.13 – Moyenne des distributions univariées $SYNTH$ vs COC_HG

On peut voir, que les probabilités dans le cas univarié et multivarié sont mieux conservées dans les données produites par $SYNTH$. Les résultats obtenus illustrent parfaitement le compromis utilité/protection. La qualité des données synthétisées sans respecter le k-anonymat, est sensiblement inférieure à celle des données k-anonymes.

5.4 Synthèse

Sur le jeu de données testé, COC_HG présente globalement, les meilleures résultats en terme de préservation des distributions de probabilités jointes et univariées. Il constitue la meilleure alternative parmi les algorithmes de k-anonymat testés

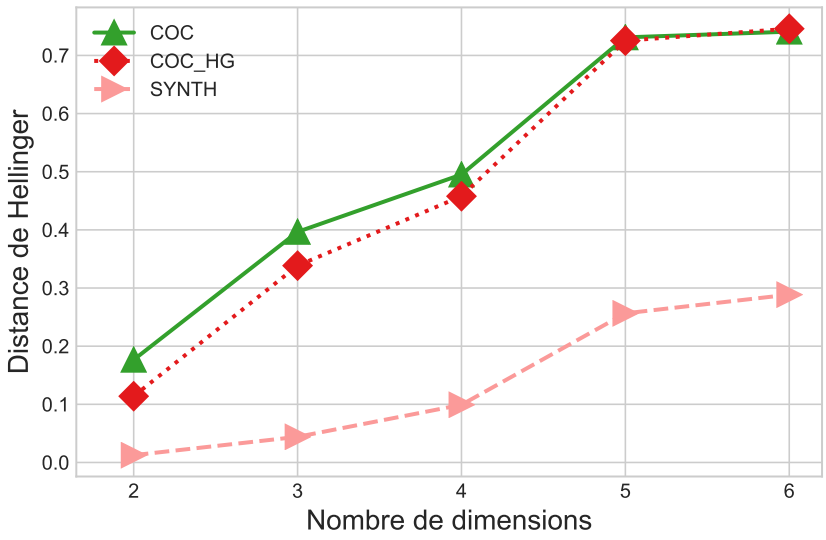


Figure 4.14 – Distribution multivariée $SYNTH$ vs COC, COC_HG

$\{COC, ARX, MICRO\}$. Cependant, la qualité des données produites avec COC_HG demeure nettement inférieure à celles des données générées via $SYNTH$.

6 Conclusion

Dans ce chapitre, nous avons exploré quelques algorithmes de l'état de l'art qui assurent le k-anonymat. Nous avons ensuite décrit notre outil de génération automatique de hiérarchies de généralisation. Enfin, nous avons proposé une nouvelle approche pour atteindre le k-anonymat. Notre solution est basée sur le co-clustering et est hybride dans le sens où elle peut être perturbatrice pour certaine classes d'équivalence et non-perturbatrice pour d'autres, selon que toutes les valeurs de la classe d'équivalence, dans l'ensemble de données original soient affectées aux clusters et prises en compte dans la généralisation, ou non. Nous avons montré par des expérimentations que notre approche offre une utilité satisfaisante et compétitive face aux algorithmes de l'état de l'art. Cependant, le principal avantage de notre approche réside dans son efficacité : elle permet de traiter un nombre important d'attributs quasi-identifiants en un temps raisonnable. Les limites de notre approche sont elle liées au manque de liberté dont dispose l'utilisateur pour fixer le paramètre k et est conditionnée par la convergence de l'approche MODL pour trouver un co-clustering informatif.

Chapitre 5

Génération de données respectueuse de la confidentialité différentielle

Sommaire

1 Introduction

Dans ce chapitre, nous étudions le problème de la génération de données respectueuse de la confidentialité différentielle. Nous considérons le modèle non-interactif et cherchons à publier des données synthétiques. D'abord, nous faisons un tour d'horizon des approches de la littérature pour la publication de données différentiellement privée. Ensuite, nous présentons notre algorithme DPCocGen, une stratégie de partitionnement en deux phases basée sur le co-clustering pour la génération de données synthétiques. Enfin, nous évaluons expérimentalement l'utilité des données publiées à l'aide de notre solution, en mesurant la préservation des propriétés statistiques et la performance prédictive des données synthétiques.

2 Publication de données différentiellement privée

Dans cette section, nous allons voir quelques méthodes de l'état de l'art, qui ont été proposées pour la publication de données différentiellement privée. Toutes ces méthodes partent d'un ensemble de données original et visent à publier soit un ou plusieurs histogrammes, soit un ensemble de données synthétiques.

La majorité des méthodes s'intéresse aux histogrammes [MCFY11, ZCP+14, RKS16]. L'objectif dans ce cas, est de calculer de manière différentiellement privée des histogrammes pour chaque attribut ou des histogrammes joints pour un petit nombre d'attributs, puis d'essayer de générer un histogramme joint pour tous les attributs à partir des histogrammes partiels.

Un histogramme représente une approximation de la distribution des données. Cette structure est obtenue en partitionnant le domaine des données et en comptant le nombre d'enregistrements appartenant à chaque partition (classe). Les histogrammes sont rendus différentiellement privés en bruitant les comptes avec un mécanisme lui-même différentiellement privé [DMNS06]. Cependant, outre les comptes, le partitionnement des données à lui seul peut potentiellement révéler de l'information. Un moyen de prévenir des divulgations issues du partitionnement consiste à utiliser un partitionnement prédéfini, et donc indépendant des données que l'on considère.

Pour des ensembles de données avec des régions faiblement peuplées, l'utilisation d'un partitionnement prédéfini peut s'avérer problématique. La précision obtenue avec un histogramme calculé de manière différentiellement privée dépend de :

1. La taille des classes (ou bins) de l'histogramme : plus les classes sont étendues, moins la valeur d'attribut est précise, augmentant ainsi l'*erreur d'approximation.*
2. Le nombre d'enregistrements appartenant aux classes : plus ce dernier et petit, plus l'ampleur du bruit est importante, augmentant ainsi l'*erreur de perturbation.*

2.1 Exemple illustratif

L'exemple suivant illustre la complexité de trouver un partitionnement optimal pour un histogramme bidimensionnel. On peut imaginer une table de N individus décrits par les attributs *age* et *income*, et l'histogramme bidimensionnel le représentant. Supposons que les valeurs de l'attribut *age* sont $20-30, 30-40$ et $40-50$, et que les valeurs de l'attribut *income* sont $0-10K, 10-20K$ et $> 20K$. Chaque cellule de l'histogramme représente le nombre d'individus qui correspondent à la valeur de *age* et *income*.

Les histogrammes 5.1 et 5.2 représentent les histogrammes calculés de manière différentiellement privée. l'histogramme 5.1 est le résultat de l'application du mécanisme Laplacien directement sur l'histogramme de départ [DMNS06]. L'histogramme 5.2 est quant à lui, obtenu en regroupant d'abord les cellules, puis en injectant un bruit Laplacien dans chaque partition. Dans les deux cas, le bruit ajouté à chaque cellule est indépendant du compte, il ne dépend que de la sensibilité (1 pour les requêtes de dénombrement) et du paramètre ε.

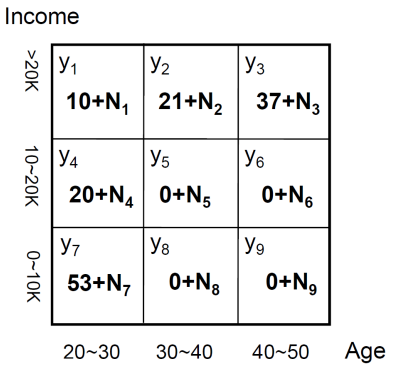


Tableau 5.1 – Histogramme avec un partitionnement fin [XXFG12]

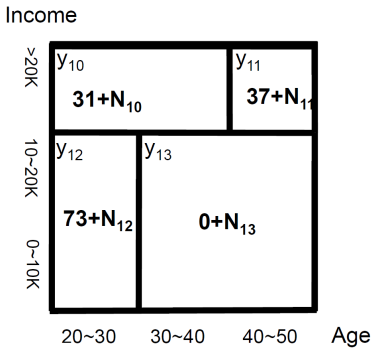


Tableau 5.2 – Histogramme avec un partitionnement optimal au sens de l'erreur de perturbation [XXFG12]

En général, un partitionnement fin introduira une petite erreur d'approximation mais engendrera une grande erreur de perturbation. Trouver le bon compromis est donc le principal enjeu. Le problème du choix du partitionnement optimal, même si l'on omettait la confidentialité, et qu'on se restreignait à deux dimensions, est un problème NP-complet [MPS99]. Plusieurs stratégies ont été proposées dans la littérature, pour améliorer la précision des requêtes de calcul d'histogramme de manière différentiellement privée.

Nous en présenterons quelques une dans la section qui suit :

2.2 Algorithmes de la littérature

Il existe dans la littérature plusieurs algorithmes de publication de données différentiellement privée. En général, ces approches tentent : (1) soit, de réduire le domaine de description des données en l'agrégeant afin de limiter l'effet de l'ajout de bruit sur les comptes, (2) soit, de transformer les données originales sous une autre forme, dans le but de les perturber plus efficacement.

