ETAT DE L'ART

Sécurisation des données sensibles dans le monde actuariel

PERIER Hugo

Sommaire

Introduction

Ce document a pour objectif de définir et décrire un certain nombre de concepts en utilisant des références scientifiques, théoriques et pratiques.
J'essayerai de le faire de manière assez générale mais prendrais mes exemples pour la plupart via les conclusions issues des jeux de données que je manipulerai dans mes projets au sein du groupe VYV.

Note : L'état de l'art sur la place de Big Data et le RGPD donc sous-partie 1 et 2 ne sont pas du tout finis. Les autres sont bien plus complètes.

I. Etat de l'art

Dans un premier temps sera exposé le nouveau paradigme instauré par le Big Data, comment il pousse les entreprises à se restructurer pour stocker et travailler des données aussi imposantes.

Je m'attarderai ensuite à expliquer au mieux le cadre légal instauré par le RGPD, les acteurs et son impact sur les entreprises qui gèrent des données, notamment des données médicales comme c'est le cas pour les assureurs.

Puis je me focaliserai sur la nature des données collectées par les assureurs, sur leur structure et leurs liens entre elles.

Je ferai par la suite un état de l'art des différentes menaces sur la vie privée, des techniques possibles pour protéger les données collectées (les MPVP) et les algorithmes qui permettent de les mettre en application.

Enfin, cette revue littéraire se clôturera par un état de l'art des différentes manières de quantifier le risque de divulgation d'une table de données. Cet aspect est nécessaire car chaque entreprise se doit de justifier ses pratiques internes, et, elle a besoin d'avoir les bons outils.

1. Place du Big data et son impact sur les méthodes de travail sur la donnée

Le terme Big data bien que défini en introduction, cependant le lien avec le monde actuariel et avec les méthodes de travail de la donnée a été peu abordé. La genèse de ce Big data a entrainé l'apparition de données « non-structurées », il s'agit de texte, d'images de vidéos ou encore des dates ou des évènements. Ce sont en fait des données sans format destinées à être analysées par des êtres humains, ainsi elles sont difficilement analysables à elles seules mais encore plus lorsque l'on parle d'amas de données (en comparaison a des données stockées dans des

tableurs). En effet, dès 2007, l'association de promotion de l'information libre APRIL ***Erreur ! S ource du renvoi introuvable.*** évoquait le nouveau paradigme lié aux informations non structurées. C'est toute une science autour de la construction, l'analyse et l'utilisation des données qui apparait. On y retrouve alors une dimension sociale, technique évidement mais aussi communautaire. Cette dernière tend vers une démocratisation d'internet, avec une considération de l'utilisateur et de son apport en tant que membre d'une communauté.

On peut alors faire le lien avec les différentes formes d'uberisation des modèles économiques, plaçant les utilisateurs au centre, collectant des données dans le but de les caractériser, les cibler, leur permettant de s'évaluer les uns et autres. Ainsi, dans le monde de l'assurance on voit apparaître de nouveaux services. En effet, de plus en plus d'acteurs se bataillent pour récupérer les précieuses données des consommateurs et le domaine de la santé n'est pas exclu. Il existait déjà des sites permettant de prendre des rendez-vous médicaux comme allodocteur.fr ou doctolib.fr cependant, on voit apparaitre depuis peu des plateformes de consultations médicale en ligne. Ainsi mesdocteurs.com et medecindirect.fr poussent encore plus loin les limites séparant le domaine médical et le web, créant un nouveau flux de données non structuré (ici vidéos issues des entretiens, ordonnances partagées etc.).

D'après l'académie des technologies *[AcTe2016]*, le Big data impacte les métiers de la gestion des données sur trois niveaux :

- Le volume, il pousse à se poser des questions sur le stockage des données (volume en pétaoctet 10^{15}).
- La variété des données, sur la forme de stockage des données non-structurées tout en conservant une facilité de navigation, utilisation.
- La vélocité : cela concerne un nombre important de données, volatiles et utiles que si exploitées rapidement.

En ce qui concerne la gestion des volumes de donnée, le stockage massivement distribué apparait comme une solution convenable à condition d'éviter tout déplacement de données et à distribuer également leur traitement. C'est là que l'on retrouve la technologie Hadoop, un Framework libre développé en java largement inspiré de publications de Google et détaillée dans le livre de Chokogoue et Juvénal **Erreur ! Source du renvoi introuvable.**. Hadoop vise a insi à séparer les données et à paralléliser le traitement de ces données sur plusieurs nœuds d'une grappe de calcul (un cluster d'ordinateurs). DataCore est un pionner dans les réseaux de

stockage SAN (Storage Area Network) c'est-à-dire des réseaux mono-tâche ayant pour but de mutualiser des ressources de stockage. Chaque année il publie une étude sur le marché du stockage. Celle de 2018 *[DataCore2018]* révèle ainsi que les sociétés (celles qui ont participées à l'études) restent préoccupées par :

- La continuité des activités, avec un besoin de disponibilité important. Les sociétés se tournent de plus en plus vers un cloud hybride. Un cloud hybride c'est une combinaison d'un Cloud public (serveurs partagés entre différents clients d'un même fournisseur, forcément en hors-site chez le fournisseur) et d'un Cloud privé (serveurs dédiés à une seule entreprise sur le site de l'entreprise ou alors hors-site).
- L'incompatibilité dans le stockage (il en existe plusieurs types : sds, hyperconvergé, san, nas…) Certains types sont plus performants et causent moins de migration comme le sds. La tendance est donc à l'adoption du sds.

2. RGPD et données personnelles ce que dit la loi

Dans cette partie, afin d'expliquer cette nouvelle loi et son étendue je mettrai en relation un certain nombre de documents et textes, un livre de Guillaume Desgens-Pasanau sur l'application française du Règlement Général sur la Protection des Données (RGPD) *[DePa2018]* ou encore les documents fournis par la CNIL *[CNIL2018]* pour citer les principaux.

a. Prémisses de la loi

Face à la croissance des technologies de l'information, une directive européenne de 1995 est venue modifier la loi française « informatique et liberté » de 1978. Cette dernière visait « la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données », d'après Guillaume Desgens-Pasanau *[DePa2018]* elle s'est centrée sur 3 axes :

- Marchandisation des flux
- Internationalisation des flux
- Traçabilité

C'est ce dernier point qui se rapproche plus de la sécurisation des données et que je vais développer. La législation française serait alors passée d'une problématique liée à un fichier

dans les années 90 à une liée à la trace, c'est-à-dire l'ensemble des données collectées à chaque connexion d'utilisateurs. Il s'est alors opéré le 24 octobre 1995 avec une directive européenne une réelle prise de conscience de la dangerosité des traitements de données personnelles par les organismes privés. En effet, la loi « informatique et liberté » du 6 janvier 1978 avait instauré un certain laxisme vis-à-vis du secteur privé qui n'avait pas pour obligation de valider ses méthodes de traitement auprès du CNIL (Comité National de l'Informatique et des Libertés). Cette inégalité a ainsi été abolie avec l'application de la directive européenne de 1995 à la législation française en 2004, ce fut la première réforme de la loi informatique et liberté. Les professionnels ont vu les formalités administratives qu'ils devaient effectuer auprès de la CNIL se réduire : la CNIL est devenue plus efficace et a eu plus de temps pour combattre les fraudes.

Cependant cette dernière réforme a instauré un régime d'autorisation préalable peu efficace. Toute entreprise, à caractère publique ou privée se devait d'obtenir l'accord de la CNIL avant de mettre en pratique des méthodes de traitement de données (comme la gestion des données client, la détection de fraude etc.) L'effet escompté ne fut pas le bon et les entreprises peu importe leur taille, dans le but de rester compétitive virent leur responsable de traitement face à un dilemme : se conformer à la CNIL ou rester dans l'illégalité et économiser les mois/années de traitement de dossier par cette dernière. Malgré certaines dispenses ou formulaires simplifiés dans plusieurs secteurs qui permettaient d'accélérer le traitement des dossiers, il se trouve que les conséquences en cas de fraude n'étaient pas particulièrement dissuasives, et c'est donc souvent cette dernière possibilité qui était choisie.

Et c'est ainsi que le Règlement Général sur la Protection des Données (RGPD) a été adopté par le Parlement européen le 14 avril 2016 et est entré en vigueur le 25 mai 2018 en France et la loi « informatique et liberté » fut remplacée par la « Loi sur la protection des données personnelles » à cette date.

b. <u>L'impact du RGPD sur les entreprises : les sanctions</u>

Grâce au RGPD on passe d'un système déclaratif à un système de contrôle. Le RGPD change la donne avec son article 83 stipulant que des « sanctions effectives, proportionnées et dissuasives seront délivrées pour toute violation du RGPD » *[DonRGPD2017]*. Nous parlons désormais d'amendes pouvant aller jusqu'à 20 millions d'euros, et pour les entreprises jusqu'à 4% de leur chiffre d'affaire mondial hors taxes sur le total de l'exercice précédent. En interne, la responsabilité du responsable de traitement ou du sous-traitant des données peut être mise en cause.

c. L'impact du RGPD sur les entreprises : changement en interne

Toute entreprise doit désormais respecter un grand nombre de principes et surtout être en mesure de démontrer à tout moment qu'elle les respecte. Les acteurs qui gèrent les données se doivent d'expliquer au délégué de la protection des données (DPD) et à la CNIL (en cas de contrôle) qu'ils respectent toutes les clauses liées à leur protection. Actuellement, selon Pierre-Louis THOUVENOT[1] la CNIL n'a que peu exercée son devoir de contrôle, certainement pour permettre aux entreprises de s'adapter, mais cela n'est qu'une question de temps. Bien entendu, tout salarié ou organisation syndicale peut engager des actions individuelles ou collectives en invoquant le non-respect du RGPD.

