Nº 11-522-X au catalogue
ISSN 1709-822X

Recueil du Symposium de 2022 de Statistique Canada : Désagrégation des données : dresser un portrait de données plus représentatif de la société

Risque d'attaque par reconstruction au moyen des données de recensement de Statistique Canada

par Mathew Abado et George Stefan

Date de diffusion : le 25 mars 2024

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2022 de Statistique Canada
Désagrégation des données : Dresser un portrait de données plus représentatif de la société

Risque d'attaque par reconstruction au moyen des données de recensement de Statistique Canada

Mathew Abado et George Stefan[1]

Résumé

La publication de davantage de données présentant un niveau élevé de désagrégation peut accroître la transparence et fournir des renseignements importants sur les groupes sous-représentés. L'élaboration d'options d'accès plus facilement disponibles augmente la quantité d'information disponible et produite par les chercheurs. Accroître l'étendue et la profondeur de l'information diffusée permet une meilleure représentation de la population canadienne, mais impose également à Statistique Canada une plus grande responsabilité quant à la manière dont cela est fait, qui se doit de préserver la confidentialité; il est donc utile d'élaborer des outils qui nous permettent de quantifier le risque lié à la granularité accrue des données.

Afin d'évaluer le risque d'une attaque par reconstruction de base de données sur les données de recensement publiées par Statistique Canada, nous appliquons la stratégie de l'U.S. Census Bureau, qui met en avant une méthode utilisant un solveur de satisfaisabilité booléenne (SAT) pour reconstruire les attributs individuels des résidents d'un îlot hypothétique du recensement américain, basé uniquement sur un tableau de statistiques récapitulatives. Nous prévoyons d'étendre cette technique pour tenter de reconstruire une petite fraction des microdonnées de recensement de Statistique Canada. Dans cet article, nous aborderons nos conclusions, les défis liés à l'élaboration d'une attaque par reconstruction et l'effet d'une mesure de confidentialité existante pour atténuer ces attaques. En outre, nous comparerons notre stratégie actuelle à d'autres méthodes potentielles utilisées aux fins de protection des données, en particulier la publication de données tabulaires perturbées par un mécanisme aléatoire, tel que la confidentialité différentielle.

Mots clés : confidentialité; attaque par reconstruction; confidentialité différentielle.

1. Introduction

1.1 Contexte

Dans le cadre du Plan d'action sur les données désagrégées (PADD), Statistique Canada prévoit mettre une plus grande quantité de données désagrégées à la disposition des chercheurs et du public. L'objectif est de veiller à ce que les données représentent mieux la diversité de la population canadienne, notamment pour qu'elles représentent mieux les personnes autochtones, les groupes racisés, issus de la diversité de genre et ayant un statut d'incapacité. Avant l'initiative du PADD, Statistique Canada publiait déjà une grande quantité de données à un niveau élevé de désagrégation. Certains de ces renseignements sont facilement accessibles et gratuits dans les tableaux de données publiés en ligne par Statistique Canada. De plus, Statistique Canada propose plusieurs autres options d'accès aux données aux chercheurs, y compris les centres de données de recherche, les outils d'accès à distance aux données et les initiatives de démocratisation des données.

[1]Mathew Abado, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6 (mathew.abado@gmail.com); George Stefan, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6 (george.stefan@mail.utoronto.ca)

Plus les options d'accès et le nombre de renseignements diffusés sont nombreux, plus le risque de divulgation d'attributs de personnes est grand. Cette divulgation contreviendrait à l'article 17 de la *Loi sur la statistique*, qui régit Statistique Canada et son obligation de recueillir et publier des renseignements. Pour étudier ce risque, on peut, entre autres méthodes, envisager une attaque par reconstruction de base de données. Ce type d'attaque consiste pour un intrus à créer un fichier de microdonnées où les statistiques correspondent à celles du fichier d'origine, par exemple, celles publiées en ligne, qui peut ensuite servir à la réidentification des répondants individuels et ainsi divulguer leurs attributs. Dinur et Nissim (2003) ont démontré le risque théorique de cette attaque en décrivant une procédure utilisant des requêtes sur une base de données privée. Ils ont constaté que pour prévenir ce type d'attaque, il faut considérablement perturber les données avant leur publication.

Les travaux réalisés à la suite de la publication de leur article pour traiter les répercussions découvertes ont mené à l'élaboration du cadre de Confidentialité différentielle. Dwork et Roth (2014) décrivent un « mécanisme de la confidentialité » ayant la forme d'un algorithme qui prend comme données d'entrée une base de données (p. ex. un fichier de microdonnées) pour produire des données de sortie (p. ex. une cellule de tableau à des fins de publication). Ce résultat correspond souvent à une requête comme « combien de lignes dans la base de données satisfont la propriété *P*? », où *P* pourrait être « résident de l'Ontario âgé de 15 à 19 ans », par exemple. Un contrôle de la divulgation peut avoir été appliqué au résultat de la requête sous forme de suppression ou de perturbation avant la sortie. La caractéristique déterminante d'un mécanisme de confidentialité différentielle est qu'il doit produire des données de sortie semblables à partir de bases de données d'entrée semblables. Cela garantit la confidentialité en ce sens que la sortie ne pourra pas révéler l'information d'un seul contributeur, puisque la sortie doit être similaire indépendamment de la réponse du contributeur. Cet aspect sera abordé plus en détail dans la section 3.2.