2.2.1 LPA [DMNS06]

La méthode publie des histogrammes en ajoutant un bruit aléatoire Laplacien à chaque compte de cellule de l'histogramme original. Elle est considérée comme un méthode de référence. Cependant, même si la méthode donne de bons résultats pour les petites dimensions, elle devient problématique pour les données en grande dimension, où le domaine de description explose et les comptes dans chaque histogramme deviennent extrêmement petits et de ce fait, beaucoup plus sensibles à l'ajout de bruit. Il en résulte une erreur de perturbation importante.

2.2.2 Privelet+ [XWG11]

Dans Privelet, l'histogramme original H est d'abord transformé en ondelette. Une ondelette est une fonction linéaire inversible qui fait correspondre H à une matrice C, telle que chaque entrée de C est une combinaison linéaire des entrées de H, et H peut être reconstruit à partir de C sans perte. On désigne les éléments de C par les cœfficients de l'ondelette.

Un bruit Laplacien est ajouté à chaque cœfficient de C de manière à assurer la ε-confidentialité différentielle. On obtient à l'issue de cette étape une ondelette bruitée C'.

Dans la dernière étape, Privelet raffine (éventuellement) l'ondelette perturbée C', puis fait correspondre C' à un histogramme H' qui sera retourné en sortie. Le raffinement peut

modifier C', mais il n'utilise aucune information issue des données originales D ou de H. Cette dernière étape ne dépend que de C'. Ceci garantit que Privelet ne divulgue aucune information de D, mise à part celle divulguée dans C'.

Une version de Privelet baptisée Privelet+ a été proposée [XWG11] dans le but de traiter des données multidimensionnelles avec des attributs discrets qui possèdent un petit nombre de valeurs. En premier lieu, Privelet+ divise l'histogramme original en sous-histogrammes à travers les dimensions des attributs, puis applique Privelet pour chaque sous-histogramme, et finalement assemble l'ensemble des sous-histogrammes perturbés en un seul histogramme.

2.2.3 NoiseFirst et StructureFirst [XZX+13]

Les auteurs proposent deux approches. La première, NoiseFirst est basée sur LPA. Dans un premier temps, un bruit aléatoire Laplacien est ajouté à chaque compte comme dans [DMNS06]. Vient ensuite une étape de post-optimisation dans laquelle les auteurs utilisent une technique de programmation dynamique de manière à construire un nouvel histogramme en fusionnant les comptes bruités. J. Xu et al. [XZX+13] s'appuient sur la loi des grands nombres pour affirmer que le calcul de la moyenne sur les comptes perturbés adjacents a pour effet d'éliminer l'impact du bruit Laplacien avec le paramètre de position à 0.

La deuxième approche, StructureFirst consiste à construire un histogramme optimal en utilisant la technique de programmation dynamique afin de déterminer les limites des classes à fusionner. La structure de cet histogramme optimal est ensuite perturbée via un mécanisme exponentiel. Enfin les moyennes des comptes agrégés des classes obtenues sont à leur tour bruités à l'aide du mécanisme Laplacien. NoiseFirst et StructureFirst induisent une complexité calculatoire importante et sont de ce fait, peu adaptés à la publication d'histogrammes issus de données en grande dimension.

2.2.4 P-HPartition [ACC12]

L'idée ici est d'effectuer un clustering hiérarchique descendant sur les classes (bins) d'un histogramme. Les classes de l'histogramme qui appartiennent à un même cluster ont des comptes similaires, et peuvent ainsi être approximés par un compte moyen (le centroïde du cluster). Les auteurs affirment qu'il suffit alors de perturber les centroïdes des clusters, qui ont l'avantage de posséder une plus petite sensibilité.

L'enjeu ici, est de rendre l'algorithme de clustering différentiellement privé. Pour cela, les auteurs proposent de tirer partie d'une structure arborescente de clustering. la partition initiale qui contient l'ensemble des n classes représente la racine de l'arbre. Chaque bissection divise une partition (représentée par un nœud dans l'arbre) en deux nœuds fils. Une bissection consiste en deux étapes :

1. Sélectionner une partition qui n'a pas été bissectée plus de d fois, d est choisi de manière indépendante des données, afin de garantir la confidentialité différentielle.
2. Bissecter la partition sélectionnée en utilisant le mécanisme exponentiel.

Dans le but de garantir que la sélection de la partition à bissecter est indépendante des données, la méthode maintient une file de toutes les partitions, et choisi toujours la première partition dans la file. Après chaque bissection, les sous-partitions résultantes sont ajoutées à la fin de la file.

2.2.5 DPCube [XXFG12]

DPCube est une méthode de publication d'histogrammes multidimensionnels (cubes). La première phase consiste à construire un histogramme multidimensionnel fin avec des classes à largeur égale, en se basant sur les domaines (indépendamment des données). Les comptes de l'histogramme sont perturbés à la manière d'un LPA et donnent une approximation de la distribution des données originales. Un ensemble de données synthétiques D_c est généré à partir des comptes de l'histogramme.

Dans la seconde phase, un partitionnement de type arbre kd-multidimensionnel est construit à partit de D_c afin d'obtenir des partitions uniformes et de réduire ainsi l'erreur induite par l'estimation. La racine de l'arbre contient tout l'espace. A chaque étape, on

choisit à l'aide d'une heuristique une dimension à diviser, ainsi qu'une valeur de division pour l'intervalle courant. Les clés du partitionnent ainsi obtenues sont utilisées pour partitionner les données originales, puis, de la même manière, les comptes agrégés après le partitionnement sont bruités à l'aide du mécanisme Laplacien.

L'apport de l'approche réside dans le fait que le partitionnement en arbre kd s'adapte indirectement à la distribution des données, grâce à l'approximation qui résulte de la première phase. La base de données n'est pas requêtée pendant la construction de l'arbre, ce qui permet d'économiser le budget de protection. Le budget de protection global ε est divisé entre les deux phases uniquement pour bruiter les comptes.

2.2.6 DP-MHMD [RKS16]

Les auteurs proposent un modèle de génération synthétique des données de mobilité humaine. Le modèle prend en entrée le jeu de données original et supprime les enregistrements aberrants. Ensuite, un test de *Chi-square* pour l'indépendance et l'homogénéité est mené afin d'obtenir des groupes d'attributs non disjoints regroupés en fonction de leur associativité. Les auteurs estiment que le groupement des attributs relève de l'information publique, il ne représente donc pas de menace pour la confidentialité différentielle. Le modèle construit des tables de contingences à partir de chaque groupe d'attributs qu'il bruite grâce au mécanisme Laplacien. Enfin, des données synthétiques sont tirées à partir des distributions de chaque groupe et sont agrégées pour former un ensemble.

2.2.7 DiffGen [MCFY11]

DiffGen est une méthode de publication de données assainies pour la tâche de fouille de données. Les auteurs s'intéressent aux données avec un attribut de classe catégoriel A^{cls} ainsi qu'un ensemble d'attributs $A^{pr} = \{A_1^{pr}, A_2^{pr}, ..., A_d^{pr}\}$ qui peuvent être continus ou catégoriels. Pour les attributs catégoriels, une taxonomie doit être fournie, un exemple de taxonomie est illustré Figure 5.1.

Tout d'abord, les attributs A^{pr} sont généralisés, créant ainsi des classes d'équivalences où tous les enregistrements au sein d'un groupe possèdent les mêmes valeurs d'attributs.

Le but étant de publier les comptes bruités des groupes, la démarche consiste à bruiter les données originales par une succession de partitionnements (figure 5.2). En partant de l'état le plus général, l'algorithme procède à des spécialisations en choisissant un attribut à spécialiser ainsi qu'une valeur d'attribut dans la partition courante. A chaque itération, DiffGen choisit de manière probabiliste un candidat pour la spécialisation, le mécanisme exponentiel est exploité avec comme fonction d'utilité pour calibrer le bruit une mesure de gain d'information. Après un certain nombre d'itérations h donné en paramètre, les spécialisations sont arrêtées et un bruit laplacien est ajouté aux comptes originaux de chaque groupe.

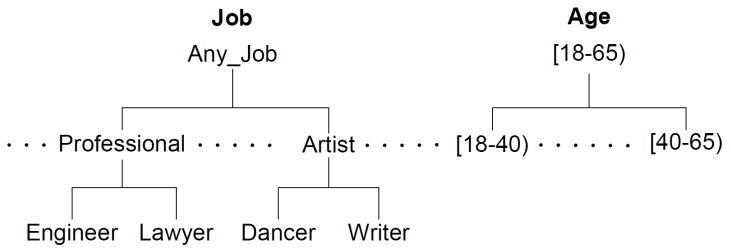


Figure 5.1 – Taxonomie des attributs Job et Age [MCFY11]

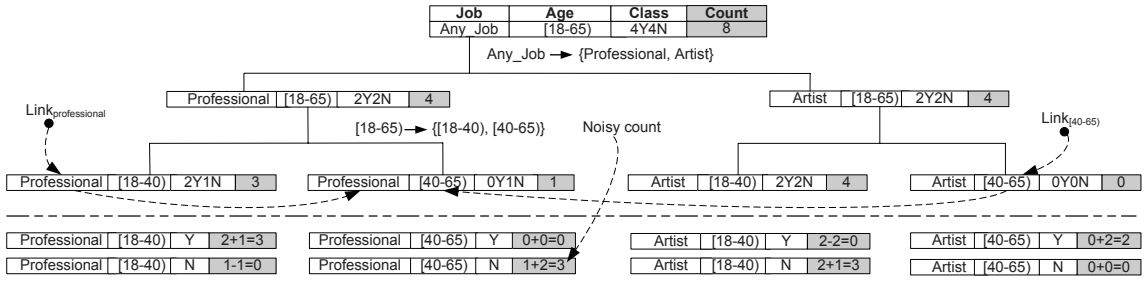


Figure 5.2 – Arbre de partionnement des enregistrements [MCFY11]

2.2.8 PrivBayes [ZCP+14]

[ZCP+14] proposent PrivBayes, une méthode de publication différentiellement privée pour les données en grande dimension. Étant donné un ensemble de données original D, PrivBayes construit un réseau bayésien N, qui d'une part modélise de manière succincte les corrélations entre les attributs de D, et d'autre part, permet d'approximer la distribution des données de D en utilisant un ensemble P de marginales à faible dimension de D.