Les obligations de la RGPD sont précises *[Diri2019]* : réalisation d'une cartographie et l'établissement d'un registre des traitements de données, l'information transparente et complète des personnes concernées (salariés, clients, fournisseurs…), la conclusion de contrats avec les sous-traitants définissant leurs obligations en matière de gestion et de protection des données personnelles, ou encore la désignation d'un Délégué à la protection des données (également appelé « DPO »).

Ces obligations poussent les entreprises à s'adapter : information/formation du personnel, modification des processus, des documents qui circulent en interne, des jeux de données, de l'accessibilité de l'information. La création de charte pour tout employé en contact avec des données clients ou l'introduction de clauses contractuelles à faire signer aux clients sont aussi à mettre en place.

Mais désormais les sous-traitant gérant des données peuvent être mis en cause avec cette loi. Les entreprises se doivent donc de bien sélectionner leurs partenaires en adaptant leur méthode de sélection aux types de données concernées. Les compétences en sécurité, anonymisation ou cryptage seront à l'avenir des critères encore plus sélectifs au sein de toutes les entreprises qui veulent avoir une activité en Europe.

D'après *[Diri2019]* la CNIL a récemment indiqué qu'elle entendait « accompagner les entreprises dans la mise en œuvre des nouvelles obligations ou des nouveaux droits résultant

[1] Pierre-Louis THOUVENOT Lead data Scientist chez VYV depuis septembre 2018

du RGPD » notamment en mettant à disposition des « outils de préparation et de mise en conformité au RGPD ».

Le site de la CNIL *[CNIL2018]* confirme cette volonté, on y retrouve une section de question-réponses sur le RGPD, un calendrier d'ateliers d'information à destination des délégués pour la protection des données (DPD) ainsi que plusieurs onglets sur les spécificités par secteurs d'activité ou les bonnes pratiques.

Grâce au RGPD les entreprises passent d'une obligation de résultat à une obligation de moyens renforcés *[DePa2018]*. On peut alors se demander comment est évaluée cette mise en place de moyens, comment les entreprises peuvent-elles montrer leur bonne foi ?

d. Méthodes d'évaluation de la sécurité des données vis-à-vis du RGPD

En regard de son site internet, la CNIL ne partage aucune information quantifiable liée au risque de partage de données, aucun barème permettant de mesurer si des données sont protégées ou non. Par exemple, elle évoque une « gestion des risques » en fonction de l'activité et des « mesures de sécurité, informatique mais aussi physique » adaptées à la sensibilité des données et du risque de divulgation.

Exemple concret, le logiciel PIA est un logiciel mis à disposition par la CNIL visant à compléter l'analyse d'impact relative à la protection des données (AIPD) prévue par le RGPD. En annexe

*[*Divers4] se trouve une capture d'écran de ce logiciel, et plus particulièrement de la zone de définition des risques. On y retrouve un simple barème et aucune normalisation de celui-ci.

C'est donc cette « approche par le risque » qui incombe donc au responsable de traitement de trouver des indicateurs suffisant pour:

- Le respect des droits des personnes liées aux données (issu de la loi « informatique et liberté »)

- Réaliser dans certain cas des études d'impact (« privacy impact assesment ») : un prérequis avant la mise en place de tout nouveau traitement de données.

3. Nature des données collectées et leurs enjeux

Je focaliserai mon mémoire sur la gestion des micro données. Ceci dit, avant de définir le concept d'anonymisation, il serait préférable de s'attarder à définir la place des données collectées et les enjeux actuels.

a. Contexte

Les données sont une ressource indispensable de production de bien et de services, moteurs de la prise de décisions en interne. Actuellement la société moderne se voit confrontée à un mouvement ou plutôt une philosophie d'accès à l'information nommée « open data ». Ce dernier pousse à une publication de données librement accessible, utilisable par tout un chacun. Plusieurs évènements sont venus renforcer ce mouvement :

- La signature d'une « Charte du G8 pour l'ouverture des données publiques » en 2013 a imposé aux collectivité la publication numérique de données publiques.

- En 2016 le Conseil Compétitivité de l'Union européenne a annoncé que tous résultats issus de recherches scientifiques financées au moins en partie par l'UE seraient publiés librement et accessible par tout un chacun dès 2020.

Ce pas en avant de l'union européenne légifère ainsi une tendance observable depuis 2010, avec une augmentation croissante des publications gouvernementales de data partagées librement par le sur le « Web of science », un service d'information universitaire. Se référer à l'annexe *[Graph1]*.

Dans un contexte économique et professionnel, l'anonymisation de données apparait comme cruciale pour :

- Maintenir une relation de confiance avec les clients (et donc continuer d'avoir leur consentement pour exploiter leur données)

- Dégager de la valeur liée à cette ressource. Car en effet, d'après la Commission Européenne les données personnelles pourraient croitre en valeur de 1 trillion d'euros par an en 2020 (concernant les données des citoyens européens).

b. Exemple concret

Les données sont au cœur de la promesse mutualiste, d'un côté on retrouve la personnalisation des services, d'un autre le respect de la vie privée. Selon Stéphane BRETON[2], l'utilisation des données sera au cœur de la relation client-assureur de demain, le groupe VYV a ainsi lancé un chantier d'identification et de recensement des sources de données durant l'été 2018.

Lorsque l'on parle de données, la première question est liée à leur recensement. Ce dernier constitue une préoccupation croissante des entreprises, peu importe leur domaine d'activité. Dans le milieu mutualiste à titre d'exemple, ces données font office de levier à la fois stratégique qu'opérationnels. Stratégique d'une part pour :

- La mise en place de nouveaux services
- La personnalisation des services
- Une anticipation de nouveaux besoins
- Une optimisation des couts (de création, vente, promotion)

Et Opérationnels d'autre part pour :

- Une compréhension et actions à entreprendre pour valoriser au mieux ce patrimoine
- Une priorisation des efforts en matière de qualité de donnée
- Un maintien de la cohérence des données dans le temps
- Une priorisation des croisements de données qu'il convient de renforcer
- Une priorisation des données à acquérir

Au sein de ce groupe de mutualistes VYV, Stéphane BRETON a pu me m'informer sur la nature des données collectées en interne. Ce ne sont pas des données confidentielles, mais elles n'ont pas pour habitude d'être partagées à l'extérieur du groupe. Ainsi, au sein de leurs livrets de données, ils se focalisent sur les données de type référentiels, de type Front Office (gestion de la relation, vente marketing) et sur les données de type Back Office (le cœur du métier) :

[2] Responsable de Projets Patrimoine et Acquisition de Données au sein du groupe assureur mutualiste VYV

entrée en assurance, vie du contrat, cotisation, gestion des sinistres et des prestations. Le [Divers1] en Annexe répertorie les données macroscopiques du groupe VYV. L'analyse de ce graphique en cercle permet d'en savoir plus sur la nature et la hiérarchie des données utilisées et diffusée en interne. Pour se faire, une clé de lecture est nécessaire au préalable.

Tout d'abord, la taille des bulles reflète la présence d'une même typologie de données dans une ou plusieurs applications des entités. Par exemple : les données personnelles (Personne physique – adhérent) sont nécessairement présentes dans toutes les entités mais souvent à plusieurs reprises dans une même entité sa taille est alors importante. Ainsi, La taille des bulles n'indique pas la volumétrie des données, indicateur qu'il pourrait être intéressant de collecter par ailleurs. La présence d'une famille de donnée dans une entité est indiquée au niveau de la légende, mais il faut tout de même se rappeler que ce graphique est une ébauche : le recensement des données n'est pas encore achevé à ce stade. Et lorsqu'il y a une superposition de données, cela est le signe d'une « connexion » théorique entre ces dernières. On peut ainsi constater l'étendue des données recueillies par les entreprises du groupe VYV ayant plus ou moins un lien avec le client ou l'entreprise et les méthodes de tarification de contrat. Quelles données sont susceptibles d'être anonymisées ?

Il serait alors intéressant de s'intéresser à la classification de ses données. En effet, lorsque l'on se questionne sur la divulgation de données privées ou personnelles, il faut tout d'abord distinguer ces deux termes. Une donnée à caractère personnel signifie qu'il s'agit d'une information relative à une personne physique identifiée *[Rh2015]*. On distingue alors deux catégories de données personnelles : les directes (patronyme, photographie, date de naissance) des indirectes. Ces dernières permettent si elles sont analysées de caractériser une personne : numéro de sécu, numéro de passeport… Au niveau de la législation française, on notera une spécification des données « sensibles », privés et personnelles, qui peuvent rassembler des données liées au domaine médical (maladie, traitement, hospitalisation…), vie sexuelle, religion, opinions politiques etc.

Faisons maintenant le lien avec les données collectées par les assurances du groupe VYV. Les différents types de données sur les ventes et CRM (Customer Relationship Management), Cotisations et Liquidation et Référentiel sont détaillés dans l'annexe au niveau du *[Divers2]* on y retrouve le même type de bulles issu du *[Divers1]*. Ainsi, si l'on se focalise les données que l'on pourrait considérer comme personnelles directes : Nom, Prénom, Sexe, Civilité, Date de décès, Date et pays de naissance, Profession, Lieux d'exercice, Employeur, Type de situation

maritale et date, Adresse postale, Adresse mail (si elle contient un patronyme). Celles indirectes : Identifiant Personne, Id gestionnaire, Rôle Personne, Tiers aidant, Coordonnées GPS. Au niveau des données sensibles, on retrouvera les données liées à la maternité (Nature du dossier, Date de création et de clôture, date présumée de grossesse, date d'accouchement, date de fin de congé, nombre d'enfants ; celles liées aux affections longues durées (nature de l'affection, date de création, date de clôture ainsi que d'autres informations liées aux éventuelles complication, traitements et opérations) ; données liées au médecin traitant (numéro professionnel, date de début et de fin) et les données liées aux remboursements / sinistres (date, montant …).