2. Données publiques

2.1 Tableaux publiés

Statistique Canada publie en ligne des centaines de tableaux de données fondés sur les données de chaque cycle quinquennal du Recensement de la population canadienne. Statistique Canada publie par ailleurs des tableaux sur une vaste gamme de sujets provenant de nombreuses sources autres que le recensement, et la méthodologie présentée dans notre article peut être étendue aux fins d'analyse de ces autres publications. Néanmoins, la présente étude portera exclusivement sur les données du recensement. Ces tableaux sont souvent des tableaux croisés de diverses caractéristiques sociales et sont disponibles à différents niveaux de détail et niveaux géographiques, jusqu'à des niveaux très fins. Cela représente la publication de centaines de millions de cellules de tableau. Des méthodes de contrôle de la divulgation statistique sont appliquées à ces tableaux avant leur publication afin de protéger la confidentialité des renseignements fournis par les répondants, conformément à l'article 17 de la *Loi sur la statistique*.

Les méthodes de contrôle de la divulgation comprennent l'arrondissement et la suppression de cellules; elles seront traitées plus en détail à la section 3 de l'article. L'agrégation de tableaux à des niveaux géographiques plus élevés ou à des variables moins détaillées fait également partie de la méthodologie de contrôle de la divulgation de Statistique Canada. Toutefois, si l'on va trop loin, cette méthode limiterait la capacité des données à représenter les expériences de groupes de population plus petits que le PADD vise justement à mieux représenter.

Compte tenu du nombre croissant de renseignements diffusés dans le cadre du PADD, il est important d'évaluer continuellement l'efficacité des méthodes de contrôle de la divulgation appliquées à ces tableaux. À titre d'exemple, il est important d'évaluer la plausibilité qu'un intrus combine l'information disponible dans ces tableaux afin de reconstruire les microdonnées confidentielles sous-jacentes et réidentifie les répondants individuels.

2.2 Augmentation des options d'accès

En plus des tableaux de données en ligne, plusieurs autres options sont proposées aux chercheurs souhaitant analyser les données du recensement et les données provenant d'autres sources. Les centres de données de recherche (CDR) sont des environnements sécurisés dans lesquels les chercheurs, qui ont demandé et obtenu une autorisation de sécurité, et qui sont donc dignes de confiance en tant que personnes réputées être employées de Statistique Canada, peuvent accéder à une grande variété de données pour leur projet. Les résultats de leurs analyses ne peuvent être publiés hors de l'environnement du CDR qu'après vérification de la confidentialité. L'accès à distance en temps réel (ADTR) est un outil de recherche en ligne qui est moins rigoureux en matière d'accès que les CDR, mais qui ne permet pas aux utilisateurs d'interagir directement avec les microdonnées. Les utilisateurs du système d'ADTR peuvent soumettre des requêtes et recevoir des résultats sous forme de tableaux sur lesquels les règles de confidentialité ont été appliquées. L'Initiative de démocratisation des données (IDD) permet aux établissements postsecondaires d'avoir accès à des microdonnées à grande diffusion pour en faciliter l'utilisation par les étudiants et les professeurs.

Ces options d'accès supplémentaires augmentent la quantité d'information accessible à un intrus potentiel, par exemple, étant donné les résultats publiés par les chercheurs à partir de ces données. Cela augmente alors le risque de divulgation de renseignements confidentiels. Bien que chacune de ces options d'accès fasse l'objet de mesures indépendantes de contrôle de la divulgation, ces mesures ne fournissent pas de garantie mathématique stricte contre la divulgation, surtout si l'on tient compte du risque global de combinaison de tous les renseignements. De plus, le nombre de renseignements diffusés étant de plus en plus grand, il devient de plus en plus difficile d'évaluer les risques de divulgation.

2.3 Sources de données externes

Le Recensement de la population et les publications qui en découlent ne sont pas les seules sources d'information sur la population canadienne. Un intrus peut avoir accès à des renseignements supplémentaires sur une collectivité ou une personne d'intérêt en particulier, susceptibles de l'aider dans une attaque par reconstruction. Ils peuvent, par exemple, connaître le sexe et l'âge de leurs voisins dans une petite collectivité et peuvent faire des estimations éclairées au sujet de l'échelle de leur salaire ou d'autres caractéristiques sociales. Un intrus cherchant à commettre ce type d'attaque peut également avoir accès à des bases de données commerciales ou administratives qui pourraient être recoupées avec son ensemble de microdonnées reconstruites dans le but de réidentifier des personnes.

