Projet Cassiopée n°53

Evaluation de bases de données anonymisées quant au risque sur la vie privée à l'aide de sources OSINT (Open Source Intelligence)

Rédigé par

Maxence DEBES

Vadim HEMZELLEC-DAVIDSON

Encadré par

Maryline LAURENT

Louis Philippe SONDECK

Avant-propos

Ce rapport a pour objectif de présenter en détail les travaux que nous avons menés dans le cadre du projet Cassiopée, lequel se focalise sur deux axes complémentaires : l'anonymisation des données et l'exploitation de l'OSINT (Open Source Intelligence).

Sous la direction de Maryline LAURENT et Louis Philippe SONDECK, tous deux reconnus pour leur expertise pointue dans le domaine de l'anonymisation, nous avons entrepris, pour la première fois, une investigation rigoureuse sur ces thématiques. Il nous semble utile de souligner que l'approche OSINT reste pour nous un terrain de découverte, tant dans sa méthodologie que dans les outils qu'elle mobilise.

Menée de janvier à juin 2025, cette étude s'inscrit avant tout dans une démarche de recherche exploratoire et prospective. En effet, l'OSINT, en perpétuelle évolution, bénéficie de progrès technologiques constants : de nouveaux logiciels d'agrégation de sources, d'intelligence artificielle pour l'analyse de contenu et de plateformes collaboratives émergent régulièrement. Ainsi, l'un des enjeux majeurs de notre travail a été de recenser ces évolutions et d'évaluer leur apport potentiel, tant pour étendre notre champ d'analyse que pour enrichir notre réflexion.

Remerciements

Nous tenons à remercier toutes les personnes qui ont contribué de près ou de loin à l'élaboration de ce rapport.

Tout d'abord, nous aimerions remercier Maryline LAURENT et Louis Philippe SONDECK, nos tuteurs de projet. Leur encadrement pédagogique, leur temps, ainsi que leurs connaissances nous on été d'une immense utilité dans l'élaboration de ce rapport.

Nous tenons également à remercier l'ensemble de l'équipe enseignante de Télécom SudParis pour la qualité de leurs apprentissages qui joue un rôle essentiel dans notre développement académique et professionnel.

Enfin, nous remercions tout particulièrement les organisateurs du projet Cassiopée pour leur engagement et leur coordination sans faille : leur implication a créé un cadre de travail stimulant et propice à la recherche, dans lequel nous avons pu mener à bien nos investigations avec rigueur.

Table des matières

1 Introduction

L'anonymisation de bases de données est un sujet particulièrement chaud pour les entreprises car c'est le seul moyen pour elle d'échapper aux contraintes réglementaires du RGPD, les données anonymisées n'étant alors plus considérées comme des données personnelles mais comme des données anonymisées. Cependant, l'anonymisation nécessite une expertise particulière notamment pour évaluer les critères de réidentification réglementaires que sont : l'individualisation, la corrélation et l'inférence. En effet, comme le précise la CNIL, il ne suffit pas de supprimer les identifiants directs (ex : nom, prénom, numéro de téléphone...) pour rendre les données anonymes, il faut s'assurer que le risque de réidentification avec des moyens raisonnables est nul.

L'objectif du projet est de faire un tour d'horizon sur les outils d'anonymisation existants, de comprendre les différents types d'attaques qui visent les bases de données, d'expérimenter et de découvrir certains outils OSINT pour tester son efficacité potentielle en terme de réidentification et de se confronter aux limites de cet outil. La question de la potentielle généralisation de chemins d'attaques basés sur l'OSINT sera également abordée. Ainsi, l'approche à définir se veut plus réaliste, et vise à préciser certains critères d'évaluation des risques qui pourront améliorer la pertinence de l'outil d'évaluation existant. Nous espérons arriver à l'issue de l'étude à un résultat concernant la possibilité et la faisabilité de ré-identification de personnes présentant des attributs sensibles dans des bases de données anonymisées à l'aide de sources OSINT, pour justement identifier si c'est un moyen "raisonnable" qu'un potentiel assaillant pourrait utiliser.

Il est important de préciser que notre cadre d'étude se limite aux bases de données, et ne concerne pas la réidentification à l'aide de l'unicité des traces de mobilité humaine ou toute autre méthode de réidentification d'individus.