4. L'anonymisation de données
    a. Contexte et enjeux de l'anonymisation de données

Les données sont une ressource indispensable de production de biens et de services, moteurs de la prise de décisions en interne. Cependant, la société moderne se voit confrontée à un mouvement ou plutôt une philosophie d'accès à l'information nommée « open data ». Cette dernière pousse à une publication de données librement accessibles, utilisables par tout un chacun. Ainsi plusieurs évènements sont venus renforcer ce mouvement :

- La signature d'une « Charte du G8 pour l'ouverture des données publiques » en 2013 a imposé aux collectivités la publication numérique de données publiques.
- La mise en place d'une publication libre de tous les articles et revues scientifiques dès 2020.

On se questionne donc légitimement sur la divulgation de données privées ou personnelles. Ainsi, dans un contexte économique et professionnel, l'anonymisation des données apparait comme cruciale pour :

- Maintenir une relation de confiance avec les clients (et donc continuer d'avoir leur consentement pour exploiter leurs données)
- Dégager de la valeur liée à cette ressource. Car en effet, d'après la Commission Européenne, les données personnelles pourraient croitre en valeur de 1 trillion d'euros par an en 2020 (concernant les données des citoyens européens).

D'après la norme ISO/TS 25237 de 2008 : l'anonymisation est *« un processus qui supprime l'association entre l'ensemble de données identifiant et le sujet des données ».* Ce qui suppose qu'à la suite de l'utilisation d'un processus d'anonymisation, les données sont sous une forme qui empêche quiconque d'identifier les caractéristiques d'un individu de la table de données de manière directe ou par croisement de tables.

D'où le dilemme pour les entreprises : si anonymiser c'est empêcher de déduire des informations personnelles en croisant des jeux de données, alors c'est ce croisement de données qui permet aux entreprises de créer de la valeur. Le but est alors de préserver une certaine qualité des données. Donc apparition de deux challenges :

- Réduire le risque de divulgation de données confidentielles et sensibles
- Réduire la perte d'information utile : donc en fonction des utilisations, une collaboration avec les professionnels du secteur d'activité est nécessaire dans le choix du processus d'anonymisation.

b. Attaques sur la vie privée

Une attaque sur la vie privée est portée par un attaquant sur une table de donnée accessible par ce dernier. Cet attaquant cherche à récupérer des informations sur un individu appelé personne cible, il peut aussi bien s'agir d'un groupe de personnes cibles. Le risque de ré-identification, c'est-à-dire le risque de retrouver des caractéristiques d'attributs personnels regroupe :

- Le risque d'individualisation, c'est-à-dire arriver à isoler un individu
- Le risque de corrélation, soit arriver à relier des data base concernant un même individu
- Le risque d'inférence : arriver à déduire des infos sur un individu en partant d'informations sur une population

Nous verrons par la suite des méthodes calculatoires du risque de ré-identification. Dans la partie qui suit la majorité des informations proviennent de la thèse de Feten BEN FREDJ *[BeFre2017]* ainsi que du document de travail de Maxime BERGEAT sur la gestion de la confidentialité sur les données individuelles *[Be2016]*.

La taxonomie des modèles d'attaque de la vie privée (AVP) en annexe *[Divers3]* permet d'avoir une vue d'ensemble sur tous les différents types d'attaques sur les bases de données. On distingue ainsi deux grands types d'AVP : celles par liens et celles par inférence

probabilistes. Afin de bien comprendre les différences entre ces attaques et comment elles sont orchestrées par les attaquants, j'ai construit un tableau de comparaison pour chaque type.

Synthèse des différentes Attaques possibles sur la vie privée :

<table>
<tr><th colspan="5">Attaque par liens : l'attaquant connait des QI de sa victime</th></tr>
<tr><th>AVP par lien de tables</th><th>AVP par lien d'enregistrement</th><th colspan="3">AVP par lien d'attribut :
L'attaquant connait le QI de sa cible, et peut croiser simplement deux tables (la sienne et une publiée) pour connaitre des attributs sensibles de sa victime.</th></tr>
<tr><td>A priori l'attaquant ignore la présence de sa victime dans la table A anonymisée. Cependant, il sait que cette dernière appartient à un groupe de k' individus d'une table B. Il suppose que la victime appartient à un groupe similaire en caractéristiques dans la table A qui contient k individus. Ainsi la victime à une probabilité k/k' d'être dans A avec les caractéristiques du groupe.</td><td>L'attaquant en plus de connaitre les QI de sa victime, sait que cette dernière fait partie de la table.</td><td>Par homogénéité
Si tous les individus d'un groupe d'une table A ont la même caractéristique (exemple : toutes les femmes de 52 ans du 20ème arrondissement ont un ulcère gastrique -> déduction directe de la victime qui se trouve appartenir à ce groupe du fait de ses QI.</td><td>Fondée sur les connaissances de base
L'attaquant connait en plus des QI de sa victime une information de base du style « 80% des personnes en activité ayant ces QI sont tendues » ; il arrive à connaitre le groupe de sa cible avec les QI puis à séparer les données sensibles grâce à son information de base.</td><td>Par similarité sémantique
Ici comme avec l'homogénéité l'attaquant suppose que sa victime a les caractéristiques des individus du groupe qu'il a identifié dans une table avec ses QI. Mais tous n'ont pas la même caractéristique, donc on fait des regroupements. Exemple : s'ils ont tous des maladies du cœur, alors la victime a certainement un problème cardiaque.</td></tr>
</table>

Attaque par inférence probabiliste
L'attaquant ne croise pas de tables. Il base son analyse sur des probabilités générales avant de comparer ces dernières aux différentes distributions au sein de la table publiée. Exemple de l'attaque par dissymétrie (usuelle) : l'attaquant se focalise sur un groupe d'individus (par QI type, âge, profession) dont il connait la fréquence des cancers de l'estomac sur la population générale. Dans une table publiée il remarque que ce même groupe à plus de cancers de l'estomac : il en conclut que les individus faisant partie de ce groupe (de cette table) ont plus de risque d'avoir un cancer de l'estomac.

c. Les méthodes de protection de la vie privée (MPVP)

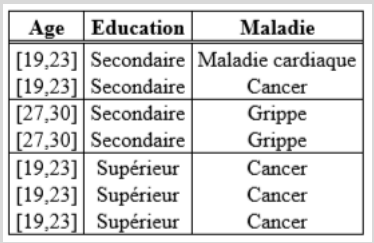
Il n'existe pas une technique éprouvée par le temps qui prend une table et la transforme en table anonymisée, cela serait idyllique et bien trop facile. Il en existe plusieurs, qui varient

par leurs degrés de fiabilité (liée au risque de ré-identification), par le cadre dans lequel on souhaite les appliquer (pour la publication, la recherche, la tarification etc.) ou par le type de data (micro, macro, continu, images, texte, catégorielles etc.).

Qu'on se le dise une bonne fois pour toutes : il n'est pas possible d'arriver à un risque nul de ré-identification. Ceci est notamment dû au développement constant de l'IT (technologie de l'information).

Dans un objectif de synthèse et de croisement de différentes sources, afin d'expliquer au mieux les différentes méthodes que l'on peut utiliser pour anonymiser un tableau, j'ai créé un tableau exposant ces dernières.

Exemple de modèles permettant de contrer des attaques de données :

Nom	Principe	Avantages	Inconvénients
Le modèle de k-anonymat	Dans une table dite k-anonyme, il y a au minimum k valeurs de quasi-identifiants dans chaque n-uplet. Exemple de table 2-Anonyme (au moins 2 enregistrements par uplet faits à partir des attributs clés, ici Age-Education).  **Age** / **Education** / **Maladie** [19,23] Secondaire Maladie cardiaque [19,23] Secondaire Cancer [27,30] Secondaire Grippe [27,30] Secondaire Grippe [19,23] Supérieur Cancer [19,23] Supérieur Cancer [19,23] Supérieur Cancer	Le degré de protection est lié à k. Cette méthode réduit le risque global en faisant baisser les fréquences d'apparition f_c associées des QI (se référer a la partie sur le calcul du risque de ré-identification) Contre uniquement les *« liaison d'enregistrements »* car il se focalise seulement sur les QI.	Le k-anonymat optimal est très couteux à trouver *[BeFre2017]*. Aucunement résistant *aux attaques : « liaison d'attribut »*, et surtout sur : « *attaques par homogénéité* » et « attaques fondées sur la connaissance de base »
Le modèle de l-diversité	On se focalise sur les attributs sensibles de la table : ils doivent être « bien »	Contre les *attaques par liaison d'attributs*.	Pour les adversaires plus expérimentés, la l-

<table>
<tr><td></td><td>représentés dans chaque classe d'équivalence.</td><td rowspan="3">Et dans l'ensemble ces techniques contrent les attaques par :
o homogénéité
o connaissances de base

La diversité des attributs sensibles réduit la menace sur la vie privée liée à ces attaques.</td><td rowspan="3">diversité ne contre pas les attaques par similarité et les attaques par inférences probabilistes (« probabilistic inference attacks ») dont celles par <u>dissymétrie</u></td></tr>
<tr><td>« l-diversité distincte »</td><td>Model simple de l-diversité où l'on cherche à obtenir des classes d'équivalence l-diverses.