3. Méthodes de confidentialité

3.1 Arrondissement et suppression

Les méthodes classiques de contrôle de la divulgation statistique comportent souvent une forme de perturbation. Une méthode possible est l'arrondissement classique, qui consiste à arrondir chaque cellule à la valeur la plus proche de base b la plus proche (peut-être $b = 5$). Par exemple, si le nombre réel de personnes parlant une langue maternelle donnée dans une subdivision de recensement (SDR) est de 3, ce chiffre serait arrondi à 5. Ce mécanisme est toutefois biaisé, car la valeur de l'espérance du résultat arrondi n'est pas nécessairement égale à la vraie valeur; par exemple, $E[arrondi_5(3)] = E[5] = 5 \neq 3$. Une solution de rechange consiste à effectuer un arrondissement aléatoire sans biais, c'est-à-dire à effectuer un arrondi stochastique sur la valeur de base supérieure ou inférieure à la valeur réelle, la probabilité étant inversement proportionnelle à la distance entre la vraie valeur et la valeur arrondie. Dans l'exemple ci-dessus, une valeur de cellule de 3 serait arrondie à 5 avec une probabilité de 0,6 et arrondie à 0 avec une probabilité de 0,4. Si le mécanisme d'arrondissement est connu ou peut être déduit par un intrus, cela impose des contraintes claires aux microdonnées sous-jacentes utilisables dans une attaque par reconstruction.

La suppression de cellules est une autre méthode courante de contrôle de la divulgation. On peut en effet supprimer les cellules de tableau jugées comme se prêtant trop à la divulgation (p. ex. les petits nombres) avant la publication. On peut aussi appliquer une suppression complémentaire afin d'éviter que la valeur de la cellule supprimée ne soit reconstruite au moyen du reste du tableau. L'intrus se retrouve ainsi avec un niveau d'incertitude quant à la valeur exacte de la cellule sensible et potentiellement de ses cellules complémentaires. Même connu par l'attaquant, le mécanisme de suppression peut encore constituer une contrainte utilisée dans le processus de reconstruction, bien qu'il

ne soit pas particulièrement utile. Dans notre projet, la suppression n'a pas été étudiée, mais la méthodologie pourrait être élargie pour l'intégrer également.

Le présent article vise à déterminer dans quelle mesure les mécanismes d'arrondissement décrits ci-dessus protègent les données en cas d'attaque. La section 4 aborde plusieurs scénarios d'attaque possibles.

3.2 Confidentialité différentielle

Développé en réponse au risque théorique d'une attaque par reconstruction de base de données, risque démontré par Dinur et Nissim (2003), le champ de la Confidentialité différentielle tente de fournir une garantie mathématique plus solide de la confidentialité. Si nous examinons un mécanisme dont les données d'entrée sont des microdonnées confidentielles et qui renvoie une valeur aux fins de publication, nous pourrions considérer ce mécanisme comme confidentiel puisque sa sortie ne peut pas changer de façon importante, car l'ensemble de données d'entrée varie graduellement. Les résultats d'un tel mécanisme seraient, par exemple, semblables qu'un contributeur donné soit présent ou non dans le fichier de microdonnées, et il est donc peu probable que les résultats divulguent ses contributions. La mesure dans laquelle la sortie du mécanisme peut varier entre les bases de données d'entrée adjacentes peut alors servir de mesure de sa capacité à protéger les données confidentielles.

Mathématiquement, un mécanisme randomisé M est de confidentialité différentielle ε si pour tous les ensembles de sortie possibles S dans la plage de M et pour toutes les bases de données d'entrée $x,\ y$, telles que x et y diffèrent dans au plus un élément, on obtient $P(M(x) \in S) \leq e^{\varepsilon} P(M(y) \in S)$ (Dwork et Roth, 2014, p. 17). En gros, tout résultat possible du mécanisme doit être approximativement aussi probable, qu'une personne donnée soit présente (disons, dans la base de données *x*) ou absente (disons, dans la base de données *y*). La valeur de ε peut être considérée comme un « budget » de perte de confidentialité. Un ε plus grand permet généralement de diffuser plus de données avec moins de perturbations; la diffusion des données peut être considérée comme une « dépense » du budget de confidentialité, et les données de meilleure qualité sont plus coûteuses.

Un exemple d'algorithme de confidentialité différentielle est le mécanisme de Laplace, qui calcule les vrais résultats d'un ensemble de requêtes de comptage disjoint (p. ex. un histogramme) et perturbe chaque coordonnée avec le bruit tiré de la loi de Laplace avec un paramètre d'échelle $1/\varepsilon$. On peut montrer que ce mécanisme préserve la confidentialité-ε différentielle. Comme pour la stratégie d'arrondissement décrite ci-dessus, l'intrus se retrouve devant un niveau d'incertitude quant à la valeur précise de la cellule.