L'étape suivante consiste à injecter du bruit dans chaque marginale de P afin d'assurer la confidentialité différentielle, et ainsi utiliser les marginales bruitées et le réseau bayésien dans le but de construire une approximation de la distribution dans D. Enfin, PrivBayes tire aléatoirement des enregistrements à partir de la distribution approximative dans le but de constituer un ensemble de données synthétique.

L'idée principale repose sur le fait que l'injection du bruit se produit au niveau des marginales à dimensions réduites de P à la place des données originales D. Les auteurs affirment que cela permet à PrivBayes de traiter des données de très grande dimension. L'enjeu majeur de cette approche réside dans la construction du réseau bayésien de manière différentiellement privée. [ZCP+14] ont recours au mécanisme exponentiel avec une fonction de score comme substitut de l'information mutuelle, pour apprendre de manière itérative les ensemble parents de chaque attribut dans le réseau bayésien.

2.3 Synthèse

le tableau 5.3 donne une comparaison des différentes approches de publication de données respectueuse de la confidentialité différentielle.

Approche	Dépendante des données	Dimension	Données mixtes	Paramètres
LPA [DMNS06]	✗	1D	✗	
NoiseFirst[XZX+13]	✗	1D	✗	
Privelet+ [XWG11]	✗	Multi-D	✗	
StructureFirst [XZX+13]	✓	1D	✗	Borne supérieure de compte
P-HPartition [ACC12]	✓	1D	✗	n^{bre} de bissections
DPCube [XXFG12]	✓	Multi-D	✗	
DP-MHMD [RKS16]	✓	Multi-D	✗	Groupement des attributs
DiffGen [MCFY11]	✓	Multi-D	✓	• Taxonomies des attributs catégoriels • n^{bre} de spécialisations
PrivBayes [ZCP+14]	✓	Multi-D	✓	Degré du réseau bayésien

Tableau 5.3 – Approches de publication de données différentiellement privée. La colonne *Dépendante des données* indique si l'approche dépend des données pour la perturbation. La colonne *Dimensions*, indique le nombre de dimensions dans les données de départ. La colonne *Données mixtes*, indique la capacité de l'approche à traiter des données avec des valeurs d'attributs continues. La colonne *Paramètres* liste les paramètres qui doivent être fournis pour chaque approche, en plus du budget de confidentialité ε.

La méthode que nous voulons mettre au point doit nous permettre de prendre en entrée des données multidimensionnelles, de traiter idéalement des données mixtes et ne doit contenir que peu ou pas de paramètres.

3 DPCocGen

Nous proposons une méthodologie pour anonymiser les microdonnées. L'objectif est de pouvoir publier une table de données anonymisée construite à partir des données d'origine tout en respectant les exigences de la confidentialité différentielle. La solution proposée combine le co-clustering avec la génération de données synthétiques pour produire des données anonymisées.

3.1 d-clustering

Dans les deux contributions précédentes, l'étape de préparation nous a permis, de transformer les données et de les rendre compatibles avec le co-clustering à deux dimensions (bi-clustering). Pour DPCocGen en revanche, nous exploitons la capacité du co-clustering MODL à produire un co-clustering à d-dimensions (d-clustering). L'idée ici est de regrouper des parties de chaque variable simultanément, de manière à construire une structure d'histogramme à d-dimension.

3.2 l'algorithme DPCocGen

Nous présentons notre algorithme DPCocGen, une stratégie de partitionnement en deux phases basée sur le co-clustering pour la génération de données synthétiques. Tout d'abord, un partitionnement sur les domaines, indépendant des données, est utilisé pour générer un histogramme multidimensionnel. Le mécanisme de Laplace est utilisé comme dans la stratégie de référence [DMNS06] pour perturber l'histogramme. Ensuite, un co-clustering MODL multidimensionnel est réalisé sur l'histogramme bruité. Cette première phase correspond à un co-clustering différentiellement privé, comme le montre la figure 5.3, et vise à produire un schéma de partitionnement. Dans la deuxième phase, DPCocGen utilise le schéma de partitionnement pour segmenter les données d'origine et calcule un compte bruité pour chacune des partitions en utilisant le mécanisme de Laplace. Enfin, des individus synthétiques sont tirés des partitions.

L'avantage de cette approche réside dans le fait que le schéma de partitionnement obtenu par le co-clustering est indirectement dépendant de la structure des données. L'intuition est que même après avoir perturbé l'histogramme multidimensionnel, la phase de co-clustering va préserver certaines des relations existantes entre les différentes dimensions. Les fusions de cellules qui en résultent limitent l'impact de l'addition de bruit dans la deuxième phase. Les données originales ne sont pas consultées lors de la construction du co-clustering, ce qui permet d'économiser le budget de protection qui est réparti entre les deux phases pour perturber les comptes. Les étapes détaillées de DPCocGen sont données dans l'Algorithme 2.

Algorithme 2 : algorithme DPCocGen

Require: Ensemble de données D, le budget global de confidentialité ε

1: **Phase 1** :
2: Construire un histogramme multidimensionnel à partir de D.
3: **Perturber** les comptes de chaque cellule en utilisant un budget de confidentialité ε_1.
4: Effectuer un co-clustering multidimensionnel à partir de l'histogramme obtenu à l'étape 3.
5: **Phase 2** :
6: Partitionner l'ensemble de données D en fonction du schéma de partitionnement obtenu à l'étape 4.
7: **Perturber** les comptes agrégés de chaque partition obtenus à l'étape 6 en utilisant un budget de confidentialité $\varepsilon_2 = \varepsilon - \varepsilon_1$.
8: Générer des individus synthétiques à partir de chaque partition en utilisant les comptes perturbés retournés de l'étape 7 pour construire un ensemble de données synthétiques D'.

Algorithme 3 : Algorithme Perturber

Require: Compte c, budget de confidentialité ε

1: $c' = c + Lap(1/\varepsilon)$
2: **if** $c' < 0$ **then**
3: $\quad c' = 0$
4: **end if**
5: Retourner c'

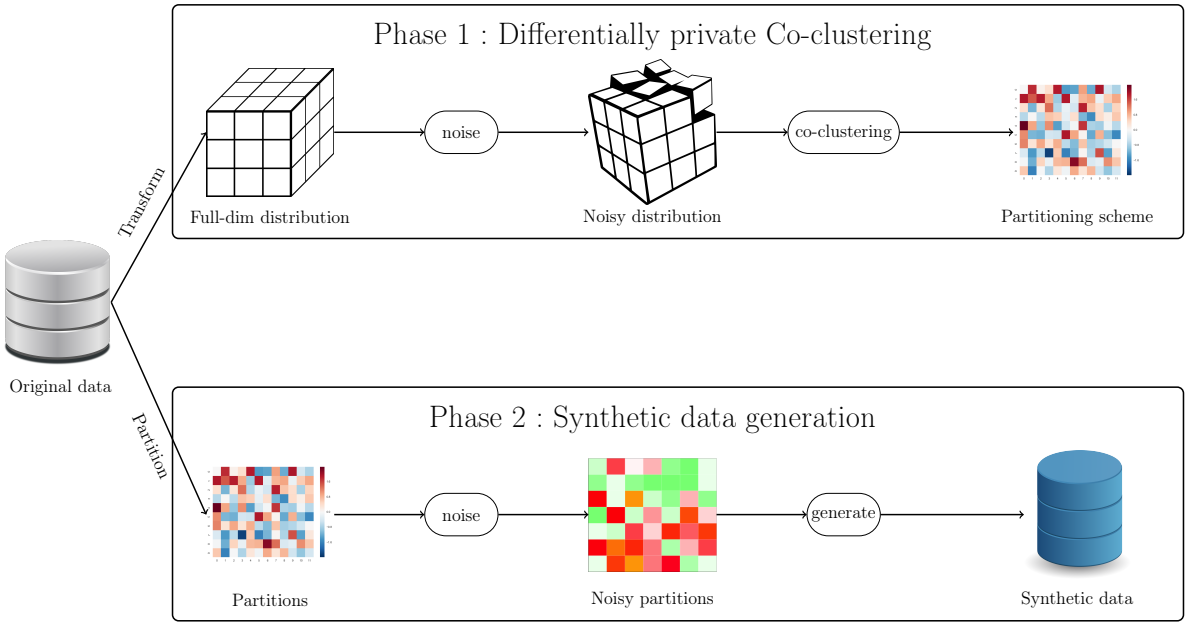


Figure 5.3 – DPCocGen : une stratégie de partitionnement en deux phases basée sur le co-clustering pour la génération de données synthétiques.