Pour chaque Attribut Sensible, on fait en sorte qu'il soit représenté l fois dans chaque groupe d'individus partageant le même QI.
Table ayant la « 3-diversité distincte » (et le 4-anonymat)
<table><tr><th>Age</th><th>Education</th><th>Maladie</th></tr><tr><td>[19,23]</td><td>Secondaire</td><td>Maladie cardiaque</td></tr><tr><td>[19,23]</td><td>Secondaire</td><td>Cancer</td></tr><tr><td>[19,23]</td><td>Secondaire</td><td>Grippe</td></tr><tr><td>[19,23]</td><td>Secondaire</td><td>Grippe</td></tr><tr><td>[27,30]</td><td>Supérieur</td><td>Cancer</td></tr><tr><td>[27,30]</td><td>Supérieur</td><td>Cancer</td></tr><tr><td>[27,30]</td><td>Supérieur</td><td>Maladie cardiaque</td></tr><tr><td>[27,30]</td><td>Supérieur</td><td>Grippe</td></tr></table></td></tr>
<tr><td>« l-diversité fondée sur l'entropie »</td><td>Technique qui traduit la notion de « bonne représentation » des attributs sensibles par une entropie de distribution de ces dernières dans chaque classe d'équivalence ≥ log (1).
L'entropie est calculée avec la formule suivante :
$\text{Entropie}(C) = -\sum_{s \in S} P(qid,s)\log(P(qid,s))$
Avec C la classe d'équivalence, P (…) la proportion des individus de la table ayant la valeur s dans la classe d'équivalence.</td></tr>
</table>

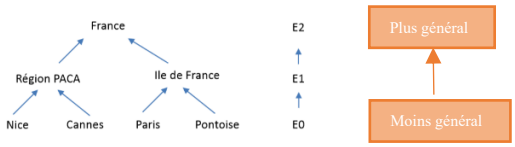
Le modèle de t-proximité (« t-closeness »)	On part du principe que l'on ne peut pas empêcher quelqu'un d'avoir accès à des informations sensibles globales sur la population. Le principe du modèle est de faire en sorte que la distribution de l'attribut sensible au sein de n'importe quelle classe d'équivalence soit proche de la distribution globale de l'attribut, ≡ la distance entre ces deux distributions doit être ≤ seuil t On privilégie la distance EMD (Earth Mover's Distance) dans la littérature. Calcul détaillé dans (Rubner, Tomasi, et Guibas 2000).		
Le modèle de δ-Présence	Afin d'empêcher que l'on puisse supposer la présence d'un enregistrement dans une table en faisant un croisement de données, le modèle de δ-présence impose que la probabilité de présence d'un enregistrement soit dans un intervalle δ = (δmin, δmax) prédéfini.	Contré par « lien de tables »	Difficile à mettre en œuvre car il suppose que l'éditeur connaisse à priori la table de rapprochement que l'attaquant est susceptible d'utiliser => problème majeur

Nous venons de voir les différents modèles de protection de la vie privée (MPVP) permettant de contrer des attaques sur la vie privée. Ces dernières contrent chacune un ou plusieurs types d'attaque. Nous pouvons conclure en vue de ce tableau et du *[Tab1]* en annexe qu'il n'existe pas de MPVP permettant de contrer tous les types d'AVP.

d. Les techniques d'anonymisation des micro-données

Concentrons-nous maintenant sur les différentes techniques d'anonymisation des micro-données qui utilisent ces MPVP. Le tableau suivant donne un bon panorama des méthodes actuelles :

Techniques d'Anonymisation des micro-données :

Nom	Principe	Type de données
La généralisation (Samarati 2001)	Elle vise à compléter le k-anonymat. Elle est non perturbatrice mais diminue la précision des données. Elle permet de confondre (c'est-à-dire tenter d'uniformiser) avec k-1 individus la table publiée. Elle transforme des QI de sorte qu'il y ait au moins k individus avec la même valeur de QI. La méthode se base sur une hiérarchie de généralisation prédéfinie : chaque QI fait l'objet d'une hiérarchie sur au moins 2 niveaux. Le lien entre 2 niveaux évoque une hypothèse de remplacement, de généralisation. Exemple : 	Continu, catégoriel

<table>
<tr><td></td><td>On peut alors remplacer les villes par les départements dans une table.
Il existe au moins 9 techniques de généralisation.</td><td></td></tr>
<tr><td>La suppression

(Lawrence H. 1980)</td><td>On retire toutes les micro-données de la table originale qui sont source d'un risque de réidentification.
Suppression globale : supprime tuple
Suppression locale : quelques données du tuple (remplacées par « nul » par ex)</td><td>Continu, catégoriel</td></tr>
<tr><td>La micro-agrégation

(Defays et Nanopoulos 1992)</td><td>Technique SDC [1] :
Renforce le k-anonymat (rassemble les enregistrements dans des groupes d'au moins k-individus : les micro-agrégats) et garantie la confidentialité des données sensibles en remplaçant la valeur de certains attributs par une mesure centrale (moyenne ou médiane usuellement) dans chaque micro-agrégat formé).

Exemple de micro-agrégation à l'attribut cholestérol, étapes :
o Division en groupes homogènes // âge pour satisfaire le 3-anonymat
o Température remplacée par la moyenne du groupe.

<table><tr><th>Num tuple</th><th>Sexe</th><th>Ville</th><th>Profession</th><th>Statut marital</th><th>Age</th><th>JH</th><th>Choles-térol</th><th>Tempé-rature</th></tr><tr><td>11</td><td>M</td><td>Paris</td><td>étudiant</td><td>célibataire</td><td>19</td><td>5</td><td>190</td><td>38</td></tr><tr><td>4</td><td>M</td><td>Nice</td><td>étudiant</td><td>célibataire</td><td>19</td><td>7</td><td>170</td><td>38</td></tr><tr><td>3</td><td>F</td><td>Paris</td><td>étudiant</td><td>célibataire</td><td>21</td><td>2</td><td>190</td><td>38</td></tr><tr><td>6</td><td>F</td><td>Cannes</td><td>étudiant</td><td>mariée</td><td>28</td><td>3</td><td>185</td><td>38,06</td></tr><tr><td>1</td><td>F</td><td>Paris</td><td>ingénieur</td><td>mariée</td><td>33</td><td>3</td><td>150</td><td>38,06</td></tr><tr><td>5</td><td>M</td><td>Paris</td><td>statisticien</td><td>marié</td><td>36</td><td>40</td><td>200</td><td>38,06</td></tr><tr><td>7</td><td>F</td><td>Pontoise</td><td>statisticien</td><td>divorcée</td><td>46</td><td>60</td><td>200</td><td>37,2</td></tr><tr><td>9</td><td>M</td><td>Cannes</td><td>statisticien</td><td>marié</td><td>58</td><td>10</td><td>260</td><td>37,2</td></tr><tr><td>2</td><td>F</td><td>Pontoise</td><td>ingénieur</td><td>veuve</td><td>65</td><td>1</td><td>290</td><td>37,2</td></tr><tr><td>10</td><td>F</td><td>Nice</td><td>professeur</td><td>veuve</td><td>63</td><td>7</td><td>290</td><td>37,2</td></tr><tr><td>8</td><td>M</td><td>Pontoise</td><td>statisticien</td><td>divorcé</td><td>81</td><td>5</td><td>300</td><td>37,2</td></tr></table>
3</td><td>Continu</td></tr>
<tr><td>La technique de</td><td>Permet de créer des tables l-diverses.</td><td>Continu, catégoriel</td></tr>
</table>

[3] A noter qu'il existe beaucoup de techniques d'anonymisation : SDC (Statistical Disclosure Control), SDL (Statistical Disclosure Limitation), PPDM (Privacy Preserving Data Mining), PPDP (Privacy Preserving Data Publishing).

<table>
<tr>
<td>« bucketisation »

(Martin et al. 2007)</td>
<td>Cette technique consiste à permuter de façon aléatoire les attributs sensibles au sein d'un même segment (bucket en anglais).

Cela permet, contrairement à la généralisation, de maintenir les valeurs originales des attributs du QI dans la table anonyme.

Mais elle supprime les corrélations entre les attributs du QI et les attributs sensibles : donc dégradation de l'info.</td>
<td></td>
</tr>
<tr>
<td>La technique « Anatomy »

(Xiao et Tao 2006)</td>
<td>Comme bucketisation : créer table l-diverse, contre désavantage de la généralisation.

<table>
<tr><th>groupe</th><th>Maladie</th><th>fréquence</th></tr>
<tr><td>1</td><td>maladie cardiaque</td><td>1</td></tr>
<tr><td>1</td><td>cancer</td><td>1</td></tr>
<tr><td>1</td><td>grippe</td><td>1</td></tr>
<tr><td>2</td><td>grippe</td><td>1</td></tr>
<tr><td>2</td><td>cancer</td><td>2</td></tr>
</table>
Elle casse le lien entre le QI et les attributs sensibles en créant deux tables séparées à partir d'une table originale. Ces tables sont reliées entre elles en créant un identifiant propre à chaque tuple.

Exemple :
Table des attributs QI / Table des attributs sensibles

<table>
<tr><th>Age</th><th>Niveau</th><th>groupe</th></tr>
<tr><td>19</td><td>Bac+2</td><td>1</td></tr>
<tr><td>19</td><td>Bac+3</td><td>1</td></tr>
<tr><td>27</td><td>Bac+3</td><td>1</td></tr>
<tr><td>30</td><td>Bac+3</td><td>2</td></tr>
<tr><td>23</td><td>Bac</td><td>2</td></tr>
<tr><td>23</td><td>Bac</td><td>2</td></tr>
</table>
</td>
<td>Continu, catégoriel</td>
</tr>
<tr>
<td>La technique de « Slicing »</td>
<td>Garantit la l-diversité, se fonde sur deux partitionnements :
o Un vertical sur les attributs (lié à leur corrélation)
o Un horizontal concerne les tuples de la table</td>
<td>Continu, catégoriel</td>
</tr>
</table>

<table>
<tr><td>(T. Li et al. 2012)</td><td>Puis à l'intérieur de chaque partition on permute aléatoirement des valeurs d'attributs pour casser le lien dans une partition verticale. Idem avec le partitionnement horizontal.

<u>Exemple de table issue du Slicing</u> :
<table><tr><th>Age</th><th>(Niveau d'études, Maladie)</th></tr><tr><td>19</td><td>(Bac+2, maladie cardiaque)</td></tr><tr><td>19</td><td>(Bac+3, grippe)</td></tr><tr><td>27</td><td>(Bac+3, cancer)</td></tr><tr><td>23</td><td>(Bac, cancer)</td></tr><tr><td>23</td><td>(Bac, grippe)</td></tr><tr><td>30</td><td>(Bac + 3, cancer)</td></tr></table></td><td></td></tr>
<tr><td>La permutation ou technique de "Swapping"

(Dalenius et Reiss 1982)</td><td>Technique SDC

Assez évidente : permutations de valeurs d'un même attribut au sein d'un sous-ensemble de tuples.