4. Risque de reconstruction de base de données

4.1 Données

Afin d'évaluer la probabilité d'une attaque par reconstruction de base de données réussie sur les tableaux de données publiés par Statistique Canada, des simulations d'attaques ont été lancées. Nous avons généré des ensembles de microdonnées synthétiques représentant de petites SDR, ayant chacune une population de 50 personnes. Chaque personne simulée avait des attributs d'âge (en années individuelles), de genre (2 catégories) et d'état matrimonial (6 catégories : *Marié(e)*, *Vivant en union libre*, *Jamais marié(e)*, *Séparé*(*e*), *Divorcé(e)*, *Veuf(-ve)*). Ces attributs ont été générés au hasard selon leur distribution conjointe dans la population canadienne réelle. Nous avons choisi d'effectuer les attaques sur ces ensembles de données synthétiques plutôt que sur des données réelles afin de faciliter les essais répétés.

Chaque ensemble de microdonnées a été agrégé pour créer quatre tableaux de contingence, de façon à imiter les tableaux disponibles en ligne au niveau de la SDR pour ces attributs. Le profil du recensement est un grand tableau à données en tableaux empilés dans lequel les chiffres de population sont ventilés par genre ainsi que par de nombreuses caractéristiques différentes. Aux fins de la présente étude, les sections du profil du recensement concernant l'âge et l'état matrimonial ont été recréées; il s'agit en fait d'un tableau de contingence où le genre et l'âge (dans les classes de cinq ans) sont croisés, et d'un autre où le genre et l'état matrimonial sont croisés. En plus du profil du recensement, nous avons créé deux autres tableaux. L'un d'eux contenait le genre croisé avec l'âge (en années individuelles) et

l'autre le genre, l'âge (en classes de cinq ans) et l'état matrimonial. Dans le cas de ce dernier tableau, seuls les six états matrimoniaux énumérés ci-dessus ont été utilisés, même si le tableau en ligne comporte des catégories plus fines.

Nous avons appliqué l'arrondissement aléatoire sans biais de base 5 comme mesure de protection de la confidentialité, mais nous avons aussi généré les versions non arrondies, c.-à-d. non protégées, de ces tableaux afin de les comparer ultérieurement.

Figure 4.1-1
Microdonnées synthétiques agrégées et arrondies pour créer le profil du recensement simulé (section Âge)

	Age	Gender	Marital Status
0	0	Men +	N/A
1	7	Men +	N/A
2	8	Men +	N/A
3	8	Women +	N/A
4	11	Women +	N/A
...	...	...	...
45	75	Men +	Never married
46	75	Women +	Married
47	76	Women +	Married
48	77	Men +	Married
49	78	Women +	Married

Census Profile (Age)

Sex / Age	Total	Men +	Women +
Total	50	25	25
0 to 14 years	5	5	5
0 to 4 years	0	0	0
5 to 9 years	5	0	0
10 to 14 years	0	0	0
15 to 64 years	35	15	20
...	...	...	...
95 to 99 years	0	0	0
100 years and over	0	0	0

4.2 Encodage des contraintes

Un problème de satisfaction de contraintes (CSP pour *constraint satisfaction problem*) est un problème formulé comme un ensemble de variables dont les états sont liés par un ensemble de contraintes mathématiques et logiques, dans lequel l'objectif est de trouver une affectation des variables qui est conforme à ces contraintes (c.-à-d. qui y *satisfait)*. Dans le cas d'une reconstruction de base de données, chaque cellule d'un tableau de contingence représente une contrainte sur les variables représentant des attributs personnels dans l'ensemble de microdonnées sous-jacent, et le problème peut être formulé comme un problème de satisfaction de contraintes de cette façon. Conformément à la procédure décrite par Garfinkel et coll. (2019), ces contraintes ont été codées au format CSP pour être utilisées dans le logiciel solveur de contraintes *Sugar* (Tamura et coll., 2009). L'annexe A en donne un exemple.

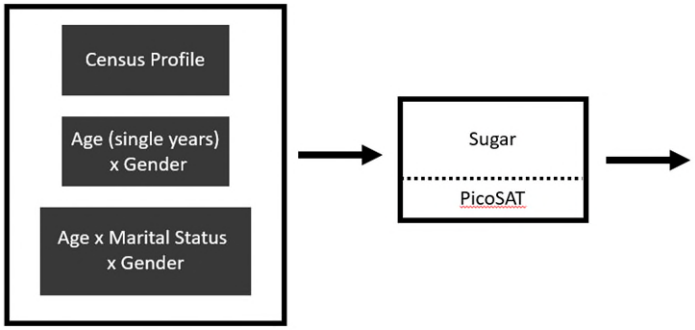
Chaque personne sous-jacente dans l'ensemble de microdonnées a une variable définie pour chacun de ses attributs inconnus, dans notre cas l'âge, le genre et l'état matrimonial. Une personne contribue par 1 à la valeur (non arrondie) d'une cellule en particulier si elle a les attributs représentés par la cellule et 0 autrement, et le décompte peut donc être représenté comme la somme de ces fonctions indicatrices. Ainsi, la valeur de chaque cellule peut être représentée par une équation plutôt longue reliant, et donc contraignant, les variables inconnues individuelles.