3.3 Garantie de confidentialité

DPCocGen suit la propriété de composition de la confidentialité différentielle. Les première et deuxième phases nécessitent un accès direct à la base de données. Les étapes 3 et 7 de l'algorithme 2 sont ε_1, ε_2-différentiellement privées. Aucun accès à la base de données d'origine n'est invoqué pendant la phase d'échantillonnage. La séquence est donc ε-différentiellement privée avec $\varepsilon = \varepsilon_1 + \varepsilon_2$.

4 Expérimentations

Dans cette section, nous menons trois expériences sur un ensemble de microdonnées réelles afin d'illustrer l'efficacité de notre proposition sur un cas pratique. L'objectif est d'explorer l'utilité des données synthétiques en mesurant la préservation des propriétés statistiques, l'erreur relative sur un ensemble de réponses à des requêtes aléatoires et leur performance prédictive.

4.1 Protocole expérimental

4.1.1 Jeu de données

Nous expérimentons avec la base de données Adult [1] qui contient 48842 observations décrites par 14 variables numériques et catégorielles, parmi lesquelles on retient les variables *Age*, *Workclass*, *Education*, *Relationship*, et *Sex*. Nous discrétisons les attributs continus en intervalles d'égale largeur, indépendamment de la distribution des données.

4.1.2 L'approche de référence

Nous mettons en œuvre la stratégie de base [DMNS06] pour générer un ensemble de données synthétiques. Un histogramme multidimensionnel prenant en compte l'ensemble des dimensions {*Age*, *Workclass*, *Education*, *Relationship*, *Sex*} est calculé et ensuite perturbé par un mécanisme différentiellement privé. Les enregistrements sont ensuite tirés à partir des comptes bruités pour former un ensemble de données.

4.1.3 PrivBayes

Nous utilisons une implémentation de PriveBayes [ZCP+14] disponible sur [jzh] afin de générer un jeu de données synthétique. Et Nous utilisons $\theta = 4$ comme suggéré par les auteurs.

4.1.4 Affectation du budget de protection de la vie privée

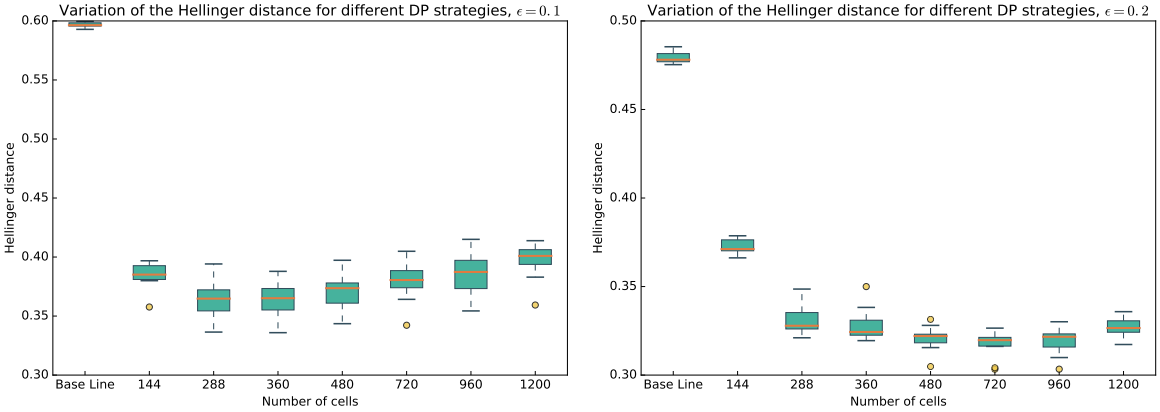
Le budget de protection de la vie privée est également réparti entre les deux phases de*DPCocGen* pour toutes les expériences, $\varepsilon_1 = \varepsilon_2 = \varepsilon/2$.

4.2 Performance descriptive

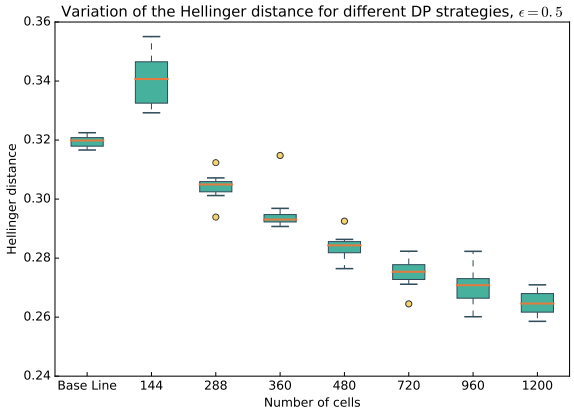
Dans cette expérience, nous nous intéressons à la préservation des propriétés statistiques des données originales dans les données produites. Tout d'abord, nous calculons le vecteur de distribution multivarié P de l'ensemble de données original, puis nous calculons le vecteur de distribution multivarié Q des données synthétiques générées avec

1. `https://archive.ics.uci.edu/ml/`

DPCocGen et le vecteur de distribution multivarié Q' des données synthétiques générées avec l'approche de référence, nommée *Baseline*. Enfin, les distances $D_{Hellinger}(P, Q)$ et $D_{Hellinger}(P, Q')$ sont mesurées. Pour chaque configuration, les distances sont calculées à partir de 50 ensembles de données synthétiques et représentées sur la Figure 5.4. Pour présenter ces résultats, nous utilisons des diagrammes en boîte où l'axe des x représente la méthode de génération de données synthétiques. La première boîte à gauche représente la stratégie de référence, les boîtes suivantes correspondent à *DPCocGen* avec différents niveaux de granularité (nombre de cellules du co-clustering). L'axe des y indique la distance de Hellinger mesurée entre la distribution calculée sur les données générées et la distribution originale.



(a) $\varepsilon = 0.1$ (b) $\varepsilon = 0.2$



(c) $\varepsilon = 0.5$

Figure 5.4 – Distance des distributions jointes

On évalue l'impact de l'agrégation du co-clustering ainsi que l'impact du budget ε. Rappelons que plus ε est petit, plus le bruit ajouté est important. Quel que soit le budget de confidentialité, la distribution de probabilité jointe des données synthétiques générées

avec *DPCocGen* est plus proche de la distribution originale que la distribution des données obtenues avec *Baseline*, sauf lorsque $\varepsilon = 0,5$ et pour le configuration *DPCocGen* avec un niveau élevé de regroupements (144 cellules). Dans ce cas de figure le partitionnement est trop grossier et ne permet pas de décrire correctement les données. Le niveau d'agrégation optimal varie en fonction de l'amplitude du bruit, mais le niveau d'agrégation le plus fin semble offrir un résultat satisfaisant pour chaque configuration testée. Nous utiliserons le niveau d'agrégation le plus fin pour les expérimentations suivantes.

4.3 Requêtes aléatoires de dénombrement

Le but de cette expérience est d'évaluer l'utilité des données produites en termes d'erreur relative lorsqu'il s'agit de répondre à des requêtes de dénombrement. Nous générons d'abord 100 requêtes aléatoires. Nous produisons des ensembles de données synthétiques en utilisant *Baseline*, *PrivBayes* et *DPCocGen*. Nous calculons toutes les requêtes et rapportons leur erreur moyenne sur 15 exécutions. Nous utilisons pour cette expérience le niveau de co-clustering le plus fin. La Figure 5.5 montre que l'erreur relative moyenne diminue lorsque le budget de confidentialité ε augmente pour les trois algorithmes. On peut aussi observer que *PrivBayes* et *DPCocGen* sont proches et font mieux que *Baseline*. Par exemple, pour le cas $\varepsilon = 0.01$ l'erreur relative moyenne mesurée avec *PrivBayes* et *DPCocGen* est inférieur à 30% et dépasse 40% pour *Baseline*. cette observation est valable indépendamment du budget de confidentialité considéré.