Il existe au moins quatre variantes : Random Swap, Rank Swap, C&C Swap et Target Swap.</td><td>Continu, catégoriel</td></tr>
<tr><td>Le recodage global

(Domingo-Ferrer et Torra 2001-2002)</td><td>Il s'agit de partitionner un attribut numérique en plusieurs intervalles de la même taille.
Chaque valeur de la variable se voit remplacée par un interval.</td><td>Continu</td></tr>
<tr><td>Les techniques de « Top Coding » et de « Bottom Coding »

(Domingo-Ferrer et Torra 2001)</td><td>Top Coding :
Reproduire dans la table anonyme toutes les données originales d'un attribut, hormis celles qui dépassent une valeur-seuil prédéfinie, remplacée par une valeur.

Bottom Coding :
Reproduire dans la table anonyme toutes les données originales d'un attribut, hormis celles qui sont inférieures à une valeur-seuil prédéfinie, remplacée par une valeur.</td><td>Continu, catégoriel</td></tr>
</table>

<table>
<tr>
<td>Le bruit aléatoire

(« Random Noise »)
(Brand 2002)</td>
<td>On considère 1 seul attribut, et on multiplie chaque valeur par une valeur aléatoire (on en connait sa distribution et sa moyenne). Deux types de bruits :

o <u>Multiplicatif</u> : on multiplie chaque valeur par ε aléatoire.
o <u>Additif</u> : on ajoute ε à chaque valeur
- Additif non corrélé (préserve les moyennes, les covariances et perturbe les corrélations et variances)
- Additif corrélé (préserve les moyennes et les corrélations.)

Exemple d'additif non corrélé :
<table>
<tr><th>Valeur originale</th><th>Valeur aléatoire</th><th>Valeur modifiée</th></tr>
<tr><td>3</td><td>2</td><td>5</td></tr>
<tr><td>1</td><td>1</td><td>2</td></tr>
<tr><td>2</td><td>5</td><td>7</td></tr>
<tr><td>7</td><td>3</td><td>10</td></tr>
<tr><td>40</td><td>-10</td><td>30</td></tr>
<tr><td>3</td><td>8</td><td>11</td></tr>
<tr><td>60</td><td>-11</td><td>49</td></tr>
<tr><td>5</td><td>4</td><td>9</td></tr>
<tr><td>10</td><td>-3</td><td>7</td></tr>
<tr><td>7</td><td>-2</td><td>5</td></tr>
<tr><td>5</td><td>3</td><td>8</td></tr>
</table>
Pour calculer ε, il faut que la moyenne des ε soit nulle et que sa variance soit proportionnelle à celle des données initiales.</td>
<td>Continu</td>
</tr>
</table>

Le *[Tab2]* en annexe met en valeur les techniques perturbatrices et celles qui ne le sont pas. Il apparait intéressant de n'utiliser que celles qui ne font pas perdre de l'information comme la généralisation, le bottom coding ou le top coding. Cependant, en mesurant la perte d'information liée aux autres méthodes on peut tout à fait les incorporer dans un algorithme d'anonymisation de données.

e. <u>L'application de ces techniques de dé-identification</u>

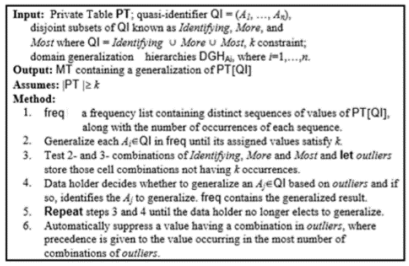
Le problème majeur de ces différentes techniques d'anonymisation est qu'elles ne sont accessibles qu'à une infime portion de la population. En effet, ces techniques sont décrites dans des documents scientifiques (thèses, articles académiques) et sont succinctement représentées

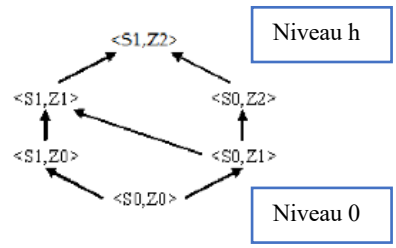
par des exemples. Il se pose alors une barrière technique entre les scientifiques spécialisés en programmation et les professionnels désirant respecter les règlementations en termes de protection de la donnée.

En termes d'outils mis à disposition qui implémentent ces techniques, on pourrait citer le logiciel mu-argus utilisé par les instituts de statistique publique en Europe.

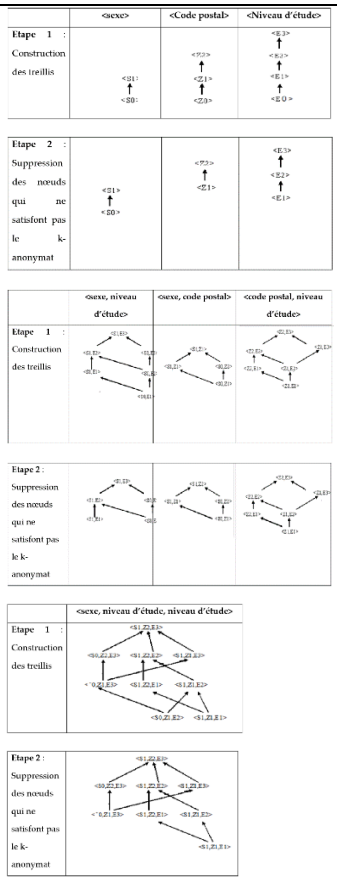
D'après Feten Ben Fredj, de façon générale, ces outils ne confèrent pas à son utilisateur un guide pratique des techniques à utiliser et leur paramétrage. D'où l'importance de connaitre les algorithmes utilisés, leurs impacts sur les données (dégradation, perte d'information) et la nécessité de calculer les risques de ré-identification avant et après application de ces derniers.

Quelques exemples d'algorithmes d'implémentation de généralisation :

Nom	Code (détail)	Infos div
N°1 : Algorithme de µ-argus	 **Input:** Private Table PT; quasi-identifier QI = $(A_1, \ldots, A_n)$, disjoint subsets of QI known as *Identifying*, *More*, and *Most* where QI = *Identifying* ∪ *More* ∪ *Most*, k constraint; domain generalization hierarchies DGH_{A_i} where i=1,…,n. **Output:** MT containing a generalization of PT[QI] **Assumes:** |PT| ≥ k **Method:** 1. freq a frequency list containing distinct sequences of values of PT[QI], along with the number of occurrences of each sequence. 2. Generalize each A_i∈QI in freq until its assigned values satisfy k. 3. Test 2- and 3- combinations of *Identifying*, *More* and *Most* and **let** *outliers* store those cell combinations not having k occurrences. 4. Data holder decides whether to generalize an A_j∈QI based on *outliers* and if so, identifies the A_j to generalize. freq contains the generalized result. 5. **Repeat** steps 3 and 4 until the data holder no longer elects to generalize. 6. Automatically suppress a value having a combination in *outliers*, where precedence is given to the value occurring in the most number of combinations of *outliers*.	À chaque itération : - l'utilisateur choisit le QI à généraliser - affichage de la liste des individus ne respectant pas la k-identité - choix de poursuivre ou non - suppression de ces individus. Ce code est incorporé dans le package **SDC Micro** sur R (en plus d'autres méthodes d'anonymisation de micro-données) **µ-Argus** et **sdcMicro**, logiciels de gestion de la confidentialité pour les données individuelles, utilisés par les instituts de statistique publique en Europe.
N°2 : Algorithme de Datafly		Automatisation de la suppression des tuples de µ-argus. L'algorithme s'arrêtera dès qu'il y aura un nombre de tuples à supprimer est au dessus ou en

	Input: Private Table PT; quasi-identifier QI = $(A_1, \ldots, A_n)$, k constraint; hierarchies DGH_{A_i} where $i=1,\ldots,n$. **Output:** MGT, a generalization of PT[QI] with respect to k **Assumes:** $\|PT\| \geq k$ **Method:** 1. freq a frequency list contains distinct sequences of values of PT[QI], along with the number of occurrences of each sequence. 2. **while there exists** sequences in freq occurring less than k times that account for more than k tuples **do** 2.1. **let** A_j be attribute in freq having the most number of distinct values 2.2. freq □ generalize the values of A_j in freq 3. freq □ suppress sequences in freq occurring less than k times. 4. freq □ enforce k requirement on suppressed tuples in freq. 5. **Return** MGT □ construct table from freq	dessous d'un seuil de tolérance entré par l'utilisateur. C'est l'algorithme qui choisit le QI considéré (dans l'ordre de celui qui contient le plus de valeurs distinctes vers celui qui en contient le moins, autrement dit, l'algorithme affecte un score de tolérance à la distorsion et commence par les QI qui sont le plus tolérants donc avec le plus de valeurs distinctes).
N°3 : Algorithme de Samarati	**Find_vector** INPUT: Table $T_i = PT[QI]$ to be generalized, anonymity requirement k, suppression threshold MaxSup, lattice VL_{DT} of the distance vectors corresponding to the domain generalization hierarchy DGH_{DT}, where DT is the tuples of the domains of the quasi-identifier attributes. OUTPUT: The distance vector sol of a generalized table GT_{sol} that is a k-minimal generalization of PT[QI] according to Definition 4.3. METHOD: Executes a binary search on VL_{DT} based on height. of vectors in VL_{DT}. 1. $low := 0$; $high := height(T, VL_{DT})$; $sol := ?$ 2. **while** $low < high$ 2.1 $try := \lfloor \frac{low+high}{2} \rfloor$ 2.2 $Vectors := \{vec \mid height(vec, VL_{DT}) = try\}$ 2.3 $reach, k := false$ 2.4 **while** $Vectors \neq \emptyset \wedge reach, k \neq$ true **do** Select and remove a vector vec from $Vectors$ **If** $satisfies(vec, k, T_i, MaxSup)$ **then** $sol := vec$; $reach, k := true$ 2.5 **if** $reach, k :=$ true **then** $high := try$ **else** $low := try + 1$ 3. **Return** sol  Treillis de généralisation sur 2 paramètres	Ici, le choix des attributs QI à généraliser ne se base plus sur un score de tolérance à la distorsion, mais : - Dans un premier temps est créé un Treillis de généralisation, arbre contenant à chaque nœud une combinaison synthétisant l'état d'avancement de la généralisation. Ainsi si on considère 2 QI, le premier avec 2 niveaux, l'autre avec 3 alors il y aura 2*3=6 nœuds. - Ensuite à chaque itération de l'algorithme, on se place à la mi-hauteur de l'arbre Treillis non exploré (donc au début de l'arbre), puis on applique la généralisation de chaque nœud 1 à 1 à la Table privée. Les nœuds permettant de satisfaire le k-anonymat (avec ou non suppression globale) sont mémorisés.