L'arrondissement et l'incertitude qui en résulte dans les nombres de chaque cellule comme dans la population totale causaient des difficultés supplémentaires. Nous avons traité l'incertitude dans la population totale en incluant des variables pour le nombre maximal possible de personnes sous-jacentes compte tenu de la population totale déclarée (arrondie). Par exemple, si la population déclarée était de 50 individus, on définit des variables pour 54 personnes possibles. Une variable binaire supplémentaire a ensuite été incluse pour chaque personne potentielle afin d'indiquer si elle existe vraiment, et ces variables d'existence ont également été résolues. Le solveur *Sugar* a ensuite été en mesure de décider, dans le cadre de la solution, combien de ces personnes potentielles il fallait exclure en « désactivant » leur variable d'existence. Nous avons traité les nombres de cellules présentant une incertitude en déterminant les bornes implicites par les valeurs arrondies et en les codant comme des inégalités plutôt que des égalités (ce qui aurait constitué les contraintes impliquées par les nombres de cellules connus).

Une fois ces contraintes codées, *Sugar* a pu générer une solution en convertissant le problème en un problème de satisfaisabilité booléenne (SAT) et en le résolvant au moyen d'un solveur SAT, dans ce cas *picoSAT* (Biere, 2008).

La solution en résultant a attribué une valeur à chacune des variables inconnues, c'est-à-dire un ensemble de microdonnées qui satisfait à toutes les contraintes données et qui aurait donc pu générer de façon plausible les tableaux donnés. Il s'agit de la base de données reconstruite. L'annexe B donne un exemple de la sortie du solveur *Sugar*.

Figure 4.2-1
Représentation schématique de la reconstruction de la base de données correspondant aux microdonnées de la figure 4.1-1. On constate que le chiffre de population total est erroné en raison de la population totale arrondie dans la publication.



	Age	Gender	Marital Status
0	2	Women +	N/A
1	4	Men +	N/A
2	5	Women +	N/A
3	9	Men +	N/A
4	12	Men +	N/A
...	...	...	...
50	123	Women +	Widowed
51	124	Women +	Widowed
52	124	Women +	Widowed
53	125	Women +	Widowed

4.3 Taux d'appariement dans les bases de données reconstruites

La mesure la plus intuitive de la réussite d'une reconstruction de base de données et, par extension, du risque posé par ce type d'attaques, est peut-être la fraction de vrais enregistrements de microdonnées qui ont été répliqués avec succès dans la base de données reconstruite. Il s'agit du *taux d'appariement*. Il faut garder à l'esprit que les taux d'appariement dépendent fortement des circonstances particulières, comme la taille de la population, le nombre d'attributs présents et les critères précis utilisés pour confirmer un appariement.

La procédure a été effectuée sur 30 ensembles de microdonnées synthétiques, comprenant chacun 50 personnes générées indépendamment de la distribution, comme le décrit la section 4.1. En moyenne, les bases de données reconstruites correspondantes contenaient des appariements exacts pour 27 % (avec un écart-type de $\sigma = 8$ %) de la population réelle. Si les critères d'appariement sont assouplis pour permettre un écart de l'âge reconstruit allant jusqu'à 2 ans, le taux d'appariement augmente à 62 % ($\sigma = 9$ %).

Il est utile de comparer ces taux d'appariement aux taux d'appariement par paires entre les ensembles de microdonnées sous-jacents identiquement distribués, mais indépendants. On peut les considérer comme le taux d'appariement accessoire entre les populations de deux collectivités différentes, en utilisant ces critères d'appariement. Elles donnent également une idée de ce qu'un intrus pourrait reconstruire s'il dispose uniquement des statistiques de la population – les renseignements que Statistique Canada a vraiment l'intention de publier – sans information supplémentaire. On obtient 870 paires entre les 30 ensembles de microdonnées synthétiques, et le taux d'appariement dans ces ensembles était de 15 % ($\sigma = 5$ %). Avec l'assouplissement du critère de correspondance selon l'âge pour tenir compte d'une différence de 2 ans, le taux atteint 50 % ($\sigma = 7$ %). Le fait que ces taux ne soient pas beaucoup plus bas que les taux d'appariement reconstruits laisse entendre que bon nombre des appariements reconstruits pourraient être accidentels.

Par ailleurs, quand on a plutôt réalisé la procédure de reconstruction sur les tableaux non arrondis, c'est-à-dire sans mesures de contrôle de la divulgation, les bases de données reconstruites contenaient des appariements exacts pour environ 90 % de la population sous-jacente. Ce taux beaucoup plus élevé que les taux d'appariement obtenus au moyen des tableaux arrondis indiquerait que l'arrondi aléatoire réussit dans une certaine mesure à entraver une attaque par reconstruction.