2

Principes d'anonymisation d'une base de données

Sommaire

2.1 Anonymisation ou pseudonymisation ?

2.1.1 Enjeux

Dans le contexte acutel où l'utilisation des bases de données devient de plus en plus fréquente, notamment pour l'analyse de celles-ci, et avec des métiers qui s'appuient de plus en plus sur le big data et l'open source, il devient indispensable de s'assurer que les informations présentes dans ces bases de données sont correctement protégées dès lors qu'elles concernent des informations sensibles sur des individus.

L'exemple le plus marquant que l'on peut citer est celui lié aux données de santé : si un laboratoire ou un organisme de santé détecte à un individu une maladie grave, cela peut avoir de sérieux impacts sur sa vie privée, au-delà des conséquences évidentes sur sa santé. En effet, depuis le 1er juin 2022 par exemple, pour les personnes ayant été atteintes d'un cancer ou d'une hépatite C, le délai de droit à l'oubli, qui permet à de potentiels emprunteurs de ne pas fournir d'informations relatives à leur état de santé et de ne pas avoir à fournir d'examen supplémentaire sous certaines conditions, est de cinq ans après la fin du protocole thérapeutique, en l'absence de rechute [1].

Si le RGPD ne comporte pas d'obligation générale d'"anonymisation" des données, la CNIL fait le point sur les techniques utilisables et sur leurs enjeux, et est en capacité d'exercer des sanctions sur

des sociétés ou organismes ne respectant pas l'anonymisation. Par exemple, la CNIL a prononcé en septembre 2024 une amende administrative de 800 000 € à l'encontre de la société CEGEDIM SANTÉ, qui traitait des données pseudonymisées et constituait un entrepôt de données de santé sans les autorisations requises par la loi et collectait illégalement des données issues du téléservice HRi (Historique des Remboursements intégré, qui permet aux professionnels de santé de consulter depuis leur logiciel métier les remboursements de santé effectués par l'assurance maladie pour un patient sur les 12 derniers mois).

Il devient donc nécessaire de définir correctement l'anonymisation et la pseudonymisation, pour que la vie privée des individus soit correctement protégée, tout en préservant au maximum l'utilité des jeux de données.

2.1.2 Définitions

Selon la norme ISO 29100, l'anonymisation est "*Le processus par lequel des informations personnellement identifiables sont altérées de façon irréversible, de sorte que la personne à laquelle se rapporte l'information ne puisse plus être identifiée directement ou indirectement*" [2].

Elle est à bien distinguer de la pseudonymisation, dont le RGPD donne une définition : "*Le traitement de données à caractère personnel de telle façon que celles-ci ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires, pour autant que ces informations supplémentaires soient conservées séparément et soumises à des mesures techniques et organisationnelles afin de garantir que les données à caractère personnel ne sont pas attribuées à une personne physique identifiée ou identifiable*" [2].

Clever Identity met un point d'orgue sur le fait que le processus de pseudonymisation est réversible [2], et qu'ainsi la suppression des données dites "identifiantes" ne suffit pas à assurer l'impossibilité de réidentification d'un individu. C'est alors tout le problème pour ceux chargés d'anonymiser les bases de données, qui doivent s'assurer de l'irréversibilité du processus.

2.2 Identifiants, Discrimination Rate

L'anonymisation et la pseudonymisation se réfèrent beaucoup à une notion d'identifiants qu'il est utile de bien définir.

2.2.1 Identifiants

2

Un *identifiant* est un attribut, ou ensemble d'attribut dans un jeu de données qui, étant connu, permet d'identifier une personne exactement dans un jeu de données anonymisé [3].

De manière similaire, un identifiant est dit *partiel* si il permet de diviser le jeu de données en différents petits groupes de 2 individus ou plus [3].

Enfin, un *zéro-identifiant* est un identifiant qui n'apporte aucune information supplémentaire et ne permet pas d'aider à l'identification d'un individu [3].