		L'algorithme passe ensuite à la partie inférieure de l'arbre, fait une généralisation sur tous les nœuds de la mi-hauteur de la partie inférieure. Si aucun n'est satisfaisant, il passe à la partie supérieure de cette partie inférieure de l'arbre restant et ainsi de suite. A la fin il en ressort la dernière liste de nœuds enregistrés : ce sont les types de généralisation qui permettent de respecter le k-anonymat choisi.

N°4 : Algorithme Incognito (LeFevre, DeWitt, et Ramakrishnan 2005) Incognito	 	Se base aussi sur l'élaboration d'un treillis et sur 3 propriétés (p.199 thèse en com.) : - Celle de généralisation - Celle des sous-ensembles - Celle du Rollup Il fonctionne avec une boucle itérative sur i, le nombre d'attributs considérés. A chaque itération : établissement de plusieurs treillis (exemple si **i**=1 et qu'il y a **n** attributs au total : **n** treillis sont faits) sur la base de la hiérarchie de généralisation pour l'étape 1, puis fusion des treillis entre eux lorsqu'on augmente le nombre d'attributs (**i**, donc étapes suivantes) par arbre. Et on supprime les nœuds ne validant pas le k-anonymat grâce aux propriétés ci-dessus. A la fin on obtient 1 treillis utilisant **n** attributs et vérifiant le k-anonymat partout.
N°5 Algorithme Généralisation ascendante (« Bottom up generalization »)	$IG/AL(S) = \begin{cases} InformationGain(S) \\ \frac{InformationGain(S)}{AnonymityLoss(S)} \end{cases}$ if $AnonymityLoss(S) \neq 0$; otherwise $InformationGain(S) = Entropy(R_a) - \sum_{si} \frac{\lvert R_{si} \rvert}{\lvert R_a \rvert} Entropy(R_a)$ $Entropie(R_x) = -\sum_{cls} \frac{freq(R_x, cls)}{\lvert R_x \rvert} \times \log_2 \frac{freq(R_x, cls)}{\lvert R_x \rvert}$ R_a/R_{si} :ensemble des enregistrements contenant la valeur a (ou si) $freq(R_x, cls)$: le pourcentage d'individus de la classe labellisée cls dans Rx. S : spécialisation choisie par l'algorithme AnonymityLoss (S) : perte d'anonymat apres application de S = Anonymat(T, après S) - Anonymat(T, avant S) Anonymat(T, après S) = taille de la plus petite classed'équivalence de T apres S. Anonymat(T,avant S) = idem mais avant d'appliquer S. **Algorithm 1** The bottom-up generalization 1: **while** R does not satisfy the anonymity requirement **do** 2: **for all** generalization G **do** 3: compute $IP(G)$; 4: **end for**; 5: find the best generalization G_{best}; 6: generalize R by G_{best}; 7: **end while**; 8: output R;	But : rendre les données propices à la classification. A chaque itération, l'algorithme considère les généralisations des attributs choisis, en partant des feuilles vers la racine (i.e. l'état le plus généralisé de l'attribut). A chaque itération, le meilleur attribut à généraliser est sélectionné s'il permet de

(Wang, Yu, et Chakraborty 2004)		se rapprocher du k-anonymat et a le plus grand IL/AG (formule à côté) : c'est le *« score de spécialisation »*. L'algorithme s'achève lorsque le k-anonymat est atteint.
N°6 La spécialisation descendante : « Top Down Specialization » ou TDS (B. C. Fung, Wang, et Yu 2005)		But : rendre les données propices à la classification. A la différence de la généralisation ascendante, TDS part de la racine vers les feuilles des hiérarchies de généralisation. Le choix de la généralisation repose sur le même score de spécialisation calculé selon la même formule.

1. Évaluer le risque de dé-anonymisation

Dans cette partie je tâcherais d'expliquer au mieux comment il est possible que quantifier le risque de divulgation porté par un individu ou le risque global porté sur tout un jeu de données. Cette partie n'aurait pas pu être réalisée sans le croisement et la mise en perspective de plusieurs documents scientifiques dont principalement le document de travail de 2016 élaboré par Maxime BERGEAT et celui élaboré par Thijs BENSCHOP, Cathrine MACHINGAUTA et Matthew WELCH en 2018. Il faudra ainsi se référer à ces travaux en annexe en ce qui concerne les démonstrations, seule la finalité de leur raisonnement sera prise en compte ici.

a. Lorsque les attributs considérés sont catégoriels

Dans cette partie on suppose les QI (quasi identifiants) comme catégoriels, la partie suivante traitera le cas particulier des variables continues. Afin que l'état de l'art de ces différents calculs ne soit pas trop fastidieux à lire, j'ai réalisé une synthèse des différentes méthodes en tableau , il en existe deux types, a priori ou a posteriori.

i. Le risque d'appariement, à postériori

Étude de l'appariement des individus (à postériori)
On considère que l'on connait deux tables, l'initiale **A** et celle anonymisée **B**. On peut alors mesurer la distance séparant chaque individu de **B** à ceux de **A**, et, pour tout **b** ∈ **B** estimer le **a** le plus proche. Pour des exemples de ré-identificateurs on pourrait citer : - Winkler, 2004 "Re-Identification methods for masked microdata" dans *International Conference on Privacy in Statistical Databases*, Springer, p.216-230. - Skinner, 2008, « Assessing disclosure risk for record linkage » dans *International Conference on Privacy in Statistical Databases*, Springer, p.166-176. Il détaillait des appariements probabilistes (puisque qu'il s'agit d'une technique utilisable pour des variables continues, je la détaillerai dans la partie suivante). Il existe une technique de calcul du risque de ré-identification par appariement dans le package R sdcMicro. Ps : il est nécessaire de standardiser les valeurs pour éviter des différences d'échelle dans le calcul des distances.

ii. Le risque par estimation des Fc, risque calculé à priori

Le risque d'une clé d'identification est directement lié au concept de rareté. Une clé d'identification est combinaison de QI. Chaque individu possède une clé d'identification unique. Une clé est dite risquée si peu d'individus de la table partagent cette même clé. C'est là que le concept de k-anonymat intervient : il garantit que au moins k individus de la tables partagent la même clé. On en vient à estimer la fréquence d'apparition de cette clé dans la population F_c à partir de celle que l'on a dans la table c'est-à-dire les f_c pour calculer le risque de ré-identification.

On note :

- C le nombre de clés d'identification distinctes dans la table considérée
- c une clé d'identification c ∈ [|1,C|].
- (F_1, ..., F_C) et (f_1, ..., f_C) les fréquences d'apparition de chacune des clés d'identification dans la population U (inconnues) et l'échantillon s (observées), respectivement.

Estimation du risque par estimation des F_c (à priori)
$r_c = \mathbb{E}\left(\frac{1}{F_c}\middle\vert f_c\right)$

Risque individuel en utilisant les poids de sondage	(Benedetti et Franconi, 1998) $\hat{r}_c = \frac{\hat{p}_c^{f_c}}{f_c} \, {}_2F_1(f_c, f_c; f_c + 1; 1 - \hat{p}_c)$ Avec : $\hat{p}_c = \frac{f_c}{\sum_{i=1, i \text{ possède la clé } c}^{n} w_i}.$ $F_1(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} \sum_{n=0}^{+\infty} \frac{\Gamma(a+n)\Gamma(b+n)}{\Gamma(c+n)} \frac{z^n}{n!}$	Méthode qui prend en compte les poids de sondage (il peut être interprété comme le nombre d'unité distinctes présente dans la population du sondage que chaque unité échantillonnée représente). D'après Maxime BERGEAT, dans le cas d'un fichier exhaustif ou d'un plan de sondage simple type aléatoire il est préférable d'utiliser le model suivant.
Modèle de Poisson	On suppose : $F_c \sim$ Poisson(λ_c), indépendamment, $F_c = 0,1,...$ et qu'on a exécuté un tirage de Bernoulli au sein de la population. $\hat{r}_c = \frac{1}{\hat{\lambda}_c(1-\pi)} \left(1 - \exp^{-\hat{\lambda}_c(1-\pi)}\right)$ $\hat{\lambda}_c = \frac{\exp(x_c' \hat{\beta}^{MV})}{\pi}$	Ne prend pas en considération le poids de sondage des individus. Ce modèle d'estimation est présent dans le package sdcMicro.

D'autres méthodes d'estimation des risques de ré-identification sur des modèles statistiques existent mais ils ne sont pas applicables avec celles utilisées par les instituts de statistique publique en Europe selon Maxime BERGEAT.