4.4 Un intrus très bien informé

Un intrus pourrait, aujourd'hui ou demain, avoir accès à plus de renseignements que ceux publiés en ligne par Statistique Canada. Ces renseignements supplémentaires pourraient, par exemple, provenir d'une future désagrégation de données, de publications découlant des autres options d'accès aux données proposées par Statistique Canada (CDR, ADTR, etc.), de la connaissance d'une collectivité donnée, ou de bases de données commerciales ou administratives. Ces renseignements peuvent être intégrés à une attaque par reconstruction de base de données. Il est difficile de savoir de façon générale la quantité de renseignements supplémentaires à la disposition d'un intrus, mais nous pouvons modéliser le cas le plus extrême pour obtenir des bornes de risque.

Dans le cas le plus extrême, un intrus a toutes les valeurs de l'ensemble de microdonnées, sauf une : par exemple, il lui manquerait seulement l'âge d'une personne. Il serait instructif de déterminer si ce dernier élément d'information est reconstructible au moyen des tableaux publiés. Cela permettrait de borner le risque de réussite d'une reconstruction complète. En effet, s'il existe plusieurs solutions possibles pour le point de données final, même avec toutes les contraintes supplémentaires imposées par les renseignements supplémentaires, alors cette ambiguïté dans la solution doit aussi être présente dans des scénarios d'attaque plus raisonnables, et la véritable solution ne peut pas être déterminée de façon unique.

De plus, il est intéressant de considérer ce cas extrême du point de vue de la Confidentialité différentielle. Étant donné que la garantie mathématique fournie par le cadre garantit expressément que la sortie sera insensible à toute valeur individuelle dans l'ensemble de microdonnées, un intrus ne sera pas en mesure de reconstruire la dernière valeur inconnue, même dans ce cas extrême. Toute sortie vue par l'intrus peut vraisemblablement avoir été générée par plusieurs valeurs d'entrée différentes possibles, et la vraie valeur d'entrée ne peut pas être choisie de façon fiable parmi celles-ci.

Les renseignements supplémentaires correspondant à la connaissance de toutes les autres cellules de microdonnées peuvent facilement être codés comme des contraintes supplémentaires au format CSP pour le solveur *Sugar*. Chaque valeur connue est codée en tant que contrainte supplémentaire et annexée à la fin des contraintes générées à partir des tableaux.

Même dans ce cas, la reconstruction du point de données final n'est pas garantie. Si la seule variable manquante est un âge, elle peut être déterminée avec précision environ la moitié du temps. S'il s'agit d'un genre, il peut être déterminé dans la plupart des cas (~75 %), ce qui est plus élevé que le taux de référence de ~50 % obtenu par supposition. S'il s'agit de l'état matrimonial, il peut être déterminé dans une minorité de cas (~25 %), ce qui est moins efficace qu'une stratégie consistant à toujours supposer que la personne en question est mariée. À titre de comparaison, si les tableaux n'étaient pas arrondis, ce scénario d'attaque extrême réussirait infailliblement.

5. Discussion

Notre étude visait à évaluer l'efficacité des méthodes classiques de contrôle de la divulgation, en particulier l'arrondissement, dans le contexte d'initiatives comme le Plan d'action sur les données désagrégées de Statistique Canada, qui vise à rendre encore plus de données désagrégées accessibles aux chercheurs et au public. Nous avons évalué l'efficacité de ces méthodes dans la protection des données en nous intéressant aux attaques par reconstruction de bases de données, dans lesquelles un intrus utilise des données publiées, parfois complétées par d'autres sources d'information, pour tenter de reconstruire les microdonnées confidentielles sous-jacentes.

Nous avons montré qu'un intrus utilisant des tableaux de données non arrondies dans une attaque par reconstruction obtient un taux de réussite élevé, mesuré par la fraction de vrais enregistrements de microdonnées répliqués dans sa base de données reconstruite (le *taux d'appariement*). Après l'application de l'arrondissement, ce taux d'appariement a été considérablement réduit, pour se situer presqu'aussi bas que celui auquel on pourrait s'attendre en n'utilisant que les statistiques de la population (c.-à-d. les paramètres du processus générant les populations de l'échantillon, qui sont les objets d'étude prévus dans ces tableaux). Même dans un scénario d'attaque extrême où l'on imagine que l'intrus a accès à toutes les cellules du tableau de microdonnées sauf une, l'arrondissement effectué sur les tableaux publiés empêche la réussite de la reconstruction du point de données final dans de nombreux cas. Ces résultats sont encourageants, mais leur interprétation nécessite une certaine subtilité.