4.4 Performance prédictive

Dans cette expérience, nous nous intéressons aux performances de classification obtenues avec un classificateur supervisé dont l'apprentissage est basé sur des ensembles de données synthétiques. Nous sélectionnons aléatoirement 80% des observations afin de générer les données synthétiques en utilisant *DPCocGen*, *Baseline* et *PrivBayes*. Nous utilisons les données générées pour construire un classifieur afin de prédire la valeur de l'attribut *Sex*, et un autre pour prédire la valeur de l'attribut *Relationship*. Les 20% restants sont utilisés pour les évaluations. Nous utilisons pour cette expérience le niveau de

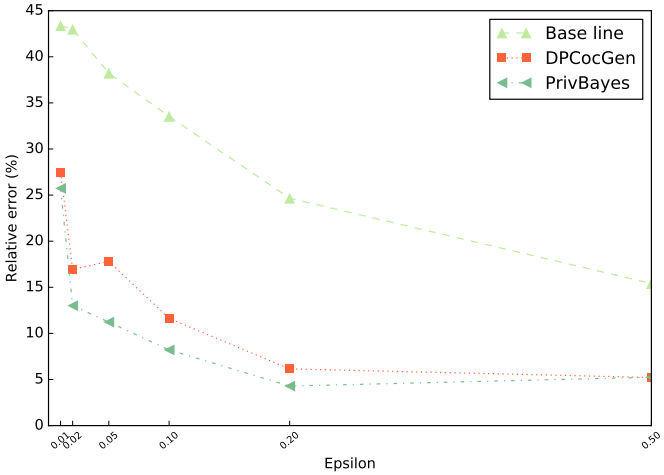
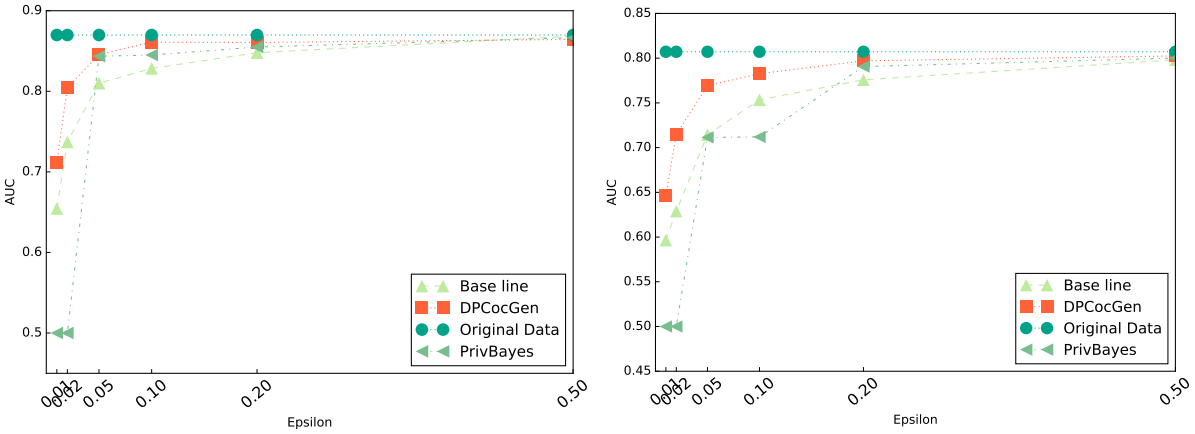


Figure 5.5 – Requêtes aléatoires de dénombrement

co-clustering le plus fin. Les résultats sont présentés Figures 5.6a et 5.6b, il s'agit d'une moyenne de 50 exécutions. La valeur du budget de confidentialité est indiquée sur l'axe des abscisses. L'axe des ordonnées indique la surface sous la courbe ROC (AUC) mesurée sur l'ensemble de test. La figure indique également, sous ma dénomination *Original Data*, les performances obtenues lorsque les données réelles sont utilisées pour l'apprentissage du modèle.



(a) Prediction de l'attribut Sex

(b) Prediction de l'attribut Relationship

Figure 5.6 – Performances prédictives

Nous retenons que les performances de classification obtenues avec *DPCocGen* sont proches de celles obtenues lorsque les données réelles sont utilisées pour apprendre le modèle. Les performances de *DPCocGen* sont toujours supérieures à celles de *Baseline* et *PrivBayes*.

5 Conclusion

Dans ce chapitre, nous avons présenté une approche pour l'anonymisation de micro-données. L'objectif était de pouvoir produire des données synthétiques qui préservent suffisamment d'information pour être utilisées à la place des données réelles. Notre approche consiste à combiner la confidentialité différentielle avec la génération de données synthétiques. Nous utilisons une technique de co-clustering, afin de partitionner l'espace de données de manière différentiellement privée. Ensuite, nous utilisons les partitions résultantes pour générer des individus synthétiques. Nous avons montré que les données synthétiques ainsi générées conservent les propriétés statistiques des données brutes. Nous avons également montré que notre approche non-paramétrique surpasse les algorithmes de référence pour la publication de données différentiellement privée.

Chapitre 6

Conclusion et Perspectives

Sommaire

1 Conclusion

Dans cette thèse, nous nous sommes intéressés au problème de la publication de données respectueuse de la vie privée et plus particulièrement à l'anonymisation de microdonnées contenant des enregistrements individuels détaillés. Les communautés des statistiques et de l'informatique ont proposé nombre de travaux dans les domaines du *Contrôle de la Divulgation Statistique* et celui de la *Publication de Données Respectueuse de la Vie Privée*. Tandis que la première communauté s'appuie sur des techniques de masquage et de perturbation pour construire un processus d'anonymisation pouvant être réitéré jusqu'à satisfaction d'un critère d'évaluation du risque de divulgation, la deuxième se focalise sur la proposition de formalismes de protection avec de solides garanties. Latanya Sweeney a été l'une des pionnières du domaine de la *Publication de Données Respectueuse de la Vie Privée*, en proposant le formalisme du k-anonymat, qui garantit qu'un individu ne peut être distingué d'aux moins $k-1$ autres individus. Dans ce contexte, les solutions

proposées ont suivi deux grandes idées : (1) fondre chaque individu dans un groupe d'individus, (2) ajouter un bruit perturbateur aux données ou à des représentations agrégées des données. L'enjeu de telles approches réside dans le compromis protection/utilité, et la difficulté consiste à maitriser un risque de divulgation tout en préservant au mieux l'information des données originales dans les données anonymisées. On s'est intéressé, dans le cadre de la thèse, à des techniques qui permettent de se prémunir des risques de divulgation tout en : (1) nécessitant un minimum de paramétrage utilisateur, (2) permettant la publication de données multidimensionnelles, (3) conservant une utilité raisonnable dans les données, de sorte que l'utilisateur ait possibilité d'effectuer des analyses de natures diverses.

En résumé, dans cette thèse, nous avons d'abord exploré uns stratégie d'anonymisation basée sur le co-clustering et la génération synthétique (Contribution 1). La méthode proposée permet d'obtenir des données anonymes dont l'utilité est très bonne mais à l'instar des méthodes concurrentes elle fournit pas de garantie formelle quant à la protection apportée. Pour pallier ce problème nous avons proposé une méthode (Contribution 2) qui vise à assurer le k-anonymat, et qui tire également partie du co-clustering. En dépit d'excellent résultats pratiques, les risques de divulgation considérés jusque là concernent uniquement la réidentification. Afin de protéger les données de menaces d'autres natures (divulgation d'appartenance à une table et attaques probabilistes), nous nous sommes intéressés au formalise de la confidentialité différentielle. Nous avons donc proposé une approche pour la génération de données synthétiques respectueuse de la confidentialité différentielle (Contribution 3).

Bilan de la contribution 1 Dans une première phase, un co-clustering des données partitionne conjointement les individus et les variables descriptives et permet de constituer des groupes d'individus. Obtenir le co-clustering optimal ne nécessite aucun paramétrage utilisateur. Le co-clustering est également un modèle générateur sur lequel on s'appuie pour construire des individus synthétiques du même format que les individus initiaux. Nous avons montré expérimentalement que les données synthétisées à l'aide de notre approche : (1) conservent les propriétés des données originales et qu'il est donc

possible d'envisager leur utilisation pour différentes tâches de fouille, (2) permettent de brouiller les liens entre les individus représentés dans l'ensemble de données original et ceux présents dans l'ensemble fictif. Une telle approche est utilisable quand le but est de fournir des données "utiles", mais nécessite une analyse a posteriori en terme de protection, et ce afin de s'assurer qu'aucun individu de l'ensemble original ne puisse être réidentifié à partir de l'ensemble synthétique.

Bilan de la contribution 2 D'abord nous avons décrit notre outil de génération automatique de hiérarchies de généralisation. Puis, nous avons proposé une nouvelle approche pour atteindre le k-anonymat. Notre solution est basée sur le co-clustering et est hybride dans le sens où elle peut être perturbatrice pour certaines classes d'équivalence et non-perturbatrice pour d'autres. Nous avons montré par des expérimentations sur le jeu de données Adult que notre approche offre une utilité satisfaisante et est compétitive face aux algorithmes de l'état de l'art tel que ARX et l'approche à base de microagrégation. Cependant, le principal avantage de notre approche réside dans son efficacité : elle permet de traiter un nombre important d'attributs quasi-identifiants en un temps raisonnable. Cette approche offre un compromis intéressant entre l'utilité offerte par la **Contribution 1** et la protection assurée par la **Contribution 3**. Elle est utilisable pour se prémunir efficacement contre le risque de réidentification.

Bilan de la contribution 3 Nous avons présenté DPCocGen un nouvel algorithme d'anonymisation respectueux de la confidentialité différentielle. Tout d'abord, un partitionnement sur les domaines de description des données est utilisé pour générer un histogramme multidimensionnel bruité. Un co-clustering multidimensionnel est ensuite effectué sur l'histogramme bruité résultant en un schéma de partitionnement. Enfin, le schéma obtenu est utilisé pour partitionner les données originales de manière différentiellement privée. Des individus synthétiques peuvent alors être tirés des partitions. Nous avons également montré expérimentalement que notre approche non-paramétrique surpasse d'autres algorithmes de la littérature tels que PrivBayes. Cette dernière contribution offre l'alternative d'anonymisation la plus stricte. Cependant, une telle garantie se paye

au prix de l'utilité des données produites. Les analyses effectuées sur les données produites avec cette approche seront par exemple beaucoup moins fidèles aux résultats sur les données originales, en comparaison à ce que l'on obtiendrait en utilisant les données obtenues via la **Contribution 1** par exemple.