Pour bien expliquer pourquoi on cherche les Fc pour calculer le risque je vais exposer un exemple. Admettons que dans une table **A** (celle anonymisée) on cherche à identifier l'individu **i** (on ne sait pas s'il est dans **A**). Dans **A** on suppose que **i** se trouve dans un groupe de **k** individus : $\mathbf{f_{ci} = 1/k}$. Il se trouve que dans une autre table, la table **B** on sait que **i** se trouve dans un groupe de **k'** individus : $\mathbf{F_{ci} = 1/k'}$. Ainsi la probabilité que **i** se trouve dans **A** :

$P(i \in A) = \frac{k}{k'} = \frac{F_{ci}}{f_{ci}}$ C'est donc le risque de déduire des informations sur i en connaissant des informations sur une autre table.

Risque calculé par ménage : il s'agit du risque, ou plutôt de la probabilité de ré-identification d'un individu d'un ménage sachant que l'on en a déjà identifié un. Si l'on note un ménage **g** avec **|g|** individus notés **(i_1, ..., $i_{|g|}$),** le risque lié à ce ménage est : $r_g^M = \mathbb{P}\left(i_1 \cup i_2 \cup \ldots \cup i_{|g|}\right)$

Grâce au package sdcMicro une manière de calculer ce risque est de rassembler les ménages grâce à un attribut commun (une adresse, un numéro de contrat par exemple) avec la fonction

selectHouseholdData() puis de calculer le risque par individus de cette sélection avec la fonction indivRisk().

Risque global (pour tout une table d'individus) : utilise une règle de fréquence minimale avec un seuil minimal par clé identifiante noté s, on peut utiliser une mesure globale du risque R en calculant la proportion d'individus à risque. Cette méthode est implémentée dans le package sdcMicro en utilisant : sdcInitial@risk$global$risk (si le sdc créé se nomme sdcInitial).
Et si on arrive à calculer le risque associé à chaque clé d'identification r_c , on peut alors calculer ce risque global comme suit : $\hat{R} = \sum_{c=1}^{C} \hat{r}_c$
On peut pondérer les r_c par la fréquence f_c des individus ayant la clef c : $R = \frac{1}{n} \sum_{c=1}^{C} f_c \times \mathbf{1}_{f_c<s}$
Il existe plusieurs variantes possibles.

On peut appliquer les techniques pour les variables continues, il suffira juste de faire des modifications préalables, comme une généralisation, ou remplacer des valeurs par des intervalles.

iii. SUDA une technique de détection des clés à risque

L'algorithme SUDA (Special Uniques Detection Algorithm) permet de classer les clés d'identification en fonction de leur degré de spécialité, on obtient ainsi une hiérarchisation des risques plus fine qu'avec la règle de fréquence minimale. L'algorithme SUDA est implémenté dans le package R sdcMicro. La majeure partie des informations de cette sous-partie sont issues de *[SDC_Practice2018]*.

Les méthodes type k-anonymat et l-diversité s'attardent à trouver ce qu'on appelle des variables clés, ou plutôt des combinaisons de QI pour lesquelles il peut exister des informations provenant d'autre sources dont le croisement permet de dégager des informations personnelles. Le problème de cette approche c'est qu'il est particulièrement difficile en pratique d'avoir une vision bien large sur toutes les données accessibles, toutes les variables, connaitre leurs liens entre elles et les risques externes.

C'est alors qu'intervient une mesure particulière sur les uniques dits « spéciaux » : il s'agit du score SUDA. Cet algorithme implémenté dans le package sdcMicro de R s'attarde à

détecter les éléments uniques. Cette méthode est dite heuristique, on ne peut donc évaluer ses performances que face à des données réelles.

Tableau comparatif :

k-anonymat / l-diversité	On trouve une clé (combinaison de QI), on créé des classes et on appliques des méthodes de protection de la vie privée (MPVP) dans le but de réduire le nombre d'individu par section.
SUDA	Ici, pour chaque individu de la table l'algorithme on répertorie tous les MSU (Minimal Sample Unique). Un MSU associé à une ligne d'une table (donc un individu) c'est un vecteur de valeurs d'attributs. La combinaison de ces valeurs pour l'individu considéré le rend unique dans toute la table. Chaque ligne possède ainsi plusieurs MSU de différentes tailles. La taille d'un MSU est le nombre de variable qu'il contient. Une fois les MSUs collectés, on y associe un score déterminant le risque de ré identification de cet element : - Plus la taille des MSUs est petite, plus le risque est grand. - Plus un individu a de MSUs, plus le risque est grand.

Exemple de calcul du score SUDA, en utilisant le *[Tab3]* en annexe.

L'enregistrement 5 possède 4 MSU (A, B, C, D)

- A : {'Secondary complete', 'Unemployed'}
- B : {'Female', 'Unemployed'}
- C : {'Female', 'Secondary Complete'}
- D : {'Rural'}

Pour vérifier si on a bien un MSU de taille k il faut vérifier une exigence minimale. Cette exigence stipule que tout sous ensemble du MSU en question (donc taille k-1) se doit de ne pas être unique dans la table considérée. Par exemple A est un MSU car {'Secondary complete'} et {'Unemployed'} ne sont pas unique dans la table.

Détail du calcul du score SUDA pour le n^{ème} élément d'une table

$$SUDA_n = \sum_{p=1}^{P_n} \prod_{i=k_p}^{M} (Var - i)$$

P_n : nombre de MSU du n^{ème} élément de la table

p : MSU considéré

k_p : taille du $p^{ème}$ MSU du $n^{ème}$ élément de la table

M : taille maximale que l'utilisateur souhaite utiliser dans le calcul du SUDA

Var : nombre total d'attributs ou de variables de la table

Plus le SUDA est élevé plus le $n^{ème}$ élément de la table a un risque d'être unique.

Une métrique d'évaluation du risque d'intrusion sur une table : le DIS metric

Dis signifie : Data Intrusion Simulation (DIS), il s'agit d'une méthode d'évaluation du risque de divulgation de données. On peut combiner DIS et SUDA d'après Elliot et Manning (2003) *[ElMa2003*] : une mesure DIS entre les éléments de la table est calculée grâce au score SUDA. Ce dernier correspond à la probabilité qu'un échantillon unique trouvé par un attaquant coïncide avec celui issu d'une table externe donc entre 0 et 1. Le calcul de cette probabilité est détaillé dans le livre de Elliot et Manning.

A noter que le package sdcMicro de R permet à la fois de calculer le SUDA et le DIS-SUDA.

b. Lorsque les attributs considérés sont continus

Les méthodes précédemment vues sont applicables sur des attributs catégoriels ou le principe de rareté a un sens. Bien entendu grâce à la MPVP de généralisation toute variable continue peut-être transformée en catégorielles (cf. thèse de Feten Ben Fredj). Sur ce type de variable on parlera de rareté dans un intervalle encadrant la valeur d'un attribut continu.

Le calcul de risque de divulgation se fait pour la plupart à postériori en utilisant une mesure de distance entre une table anonymisée et celle initiale (cf. plus haut lorsque j'évoquais l'étude de l'appariement des individus).

On définira la rareté d'une valeur selon deux manières différentes :

- En terme relatif avec le « record linkage » :

  C'est une manière d'évaluer le risque de perturbation et non celui initial, à posteriori donc. Deux méthodes sont alors possibles :

  - Rassembler les individus similaires sur une base déterministe. La technique la plus simple consiste à choisir un identifiant unique, notons le ID_A. Puis, le calcul des distances entre les ID_A des données brutes et ceux de celles traitées fait, si la distance limite n'est pas dépassée : il y a correspondance, sinon non.

- Coupler les individus sur une base probabiliste (probabilité que les individus de la table brute soient reliés aux bonnes personnes de la table anonymisée). La différence avec la méthode précédente est la manière d'aborder la méthode de couplage en prenant en compte un plus grand nombre d'indicateurs potentiels. Pour détailler de façon succincte cette méthode, d'après une étude probabiliste : ***Erreur ! Source du renvoi introuvable.***, il s'agit de calculer une pondération p our chaque identificateur en fonction de sa capacité estimée à identifier correctement une correspondance ou une non-correspondance. On supposera 0 comme un désaccord total, 1 un accord complet, et dans]0 ;1[un accord partiel. On peut créer autant de barème qu'il y a de variables prises en compte. Il suffit alors d'utiliser ces poids/notes pour calculer la probabilité que deux enregistrements donnés se rapportent à la même entité.

  Cette méthode est jugée moins efficace que la précédente d'après *[SDC_Practice2018]*.

- En terme absolu avec le « intervalle measure ». Cette méthode consiste à créer, pour chaque variable considérée dans la table brute un intervalle de risque. Cet intervalle est calculé en fonction de l'écart type des valeurs de la table et d'un paramètre d'échelonnage. Puis, en observant les valeurs de la table altérée : lorsqu'une valeur est en dehors de l'intervalle elle est considérée comme sûre, sinon il sera nécessaire d'altérer un peu plus sa valeur. Cette méthode est implémentée dans le package sdcMicro sous le nom de la fonction dRisk().

Un autre risque lié aux valeurs continues est celui des valeurs aberrantes. En effet même si on ajoute un certain degré d'incertitude autour de la valeur d'une aberration, cela en reste une. La question est donc de les identifier.

- Une première méthode est d'utiliser les quantiles. Par exemple, on pourra considérer qu'un individu de la table ayant un attribut supérieur au quantile à 90% comme étant un individu à risque car appartenant aux 10% des individus ayant les valeurs les plus grandes sur cette variable.

- La seconde méthode se base sur le « intervalle measure » décris précédemment. On créé des intervalles de risques autour des valeurs de la table brute et on évalue le risque dans la table altérée. Mais là où cela devient intéressant, c'est que d'après Templ et Meindl *[TeMe2008]* ont élaboré une méthode de création d'intervalles fondée sur le « squared Robust Mahalanobis Distance » RMD en abrégé. Cette dernière prend en compte les valeurs aberrantes dans le sens ou elle crée des intervalles de risque plus larges pour ces dernières. Cette méthode est implémentée dans le package sdcMicro sous le nom de dRiskRMD().