Premièrement se pose la question de la mesure du risque : le cadre de risque de reconstruction de base de données présenté ici traite-t-il de tous les types de risque de divulgation, et comment le taux d'appariement doit-il être interprété comme une mesure de ce risque? Le cadre de reconstruction de base de données est très souple, et de nombreux types d'attaques peuvent être formulés comme des problèmes de satisfaction de contraintes de la même manière, mais pas tous les types. Nous ne savons pas, par exemple, comment modéliser la reconnaissance spontanée de cette façon.
Le taux d'appariement est une mesure intuitive du risque dans la reconstruction de base de données, mais il peut être très sensible aux paramètres de l'attaque. Ainsi, les reconstructions de grandes populations avec des attributs moins nombreux et moins détaillés sont susceptibles d'avoir un grand nombre d'appariements accessoires. Il est donc important d'avoir une référence pour la comparaison. L'interprétation du taux d'appariement est subtile elle aussi : à partir de quel point est-il trop élevé? Un intrus qui sait que son attaque peut reconstituer 100 % de la population réelle constitue un problème, mais s'il est capable de reconstituer seulement, disons, 25 % de la population, il ne saura peut-être toujours pas *quels* enregistrements reconstruits sont exacts. Cela pourrait tout de même être considéré comme un risque de divulgation, mais des mesures supplémentaires de ce risque seraient nécessaires pour brosser un tableau complet de la situation. Malgré ces limites et subtilités, le taux d'appariement pourrait jouer un rôle important dans un protocole plus large d'évaluation des risques de divulgation.

Ajoutons que les simulations présentées dans l'article n'étaient pas des reconstructions à grande échelle de la base de données de la population canadienne. La portée était limitée pour ce qui est de la taille de la population et de la géographie, du nombre d'attributs utilisés et des tableaux utilisés aux fins des contraintes. Il est possible qu'un intrus ayant suffisamment d'expertise, de motivation et de puissance de calcul monte une attaque à plus grande échelle, et que le plus grand nombre de contraintes calculées à partir des tableaux supplémentaires, des variables et des géographies emboîtées donne des taux d'appariement plus élevés. Il sera crucial de garder à l'esprit cette possibilité si des tableaux comportant plus de variables et concernant plus de niveaux géographiques sont mis à disposition du public en plus grand nombre, par l'intermédiaire d'initiatives comme le PADD, par exemple.

Il serait tout aussi intéressant de se pencher sur d'autres mécanismes de contrôle de la divulgation. Les mécanismes de confidentialité différentielle sont particulièrement bien adaptés à la prévention des attaques par reconstruction de bases de données. Ainsi, on pourrait appliquer le mécanisme de Laplace aux tableaux publiés et déterminer le degré auquel ce mécanisme protège les données contre la reconstruction de base de données pour le comparer au mécanisme d'arrondissement. L'algorithme utilisé dans les attaques par reconstruction étudiées dans l'article repose sur des bornes finies pour la perturbation additive de chaque cellule, mais la fonction de densité de probabilité de Laplace est non bornée. Un intrus pourrait choisir des bornes artificielles; cependant, il serait difficile pour lui de les choisir suffisamment larges pour capturer le bruit, mais suffisamment petites pour être informatives. Il serait utile d'étudier le taux d'appariement d'une tentative de reconstruction dans ces conditions. Cependant, la garantie mathématique fournie par le mécanisme de confidentialité différentielle empêche un intrus d'être totalement sûr de toute reconstruction : si une base de données reconstruite est une solution plausible, alors toute base de données similaire doit aussi être plausible, et donc toute ligne particulière (c.-à-d. une personne) peut être ou ne pas être un vrai appariement. Le degré de protection offert par d'autres mécanismes de contrôle de la divulgation, qu'ils soient de confidentialité différentielle ou autres, doit ensuite être évalué en fonction de préoccupations relatives à la qualité, par exemple la capacité des données à représenter les besoins et la diversité de la population canadienne.

Bibliographie

Biere, A. (2008), « PicoSAT essentials », *Journal on Satisfiability, Boolean Modeling and Computation*, 4(2-4), p. 75-97.

Dinur, I. et K. Nissim. (2003, juin), « Revealing information while preserving privacy » dans Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, p. 202-210.

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

Garfinkel, S., J.M. Abowd et C. Martindale. (2019), « Understanding database reconstruction attacks on public data », *Communications of the ACM*, 62(3), p. 46-53.

Loi sur la statistique, L.R.C. (1985), ch. S-49, s 17. https://laws-lois.justice.gc.ca/fra/lois/s-19/

Statistique Canada. 2017, *London, SR (subdivision de recensement), Ontario et Canada (pays)* (tableau). *Profil du recensement*. Recensement de 2016, produit nº 98-316-X2016001 au catalogue de Statistique Canada. Ottawa. Diffusé le 29 novembre 2017. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=F (Site consulté le 22 novembre 2022).