Subjects	ZIP Code	Age	Salary	Disease
Subject 1	35000	22	4K	Cancer
Subject 2	35000	35	5K	Diabetes
Subject 3	35000	63	3K	Malaria
Subject 4	35000	22	13K	Cancer
Subject 5	35000	22	8K	Cancer
Subject 6	35000	35	15K	Malaria
Subject 7	35000	45	9K	Malaria
Subject 8	35000	35	7K	Diabetes
Subject 9	35000	40	11K	Diabetes

Figure 1 – Exemple de jeu de données

Ici par exemple, le code postal n'apporte rien : c'est un zéro-identifiant. Par contre, sachant que tous les individus du jeu de données ont des salaires différents, on dit alors que le salaire est un identifiant. Enfin, l'âge est un identifiant partiel, car tous les individus n'ont pas le même âge, même si certains ont le même âge.

Il existe également une classification des identifiants selon leur visibilité dans les jeux de données [4], qui classe les identifiants dans 4 grandes catégories:

— "Internes Restreints" (IR): Ce sont des attributs liés à des opérations spécifiques qui ne sont pas utilisés en dehors du contexte de ces opérations. Ils sont par conséquent difficiles à trouver dans un autre ensemble de données (par exemple, certains résultats de tests médicaux, données techniques, etc.).

— "Internes Élargis" (IE): Ce sont des attributs utilisés par plusieurs services au sein d'une organisation, mais pas en dehors du contexte de cette organisation (par exemple, numéro de personnel, nom du département, etc.).

— "Externes Restreints" (ER): Ce sont des attributs qui peuvent être trouvés en dehors de l'organisation, mais qui nécessitent une recherche approfondie sur les réseaux sociaux (par exemple, fumeur/non-fumeur, etc.).

— "Externes Élargis" (EE): Ce sont des attributs qui sont largement utilisés dans différents contextes et qui sont facilement accessibles à partir des réseaux sociaux ou des moteurs de recherche (par exemple, données démographiques, nom de famille, prénom, âge, données de localisation, etc.).

Identifier à quelle catégorie se réfère un attribut, c'est savoir où chercher d'éventuelles bases de données où de la corrélation pourrait être faite.

2.2.2 Discrimination Rate

Puisque l'on vient de définir des notions d'information, il devient alors possible de faire des calculs d'entropie et de définir ce que l'on appelle un DR (Discrimination Rate):

Soient X et Y deux variables aléatoires discrètes. Le DR de Y relativement à X mesure la capacité de Y à affiner l'ensemble des issues de X, et se calcule comme suit:

$$DR_X(Y) = \frac{H(X) - H(X \mid Y)}{H(X)} = 1 - \frac{H(X \mid Y)}{H(X)} \tag{1}$$

Intuitivement, on retrouve quelques résultats:

$0 \leq DR_X(Y) \leq 1$, et:

— $DR_X(Y) = 0$ lorsque Y est un zéro-identifiant, c'est-à-dire lorsque $H(X \mid Y) = H(X)$: l'incertitude résiduelle est maximale.

— $DR_X(Y) = 1$ lorsque Y est un identifiant, c'est-à-dire lorsque $H(X \mid Y) = 0$: l'incertitude résiduelle est nulle.

Ici, Y se réfère à l'attribut clé que l'on a en notre possession, et X à l'attribut sensible.

2.3 Individualisation, inférence, corrélation et techniques d'anonymisation

2

2.3.1 Individualisation, inférence, corrélation

Dans des jeux de données anonymisés, un attaquant peut avoir plusieurs objectifs selon l'exploitabilité des données. On peut définir ces objectifs comme suit [2] :

L'*individualisation* : qui correspond à la possibilité d'isoler une partie ou la totalité des enregistrements identifiant un individu dans l'ensemble de données.

La *corrélation* : la capacité de relier entre elles, au moins deux enregistrements se rapportant à la même personne concernée ou à un groupe de personnes concernées (soit dans la même base de données, soit dans deux bases de données différentes). Si une attaque permet d'établir que deux enregistrements correspondent à un même groupe d'individus, mais ne permet pas d'isoler des individus au sein de ce groupe, la technique résiste à l'individualisation, mais non à la corrélation.

L'*inférence* : qui est la possibilité de déduire, avec un degré de probabilité élevé, la valeur d'un attribut à partir des valeurs d'un ensemble d'autres attributs.