Conclusion

Au sein de ce document j'ai tâché au mieux d'expliquer le contexte lié à la gestion des données sensibles. Le RGPD et son application en France ont récemment ajouté des contraintes importantes au sein de toute entreprise détenant des informations personnelles liées à leur client. Heureusement les techniques de protections de données se perfectionnent de jours en jours. Mes recherches ont permis de comprendre les principales et de découvrir comment elles pourraient être mises en pratique. Vérifier ces dernières, et étudier comment les assureurs se parent face à ce nouveau contexte constituerons la suite de mes recherches au sein du DIR4.

Table des matières

Bibliographie

I. Livres

[Ch2018]
Chokogoue, Juvénal (2018) Maîtrisez l'utilisation des technologies Hadoop : Initiation à l'écosystème Hadoop Ed. 1, : Eyrolles.

[DePa2018]
Guillaume Desgens-Pasanau, Sophie Nerbonne (préface) (2018) La protection des données personnelles. Le RGPD et la nouvelle loi française, 3ème edition.: LexisNexis.

[ElMa2003]
Elliot , M. J., & Manning, A. M. (2003). Using DIS to Modify the Classification of Special Uniques. Invited Paper. Joint ECE/Eurostat Work Session on Statistical Data Confidentiality. Luxemboug 2-9 April 2003.

II. Articles Académiques et Scientifiques

[AcTe2016]
Académie des technologies (2016) *Big Data : un changement de paradigme peut en cacher un autre*, : EDP Sciences.

[APRIL2007]
Association de promotion de l'information libre (APRIL), l'Aproged et le Cigref (2007) Le livre blanc du Logiciel Libre.

[Be2016]
Maxime BERGEAT (2016) La gestion de la confidentialité pour les données individuelles, document de travail, DMCSI (PARIS 12)

[BeFre2017]
Feten Ben Fredj (2017) Méthode et outil d'anonymisation des données sensibles, Conservatoire national des arts et metiers - CNAM,: HAL.

[DataCore2018]
DataCore Software (2018) The State of Software-Defined Storage, Hyperconverged and Cloud Storage: Seventh Annual Market Survey, Livre blanc, Disponible à: https://www.datacore.com/document/state-of-sds-hci-cloud-storage-seventh-annual/ (Accédé: 20 Avril 2019).

[HiLo2011]
Martin Hilbert, Priscila López (2011) 'The world's technological capacity to store, communicate, and compute information', Science, 332(6025), pp. 60-65 [Online]. Disponible à: https://science.sciencemag.org/content/332/6025/60 (Accédé: Octobre 2018).

[IDC_2016]
Étude de l'IDC de 2016 relayée par plusieurs sites: lebigdata.fr, comarketing-news.fr, journaldunet.com etc. (Accédé: Novembre 2018)

[ReGaRy2018]
David Reinsel – John Gantz – John Rydning (2018) The Digitization of the World From Edge to Core, Livre blanc de l'IDC sponsorié par Seagate. Disponible à: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf (Accédé: Novembre 2018).

[SaBeBlSt2015]
Adrian Sayers, Yoav Ben-Shlomo, Ashley W. Blom et Fiona Steele (2015) Probabilistic record linkage : International Epidemiological Association, Oxford University Press (OUP). Disponible en telechargement à : https://www.researchgate.net/publication/287791889_Probabilistic_record_linkage (Accédé: 20 Avril 2019)

[TeMe2008]
Templ, M. & Meindl, B. (2008) Robust Statistics Meets SDC: New Disclosure Risk Measures for Continuous Microdata Masking. In Privacy in Statistical Databases, PSD 2008 (eds. Domingo-Ferrer J. and Saygin Y.), vol. 5262 of Lecture Notes in Computer Science, pp. 177-189. Berlin/Heidelberg: Springer.

III. Sites internet

[CNIL2018]
CNIL (2018) Règlement européen sur la protection des données : ce qui change pour les professionnels, Disponible à: https://www.cnil.fr/fr/reglement-europeen-sur-la-protection-des-donnees-ce-qui-change-pour-les-professionnels (Accédé: Mai 2019).

[DonRGPD2017]
Coheris (2017) La sanction RGPD : Les amendes liées aux violations du Règlement, Disponible à : https://donnees-rgpd.fr/les-regles/sanction-rgpd/ (Accédé: Mai 2018).

[Diri2019]
(2019) RGPD ce qui a changé pour les entreprises, Disponible à: http://www.dirigeant.fr/011-1898-RGPD-ce-qui-a-change-pour-les-entreprises.html (Accédé: 10 Avril 2019).

[HeJo2008]
Hellerstein, Joe (9 November 2008). "Parallel Programming in the Age of Big Data". Gigaom Blog. Website: http://gigaom.com/2008/11/09/mapreduce-leads-the-way-for-parallel-programming/

[Ro2016]
Ropert, R.S. (7 Avril 2016) Internet of things, Disponible à : https://en.idate.org/internet-of-things-2/ (Accédé : Octobre 2019).

[Rh2015]
David-Julien Rahmil (2015) Vie privée et données personnelles, quelles différences ? Disponible:https://digital-society-forum.orange.com/fr/les-forums/474vie_privee_et_donnees_personnelles_quelles_differences (Accédé: Avril 2019).

[SDC_Practice2018]
Thijs Benschop, Cathrine Machingauta, Matthew Welch (2018) Measuring Risk, Disponible à: https://sdcpractice.readthedocs.io/en/latest/measure_risk.html (Accédé: Avril 2019).

Annexes

I. Tableaux

[Tab1]
Synthèse des modèles de protection de la vie privée (MPVP), Source : Feten Ben Fredj (2017) Méthode et outil d'anonymisation des données sensibles, Conservatoire national des arts et métiers - CNAM,: HAL.

Modèle de protection De la vie privée	Liaison d'enregistrement	Liaison d'attribut			Liaison de table
		L'attaque d'homogénéité	L'attaque de connaissance acquise	L'attaque d'inférence Probabiliste	
k-anonymat	*				
l-diversité	*	*	*		
(l,c)-diversité	*	*	*	*	
l-diversité d'entropie	*	*	*	*	
t-ferméture		*			
δ-Présence					*

[Tab2]
Tableau des types de techniques d'anonymisation, Source : Feten Ben Fredj (2017) Méthode et outil d'anonymisation des données sensibles, Conservatoire national des arts et métiers – CNAM : HAL.

Technique	perturbatrice	non perturbatrice
Généralisation		*
Suppression		*
Micro-agrégation	*	
« Bucketization »	*	
« Anatomy »		*
« Slicing »	*	
« Swapping »	*	
Recodage global		*
« Bottom coding »		*
« Top coding »		*
Bruit aléatoire	*	

Tableau 23. Types de techniques d'anonymisation

[Tab3]
Jeu de données issue fait par Thijs Benschop, Cathrine Machingauta, Matthew Welch (2018) Measuring Risk, Disponible à: https://sdcpractice.readthedocs.io/en/latest/measure_risk.html (Accédé: Avril 2019).

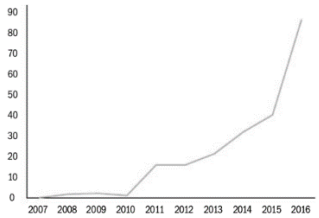
No	Residence	Gender	Education level	Labor status	Weight	f_k	F_k	risk
1	Urban	Female	Secondary incomplete	Employed	180	2	360	0.0054
2	Urban	Female	Secondary incomplete	Employed	180	2	360	0.0054
3	Urban	Female	Primary incomplete	Non-LF	215	1	215	0.0251
4	Urban	Male	Secondary complete	Employed	76	2	152	0.0126
5	Rural	Female	Secondary complete	Unemployed	186	1	186	0.0282
6	Urban	Male	Secondary complete	Employed	76	2	152	0.0126
7	Urban	Female	Primary complete	Non-LF	180	1	180	0.0290
8	Urban	Male	Post-secondary	Unemployed	215	1	215	0.0251
9	Urban	Female	Secondary incomplete	Non-LF	186	2	262	0.0074
10	Urban	Female	Secondary incomplete	Non-LF	76	2	262	0.0074

II. Graphiques

[Graph1]

Van Schalkwyk, Francois (2017) *The Social Dynamics of Open Data*, : African Minds.

Figure 1 Number of research publications on open government data indexed in the Web of Science 2007–2016 (n=216)



III. Divers

[Divers1]

Patrimoine macroscopique du groupe VYV_Source : document interne de la DDSN du groupe (Direction des Données et Stratégie Numérique

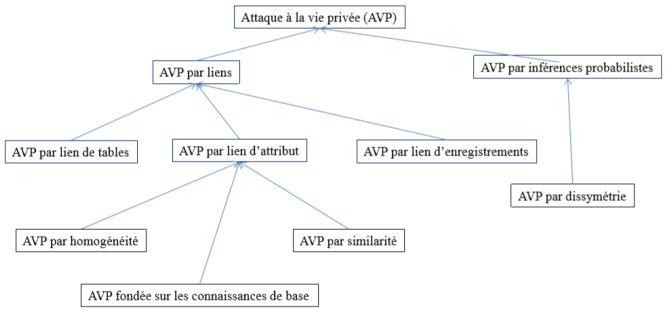
NON DISPONIBLE À LA DIFFUSION EXTERNE

[Divers2]

Patrimoine macroscopique du groupe VYV détaillé, Source : document interne de la DDSN du groupe (Direction des Données et Stratégie Numérique)

NON DISPONIBLE À LA DIFFUSION EXTERNE

[Divers3]
Taxonomie des modèles d'attaque de la vie privée (AVP), Source : Feten Ben Fredj (2017) Méthode et outil d'anonymisation des données sensibles, Conservatoire national des arts et métiers - CNAM,: HAL.



[Divers4]
Capture d'écran du logiciel AIPD fourni par la CNIL (2018), Disponible à : https://www.cnil.fr/fr/reglement-europeen-sur-la-protection-des-donnees-ce-qui-change-pour-les-professionnels (Accédé: Mai 2019).