Statistique Canada. 2017, *London, SR (subdivision de recensement), Ontario et Canada (pays)* (tableau). *Âge (en années) et âge moyen (127) et sexe (3) pour la population du Canada, provinces et territoires, divisions de recensement et subdivisions de recensement*. Recensement de 2016, produit nº 98-400-X2016004 au catalogue de Statistique Canada. Ottawa. Diffusé le 3 mai 2017. https://www150.statcan.gc.ca/n1/fr/catalogue/98-400-X2016004 (Site consulté le 22 novembre 2022).

Statistique Canada. 2017, *London, SR (subdivision de recensement), Ontario et Canada (pays)* (tableau). *État matrimonial (13), âge (16) et sexe (3) pour la population âgée de 15 ans et plus du Canada, provinces et territoires, divisions de recensement, subdivisions de recensement et aires de diffusion*. Recensement de 2016, produit nº 98-400-X2016034 au catalogue de Statistique Canada. Ottawa. Diffusé le 2 août 2017. https://www150.statcan.gc.ca/n1/fr/catalogue/98-400-X2016034 (Site consulté le 22 novembre 2022)

Tamura, N., A. Taga, S. Kitagawa et M. Banbara. (2009), « Compiling finite linear CSP into SAT » *Constraints*, 14(2), p. 254-272.

Remerciements

Les auteurs tiennent à remercier leurs collègues pour leur précieuse contribution à la réussite de ce projet. Nous remercions en particulier Steven Thomas pour son soutien et ses conseils à chaque étape du projet, ainsi que pour son examen des diapositives et du manuscrit. Nous remercions Mark Stinner de nous avoir fait profiter de son expertise dans des discussions très éclairantes et nous voir aidé à poser les bonnes questions. Nous remercions également Héloïse Gauvin pour son aide dans la révision et la traduction des diapositives. Enfin, nous remercions Tyler Kirkland et Peter Wright d'avoir examiné le manuscrit et proposé des améliorations.

Annexe A Exemple de code *Sugar*

```
; Extrait du code Sugar montrant la définition de la variable et
;l'encodage :des contraintes.
; Compte tenu d'une population totale déclarée de 50 personnes, définir des
variables pour 54 personnes.
; ...
; Définir des variables à résoudre pour la personne potentielle nº 53.
(int AGE053 0 125)      ; âge inconnu, nombre entier entre 0 et 125.
(int GENDER053 0 1)     ; Genre inconnu à deux catégories, codé 0 ou 1.
(int MARST053 0 5)      ; État matrimonial inconnu à six catégories, codé
                        ; de 0 à 5.
(int EXISTS053 0 1)     ; On ne sait pas si la personne nº 53 existe ou bien
; s'il y moins de personnes. Coder 1    ; pour l'existence, 0 sinon ; Exemple
de codage d'une cellule indiquant 5 personnes âgées de 30 à 34 ans.
; Supposons qu'un arrondissement aléatoire sans biais est appliqué, alors 5
```

```
; représente de 1 à 9 personnes.
; Insister pour que le nombre de cellules calculé dans le bloc intérieur soit
; >= 1.
( >=
  ; Additionner les variables indicatrices pour obtenir le nombre de
  ; cellules.
  ( +
    ; Si la personne nº 0 existe et est dans la fourchette d'âge, ajouter 1,
;sinon ajouter 0.
    (si (et (= EXISTS000 1) (>= AGE000 30) (<= AGE000 34)) 1 0)
    ; ... (variables indicatrices semblables pour chaque personne)
    (si (et (= EXISTS053 1) (>= AGE053 30) (<= AGE053 34)) 1 0)
  ]
  1 ; Insister sur le nombre de cellules >= 1.
)
; Insister pour que le nombre de cellules soit <= 9 aussi.
( <=
  ( +
    (si (et (= EXISTS000 1) (>= AGE000 30) (<= AGE000 34)) 1 0)
    ; ... (variables indicatrices semblables pour chaque personne)
    (si (et (= EXISTS053 1) (>= AGE053 30) (<= AGE053 34)) 1 0)
  )
  9 ; Insister sur le nombre de cellules <= 9.
)
```

Annexe B : Exemple de sortie de *Sugar*

```
; extrait de sortie de Sugar avec commentaires ajoutés.
s SATISFAISANT    ; il existe au moins une solution, c.-à-d. Une attribution
                  ; de toutes les variables conformes aux contraintes
                  ; données.
                  ; Sugar choisit l'une de ces solutions pour les données de
                  ; sortie.
a AGE000    2     ; la personne nº 0 est reconstruite pour être âgée
                  ; de 2 ans.
a GENDER000 0     ; la personne nº 0 est reconstruite avec le genre codé 0.
a MARST000  5     ; la personne nº 0 est reconstruite avec l'état matrimonial
; 5.
a EXISTS000 1     ; la personne n° 0 est reconstruite pour exister
;réellement.
; ...
a EXISTS053 0     ; la personne n° 53 est reconstruite pour ne PAS exister,
                  ; c'est-à-dire que l'ensemble de données reconstruit
;                 ; compte moins de 54 personnes.
```