2 Perspectives

Notre travail peut être poursuivi dans au moins trois directions :

— Enrichir la **Contribution 1** en : (1) implémentant l'automatisation de la discrétisation des variables numériques avec le choix optimal de découpage en intervalles. Nous pouvons nous appuyer pour cela sur les travaux de [BBRC18] ce qui rendrait la méthodologie d'anonymisation libre de toute paramétrage utilisateur, (2) autoriser la génération de valeurs numériques. Les attributs numériques sont actuellement traités comme des modalités catégorielles (intervalles)(3) l'adaptant à des données où il existe plusieurs instances par identifiant, et traiter, le cas où l'on trouve plusieurs observations par couple (instance, variable) comme dans les séries temporelles ou les trajectoire Spatio-temporelles par exemple.

— Étendre la **Contribution 2** afin de garantir un formalisme pour la protection des attributs sensibles tel que la t-proximité et la l-diversité. Une telle modification impliquerait d'accorder une attention particulière aux attributs sensibles pendant la phase de synthèse en s'appuyant par exemple sur la distribution globale plutôt que sur la distribution spécifique à chaque cluster d'individus.

— La **Contribution 3** pourrait être améliorée en rendant la phase de co-clustering différentiellement privée pendant la phase de l'optimisation du critère MODL. Au lieu de calculer un histogramme multidimensionnel bruité à l'aide du mécanisme Laplacien, on n'accéderait qu'une seule fois aux données et la phase de construction du d-clustering serait perturbée via un mécanisme exponentiel. Le défi majeur serait alors de qualifier la sensibilité d'un tel processus et de limiter le nombre d'itérations de l'algorithme, dans le but de répartir le budget de protection ε sur

l'ensemble des itérations.

3 Liste des publications de cette thèse

— Fessant Françoise, Benkhelif Tarek, & Clérot Fabrice. "Anonymiser des données multidimensionnelles à l'aide du coclustering." EGC. 2017. (Prix du meilleur article académique de l'édition 2017)

— Benkhelif Tarek, Fessant Françoise, Clérot Fabrice, & Raschia Guillaume. (2017, August). Co-clustering for Microdata Anonymization. In International Conference on Database and Expert Systems Applications (pp. 343-351). Springer, Cham.

— Benkhelif Tarek, Fessant Françoise, Clérot Fabrice, & Raschia Guillaume. (2017, September). Co-clustering for differentially private synthetic data generation. In International Workshop on Personal Analytics and Privacy (pp. 36-47). Springer, Cham.

Bibliographie

[ACC12] Gergely Acs, Claude Castelluccia, and Rui Chen. Differentially private histogram publishing through lossy compression. In *2012 IEEE 12th International Conference on Data Mining*, pages 1–10. IEEE, 2012.

[Agg05] Charu C. Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB '05, pages 901–909. VLDB Endowment, 2005.

[Ant16] László Antal. *Statistical Disclosure Control for Frequency Tables*. PhD thesis, The University of Manchester, 2016.

[ARPDMT16] Vanessa Ayala-Rivera, A Omar Portillo-Dominguez, Liam Murphy, and Christina Thorpe. Cocoa : A synthetic data generator for testing anonymization techniques. In *International Conference on Privacy in Statistical Databases*, pages 163–177. Springer, 2016.

[AS00] Rakesh Agrawal and Ramakrishnan Srikant. *Privacy-preserving data mining*, volume 29. ACM, 2000.

[AW89] Nabil R. Adam and John C. Worthmann. Security-control methods for statistical databases : A comparative study. *ACM Comput. Surv.*, 21(4) :515–556, December 1989.

[BA05] R. J. Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *21st International Conference on Data Engineering (ICDE'05)*, pages 217–228, April 2005.

[BBR17] Aichetou Bouchareb, Marc Boullé, and Fabrice Rossi. Co-clustering de données mixtes à base des modèles de mélange. In *Conférence Interna-*

tionale Francophone sur l'Extraction et gestion des connaissances (EGC 2017), pages 141–152, 2017.

[BBRC18] Aichetou Bouchareb, Marc Boullé, Fabrice Rossi, and Fabrice Clérot. Un modèle bayésien de co-clustering de données mixtes. In *EGC*, pages 275–280, 2018.

[BEC+12] Vincent D Blondel, Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. Data for development : the d4d challenge on mobile phone data. *arXiv preprint arXiv :1210.0137*, 2012.

[BGR12] Marc Boullé, Romain Guigourès, and Fabrice Rossi. Clustering hiérarchique non paramétrique de données fonctionnelles. In *Extraction et gestion des connaissances*, pages 101–112, 2012.

[BLR08] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, pages 609–618, New York, NY, USA, 2008. ACM.

[Bou07] Marc Boullé. *Recherche d'une représentation des données efficace pour la fouille des grandes bases de données*. PhD thesis, Télécom ParisTech, 2007.

[Bou08] Marc Boullé. Khiops : outil de préparation et modélisation des données pour la fouille des grandes bases de données. In *EGC*, pages 229–230, 2008.

[Bou11] Marc Boullé. Estimation de la densité d'arcs dans les graphes de grande taille : une alternative à la détection de clusters. In *EGC*, pages 353–364, 2011.

[Bou12] Marc Boullé. Sélection bayésienne de modèles avec prior dépendant des données. In *EGC*, pages 29–34, 2012.

[Bra02] Ruth Brand. Microdata protection through noise addition. In *Inference control in statistical databases*, pages 97–116. Springer, 2002.

[BSA13] Gary Benedetto, Martha Stinson, and John M Abowd. The creation and use of the sipp synthetic beta. 2013.

[CC00] Yizong Cheng and George M Church. Biclustering of expression data. In *Ismb*, volume 8, pages 93–103, 2000.

[CDM+05] Shuchi Chawla, Cynthia Dwork, Frank McSherry, Adam Smith, and Hoeteck Wee. Toward privacy in public databases. In *Proceedings of the Second International Conference on Theory of Cryptography*, TCC'05, pages 363–385, Berlin, Heidelberg, 2005. Springer-Verlag.

[CFM+13] Rui Chen, Benjamin CM Fung, Noman Mohammed, Bipin C Desai, and Ke Wang. Privacy-preserving trajectory data publishing by local suppression. *Information Sciences*, 231 :83–97, 2013.

[CR10] Gregory Caiola and Jerome P Reiter. Random forests for generating partially synthetic, categorical data. *Trans. Data Privacy*, 3(1) :27–42, 2010.

[Cul93] Mary J Culnan. " how did they get my name ?" : An exploratory investigation of consumer attitudes toward secondary information use. *MIS quarterly*, pages 341–363, 1993.

[Dal77] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15 :2–1, 1977.

[DFGN10] Josep Domingo-Ferrer and Úrsula González-Nicolás. Hybrid microdata using microaggregation. *Inf. Sci.*, 180(15) :2834–2844, August 2010.

[DFMBMSS06] Josep Domingo-Ferrer, Antoni Martínez-Ballesté, Josep Maria Mateo-Sanz, and Francesc Sebé. Efficient multivariate data-oriented microaggregation. *The VLDB Journal—The International Journal on Very Large Data Bases*, 15(4) :355–369, 2006.

[DFMS02] Josep Domingo-Ferrer and Josep Maria Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and data Engineering*, 14(1) :189–201, 2002.

[DFSSC16] Josep Domingo-Ferrer, David Sanchez, and Jordi Soria-Comas. *Database Anonymization :Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*. Morgan & Claypool, 2016.

[DFT05] Josep Domingo-Ferrer and Vicenç Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2) :195–212, 2005.

[DKM+06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves : Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer, 2006.

[DL89] George Duncan and Diane Lambert. The risk of disclosure for microdata. *Journal of Business & Economic Statistics*, 7(2) :207–217, 1989.

[DM15] Josep Domingo-Ferrer and Krishnamurty Muralidhar. New directions in anonymization : Permutation paradigm, verifiability by subjects and intruders, transparency to users. *CoRR*, abs/1501.04186, 2015.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

[DN93] D Defays and Ph Nanopoulos. Panels of enterprises and confidentiality : the small aggregates method. In *Proceedings of the 1992 symposium on design and analysis of longitudinal surveys*, pages 195–204, 1993.

[Dre10] Jörg Drechsler. Using support vector machines for generating synthetic datasets. In *International Conference on Privacy in Statistical Databases*, pages 148–161. Springer, 2010.

[Dre11] Jörg Drechsler. My understanding of the differences between the cs and the statistical approach to data confidentiality. In *4th IAB workshop on confidentiality and disclosure*, 2011.

[dVFLS12] Sabrina De Capitani di Vimercati, Sara Foresti, Giovanni Livraga, and Pierangela Samarati. Data privacy : Definitions and techniques. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 20(6) :793–818, 2012.

[Dwo06a] Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer Berlin Heidelberg, 2006.

[Dwo06b] Cynthia Dwork. *Differential Privacy*, pages 1–12. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[Dwo10] Cynthia Dwork. Differential privacy in new settings. In *Proceedings of the Twenty-first Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '10, pages 174–183, Philadelphia, PA, USA, 2010. Society for Industrial and Applied Mathematics.

[Fre17] Feten Ben Fredj. *Méthode et outil d'anonymisation des données sensibles*. PhD thesis, Conservatoire national des arts et metiers-CNAM, 2017.

[FS69] Ivan P Fellegi and Alan B Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328) :1183–1210, 1969.

[FWCY10] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing : A survey of recent developments. *ACM Comput. Surv.*, 42(4) :14 :1–14 :53, June 2010.

[Geh06] J. Gehrke. Models and methods for privacy-preserving data analysis and publishing. In *Data Engineering, 2006. ICDE '06. Proceedings of the 22nd International Conference on*, pages 105–105, April 2006.

[GN08] Gérard Govaert and Mohamed Nadif. Block clustering with bernoulli mixture models : Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6) :3233–3245, 2008.

[GN13] Gérard Govaert and Mohamed Nadif. *Co-Clustering : Models, Algorithms and Applications*. John Wiley & Sons, 2013.

[GRS12] Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6) :1673–1693, 2012.

[Gui13] Romain Guigourès. *Utilisation des modèles de co-clustering pour l'analyse exploratoire des données*. PhD thesis, Université Panthéon-Sorbonne-Paris I, 2013.

[Har75] John A Hartigan. Clustering algorithms. 1975.

[HDC05] Zhengli Huang, Wenliang Du, and Biao Chen. Deriving private information from randomized data. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 37–48. ACM, 2005.

[HDFF+12] Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul De Wolf. *Statistical disclosure control*. John Wiley & Sons, 2012.

[HM03] Stephen Lee Hansen and Sumitra Mukherjee. A polynomial algorithm for optimal univariate microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, 15(4) :1043–1044, 2003.

[HvdWR+03] Anco Hundepool, Aad van de Wetering, Ramya Ramaswamy, Luisa Franconi, Alessandra Capobianchi, Peter-Paul de Wolf, Josep Domingo, Vicenç Torra, Ruth Brand, and Sarah Giessing. μ-argus user manual version 4.0. *Statistics Netherlands, Voorburg*, 2003.

[(ID14] International Data Corporation (IDC). The digital universe of opportunities. In *https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf*, 2014.

[(ID17] International Data Corporation (IDC). European data market smart 2013/0063. In *https://www.key4biz.it/wp-content/uploads/2018/04/SMART20130063_Final-Report_030417_2.pdf*, 2017.

[Iye02] Vijay S Iyengar. Transforming data to satisfy privacy constraints. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 279–288. ACM, 2002.

[JJR02] Markus Jakobsson, Ari Juels, and Ronald L. Rivest. Making mix nets robust for electronic voting by randomized partial checking. In *Proceedings of the 11th USENIX Security Symposium*, pages 339–353, Berkeley, CA, USA, 2002. USENIX Association.

[JLR+06] Daniel Jeske, Pengyue Lin, Carlos Rendon, Rui Xiao, and Behrokh Samadi. Synthetic data generation capabilties for testing data mining tools. 2006.

[jzh] jzhang027. *PrivBayes*. disponible à `https://sourceforge.net/projects/privbayes`.

[KB08] Alexei Kounine and Michele Bezzi. Assessing disclosure risk in anonymized datasets. *Proceedings of FloCon*, 2008.

[KL51] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1) :79–86, 03 1951.

[KPE+12] Florian Kohlmayer, Fabian Prasser, Claudia Eckert, Alfons Kemper, and Klaus A Kuhn. Flash : efficient, stable and optimal k-anonymity. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 708–717. IEEE, 2012.

[KRR+11] Satkartar K Kinney, Jerome P Reiter, Arnold P Reznek, Javier Miranda, Ron S Jarmin, and John M Abowd. Towards unrestricted public use business microdata : The synthetic longitudinal business database. *International Statistical Review*, 79(3) :362–384, 2011.

[Kun09] Meglena Kuneva. Keynote speech, roundtable on online data collection, targeting, and profiling. brussels, march 31, 2009, 2009.

[LC11] Jaewoo Lee and Chris Clifton. How much is enough ? choosing ε for differential privacy. In *International Conference on Information Security*, pages 325–340. Springer, 2011.

[LDR05] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Incognito : Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, pages 49–60, New York, NY, USA, 2005. ACM.

[LDR06a] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 25–25. IEEE, 2006.

[LDR06b] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Workload-aware anonymization. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 277–286. ACM, 2006.

[LL07] Ninghui Li and Tiancheng Li. t-closeness : Privacy beyond k-anonymity and l-diversity. In *In Proc. of IEEE 23rd Int'l Conf. on Data Engineering (ICDE'07)*, 2007.

[LLV10] N. Li, T. Li, and S. Venkatasubramanian. Closeness : A new privacy measure for data publishing. *IEEE Transactions on Knowledge and Data Engineering*, 22(7) :943–956, July 2010.

[LM05] Michael Laszlo and Sumitra Mukherjee. Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, 17 :2005, 2005.

[LP08] Yehuda Lindell and Benny Pinkas. Secure multiparty computation for privacy-preserving data mining. *IACR Cryptology ePrint Archive*, 2008 :197, 2008.

[LQS] Ninghui Li, Wahbeh H Qardaji, and Dong Su. Provably private data anonymization : Or, k-anonymity meets differential privacy.

[LTX08] Jiexing Li, Yufei Tao, and Xiaokui Xiao. Preservation of proximity privacy in publishing numerical sensitive data. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 473–486, New York, NY, USA, 2008. ACM.

[LZHH13] Bronwyn Loong, Alan M Zaslavsky, Yulei He, and David P Harrington. Disclosure control using partially synthetic data for large-scale health surveys, with applications to cancors. *Statistics in medicine*, 32(24) :4139–4161, 2013.

[MCFY11] Noman Mohammed, Rui Chen, Benjamin Fung, and Philip S Yu. Differentially private data release for data mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 493–501. ACM, 2011.

[McS09] Frank D McSherry. Privacy integrated queries : an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30. ACM, 2009.

[MKGV07] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity : Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1), March 2007.

[MNC07] Maurizio Atzori Mehmet Nergiz and Christopher Clifton. Hiding the presence of individuals from shared databases. In *2007 ACM SIGMOD International Conference on Management of Data*, Beijing, China, 06 2007.

[MPS99] S Muthukrishnan, Viswanath Poosala, and Torsten Suel. On rectangular partitionings in two dimensions : Algorithms, complexity and applications. In *International Conference on Database Theory*, pages 236–256. Springer, 1999.

[MS08] Krishnamurty Muralidhar and Rathindra Sarathy. Generating sufficiency-based non-synthetic perturbed data. *Transactions on Data Privacy*, 1(1) :17–33, 2008.

[MSV12] Sergio Martínez, David Sánchez, and Aïda Valls. Towards k-anonymous non-numerical data via semantic resampling. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 519–528. Springer, 2012.

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 94–103. IEEE, 2007.

[MW04] Adam Meyerson and Ryan Williams. On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228. ACM, 2004.

[NCN09] Mehmet Ercan Nergiz, Christopher Clifton, and Ahmet Erhan Nergiz. Multirelational k-anonymity. *Knowledge and Data Engineering, IEEE Transactions on*, 21(8) :1104–1117, 2009.

[Nis08] Kobbi Nissim. Private data analysis via output perturbation. In *Privacy-Preserving Data Mining*, pages 383–414. Springer, 2008.

[NRS07a] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Thirty-ninth Annual ACM Symposium on Theory of Computing*, STOC '07, pages 75–84, New York, NY, USA, 2007. ACM.

[NRS07b] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-*

ninth annual ACM symposium on Theory of computing, pages 75–84. ACM, 2007.

[NS06] Arvind Narayanan and Vitaly Shmatikov. How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*, 2006.

[ODf01] Anna Oganian and Josep Domingo-ferrer. On the complexity of optimal microaggregation for statistical disclosure control. *Statistical Journal of the United Nations Economic Comission for Europe*, 18 :345–354, 2001.

[RD10] Jerome P Reiter and Jörg Drechsler. Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality. *Statistica Sinica*, pages 405–421, 2010.

[Rei82] Steven P Reiss. Data-swapping–a technique for disclosure control. *J. Statistical Planning and Inference*, 6(1) :73–85, 1982.

[Rei03] Jerome P Reiter. Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29(2) :181–188, 2003.

[Rei05] Jerome P Reiter. Releasing multiply imputed, synthetic public use microdata : An illustration and empirical study. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 168(1) :185–205, 2005.

[RKS16] Harichandan Roy, Murat Kantarcioglu, and Latanya Sweeney. Practical differentially private modeling of human movement data. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 170–178. Springer, 2016.

[RSH07] Vibhor Rastogi, Dan Suciu, and Sungho Hong. The boundary between privacy and utility in data publishing. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, VLDB '07, pages 531–542. VLDB Endowment, 2007.

[RTG00] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2) :99–121, 2000.

[Rub93] Donald B Rubin. Statistical disclosure limitation. *Journal of official Statistics*, 9(2) :461–468, 1993.

[Sam01a] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Trans. on Knowl. and Data Eng.*, 13(6) :1010–1027, November 2001.

[Sam01b] Pierangela Samarati. Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering*, 13(6) :1010–1027, 2001.

[SCDFSM14] Jordi Soria-Comas, Josep Domingo-Ferrer, David Sánchez, and Sergio Martínez. Enhancing data utility in differential privacy via microaggregation-based k-anonymity. *The VLDB Journal*, 23(5) :771–794, 2014.

[Sch84] Ferdinand David Schoeman. *Philosophical dimensions of privacy : An anthology*. Cambridge University Press, 1984.

[SDW06] Natalie Shlomo and Ton De Waal. Protection of micro-data subject to edit constraints against statistical disclosure. 2006.

[SDX11] H Jeff Smith, Tamara Dinev, and Heng Xu. Information privacy research : an interdisciplinary review. *MIS quarterly*, 35(4) :989–1016, 2011.

[SE02] Chris J Skinner and MJ Elliot. A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society : series B (statistical methodology)*, 64(4) :855–867, 2002.

[Ski09] Chris Skinner. Statistical disclosure control for survey data. February 2009.

[SM17] H Surendra and HS Mohan. A review of synthetic data generation methods for privacy preserving data publishing. *International Journal of Scientific and Technology Research*, 6 :95–101, 2017.

[Sol08] Daniel Solove. Understanding privacy. 2008.

[SRN+16] Joshua Snoke, Gillian Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. General and specific utility measures for synthetic data. *arXiv preprint arXiv :1604.06651*, 2016.

[SRN+18] Joshua Snoke, Gillian M Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 181(3) :663–688, 2018.

[SSDF08] Agusti Solanas, Francesc Sebé, and Josep Domingo-Ferrer. Micro-aggregation-based heuristics for p-sensitive k-anonymity : one step beyond. In *Proceedings of the 2008 international workshop on Privacy and anonymity in information society*, pages 61–69. ACM, 2008.

[Swe97] Latanya Sweeney. Guaranteeing anonymity when sharing medical data, the datafly system. In *Proceedings of the AMIA Annual Fall Symposium*, page 51. American Medical Informatics Association, 1997.

[Swe02a] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05) :571–588, 2002.

[Swe02b] Latanya Sweeney. K-anonymity : A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5) :557–570, October 2002.

[Tem08] Matthias Templ. Statistical disclosure control for microdata using the r-package sdcmicro. *Transactions on Data Privacy*, 1(2) :67–85, 2008.

[TFBJ04] Traian Marius Truta, Farshad Fotouhi, and Daniel Barth-Jones. Assessing global disclosure risk in masked microdata. In *Proceedings of the 2004 ACM workshop on Privacy in the electronic society*, pages 85–93. ACM, 2004.

[Tor04] Vicenç Torra. Microaggregation for categorical variables : a median based approach. In *Privacy in statistical databases*, pages 162–174. Springer, 2004.

[Tro] J.K. Trotter. *Public NYC Taxicab Database Lets You See How Celebrities Tip.* `http://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546`.

[TV06] T.M. Truta and B. Vinay. Privacy protection : p-sensitive k-anonymity property. In *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, pages 94–94, 2006.

[VBE18] Michael Veale, Reuben Binns, and Lilian Edwards. Algorithms that remember : Model inversion attacks and data protection law. *arXiv preprint arXiv :1807.04644*, 2018.

[War65] Stanley L. Warner. Randomized response : A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309) :63–69, 1965.

[WB90] Samuel D Warren and Louis D Brandeis. The right to privacy. *Harvard law review*, pages 193–220, 1890.

[Wes68] Alan F Westin. Privacy and freedom. *Washington and Lee Law Review*, 25(1) :166, 1968.

[WF06] Ke Wang and Benjamin C. M. Fung. Anonymizing sequential releases. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 414–423, New York, NY, USA, 2006. ACM.

[WFY05] Ke Wang, Benjamin CM Fung, and Philip S Yu. Template-based privacy preservation in classification problems. In *Data Mining, Fifth IEEE International Conference on*, pages 8–pp. IEEE, 2005.

[WFY07] Ke Wang, BenjaminC.M. Fung, and PhilipS. Yu. Handicapping attacker's confidence : an alternative to k-anonymization. *Knowledge and Information Systems*, 11(3) :345–368, 2007.

[Win04] William E Winkler. Masking and re-identification methods for public-use microdata : Overview and research problems. In *International Workshop on Privacy in Statistical Databases*, pages 231–246. Springer, 2004.

[WLFW06] Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu, and Ke Wang. (α, k)-anonymity : An enhanced k-anonymity model for privacy preserving data publishing. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 754–759, New York, NY, USA, 2006. ACM.

[WROK09] Mi-Ja Woo, Jerome P Reiter, Anna Oganian, and Alan F Karr. Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, 1(1) :7, 2009.

[XDSH11] Heng Xu, Tamara Dinev, Jeff Smith, and Paul Hart. Information privacy concerns : Linking individual perceptions with institutional privacy assurances. *Journal of the Association for Information Systems*, 12(12) :798, 2011.

[XT06a] Xiaokui Xiao and Yufei Tao. Anatomy : Simple and effective privacy preservation. In *Proceedings of the 32Nd International Conference on Very Large Data Bases*, VLDB '06, pages 139–150. VLDB Endowment, 2006.

[XT06b] Xiaokui Xiao and Yufei Tao. Personalized privacy preservation. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, SIGMOD '06, pages 229–240, New York, NY, USA, 2006. ACM.

[XWG09] Xiaokui Xiao, Guozhang Wang, and Johannes Gehrke. Interactive anonymization of sensitive data. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 1051–1054. ACM, 2009.

[XWG11] Xiaokui Xiao, Guozhang Wang, and Johannes Gehrke. Differential privacy via wavelet transforms. *IEEE Transactions on Knowledge and Data Engineering*, 23(8) :1200–1214, 2011.

[XXFG12] Yonghui Xiao, Li Xiong, Liyue Fan, and Slawomir Goryczka. Dpcube : differentially private histogram release through multidimensional partitioning. *arXiv preprint arXiv :1202.5358*, 2012.

[XZX+13] Jia Xu, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, Ge Yu, and Marianne Winslett. Differentially private histogram publication. *The VLDB Journal*, 22(6) :797–822, 2013.

[YZW05] Zhiqiang Yang, Sheng Zhong, and Rebecca N. Wright. Anonymity-preserving data collection. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 334–343, New York, NY, USA, 2005. ACM.

[ZCP+14] Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes : Private data release via bayesian networks. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1423–1434. ACM, 2014.

[ZKSY07] Qing Zhang, N. Koudas, D. Srivastava, and Ting Yu. Aggregate query answering on anonymized tables. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 116–125, April 2007.

[ZZ07] Nan Zhang and Wei Zhao. Privacy-preserving data mining systems. *Computer*, (4) :52–58, 2007.

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE NANTES

Titre : Publication de données individuelles respectueuse de la vie privée

Mots clés : protection de la vie privée, k-anonymat, confidentialité différentielle.

Résumé : Il y a une forte demande économique et citoyenne pour l'ouverture des données individuelles. Cependant, la publication de telles données représente un risque pour les individus qui y sont représentés. Cette thèse s'intéresse à la problématique de l'anonymisation de tables de données multidimensionnelles contenant des données individuelles dans un objectif de publication.

On se concentrera plus particulièrement sur deux familles d'approches pour l'anonymisation: la première vise à fondre chaque individu dans un groupe d'individus, la deuxième est basée sur l'ajout d'un bruit perturbateur aux données originales. Deux nouvelles approches sont développées dans le cadre de l'anonymisation par groupe, elles consistent à agréger les données à l'aide d'une technique de co-clustering puis à utiliser le modèle produit, pour générer des enregistrements synthétiques, dans le cas de la première solution.

La deuxième proposition quant à elle, cherche à atteindre le formalisme du k-anonymat. Enfin, nous présentons DPCocGen un nouvel algorithme d'anonymisation respectueux de la confidentialité différentielle. Tout d'abord, un partitionnement sur les domaines est utilisé pour générer un histogramme multidimensionnel bruité, un co-clustering multidimensionnel est ensuite effectué sur l'histogramme bruité résultant en un schéma de partitionnement. Enfin, le schéma obtenu est utilisé pour partitionner les données originales de manière différentiellement privée. Des individus synthétiques peuvent alors être tirés des partitions.

Title : Privacy preserving microdata publishing

Keywords : privacy preserving data publishing, k-anonymity, differential privacy.

Abstract : There is a strong economic and civic demand for the opening of individual data. However, the publication of such data poses a risk to the individuals represented in it. This thesis focuses on the problem of anonymizing multidimensional data tables containing individual data for publishing purposes.

In particular, two data anonymization approaches families will be focused on: the first aims to merge each individual into a group of individuals, the second is based on the addition of disruptive noise to the original data. Two new approaches are developed in the context of group anonymization. They aggregate the data using a co-clustering technique and then use the produced model, to generate synthetic records, in the case of the first solution.

While the second proposal seeks to achieve the formalism of k-anonymity. Finally, we present a new anonymization algorithm "DPCocGen" that ensures differential privacy. First, a data-independent partitioning on the domains is used to generate a perturbed multidimensional histogram, a multidimensional co-clustering is then performed on the noisy histogram resulting in a partitioning scheme. Finally, the resulting schema is used to partition the original data in a differentially private way. Synthetic individuals can then be drawn from the partitions.