2.3.2 Techniques d'anonymisation

Notre étude et nos expérimentations laissent à penser qu'en cas de bonne anonymisation des données, conformément aux critères du G29 (groupe des CNIL européennes), l'inférence (à des probabilités au-delà de 1/2) reste le moyen présentant le plus de succès. En effet, la réidentification d'individus au sens brut, corresponsant à l'individualisation, nécessite une expertise forte en cas de bases de données bien anonymisées, comme le montreront les cas détaillés dans 5.2. De manière générale, les méthodes d'anonymisation doivent, selon le G29, être robustes à l'individualisation, l'inférence et la corrélation.

Le G29 analyse l'efficacité et les limites des techniques d'anonymisation existantes dans le contexte juridique de la protection des données dans l'Union européenne et formule également des recommandations pour l'utilisation de ces techniques, en tenant compte du risque résiduel d'identification [5].

On peut isoler deux grandes familles de techniques d'anonymisation, que sont la *randomisation* et la *généralisation*, décrites comme suit par le G29 [5]:

La *randomisation* est une famille de techniques qui visent à altérer la véracité (l'exactitude) des données afin d'affaiblir le lien entre les données et l'individu concerné. La randomisation se pratique avec de l'injection de bruit (modifier les attributs pour les rendre moins précis), ou encore de la substitution de données (comme la permutation, qui mélange les valeurs des attributs entre les enregistrements).

La *généralisation*, quant à elle, consiste à généraliser ou diluer les attributs des personnes concernées en modifiant leur échelle ou leur ordre de grandeur respectif. Cela signifie rendre les informations moins précises en les remplaçant par des catégories plus larges (par exemple, remplacer une ville par une région ou un pays, ou une date de naissance exacte par une fourchette d'âge).

On dit qu'on applique de la *k-anonymisation* lorsque dans un jeu de données, on regroupe les individus de manière à ce que chaque personne soit indistinguable d'au moins $k-1$ autres personnes au sein d'un groupe.

	ZIP Code	Age	Location 1	Location 2
1	35877	22	Diabetes clinic	Church
2	35512	35	HIV clinic	Synagogue
3	35620	63	Diabetes clinic	Mosque
4	35517	22	Cardiology clinic	Church
5	35510	22	Cancer clinic	Church
6	35830	39	Cancer clinic	Mosque
7	35842	45	HIV clinic	Mosque
8	35618	35	Cardiology clinic	Synagogue
9	35655	40	HIV clinic	Synagogue

	ZIP Code*	Age*	Location 1	Location 2
1	*****	2*	Diabetes clinic	Church
4	*****	2*	Cardiology clinic	Church
5	*****	2*	Cancer clinic	Church
2	*****	3*	HIV clinic	Synagogue
6	*****	3*	Cancer clinic	Mosque
8	*****	3*	Cardiology clinic	Synagogue
3	*****	≥ 40	Diabetes clinic	Mosque
7	*****	≥ 40	HIV clinic	Mosque
9	*****	≥ 40	HIV clinic	Synagogue

Figure 2 – Jeu de données non anonymisé (à gauche) et sa variante anonymisée (à droite)

Ici par exemple, on a rendu le jeu de données 3-anonyme en appliquant de la généralisation sur les âges et un cas extrême de randomisation sur les codes postaux (supression de l'attribut qui devient totalement masqué).

3
Principes de base de l'OSINT

Sommaire

3.1 Qu'est ce que l'OSINT ?

L'OSINT (Open Source Intelligence) est le recueil et l'analyse d'informations obtenues à partir de sources d'informations publiques. Les jeux de données Open Source étant de plus en plus démocratisés, leur anonymisation est nécessaire pour ne pas mettre en péril la vie privée des individus.

Il s'agit donc d'essayer d'intégrer l'OSINT à notre étude pour estimer s'il est raisonnable d'évaluer les risques de prendre en compte l'OSINT en tant que moyen potentiel de réidentification d'individus dans des jeux de données présentant des attributs sensibles.

L'OSINT est un domaine extrêmement large qui peut cibler à la fois des personnes physiques et des entreprises. Nous nous sommes ici restreints aux individus, et aux jeux de données présentant des attributs potentiellement sensibles, pour justifier la pertinence d'un "risque" de réidentification.

3.2 Obtenir des informations sur des personnes physiques

Il existe en France (et probablement dans d'autres pays) une branche de l'OSINT qui ne s'effectue pas grâce aux outils numériques, mais grâce à l'administration publique.

En effet, sur ozint.eu, plateforme communautaire dédiée à l'OSINT, on apprend qu'il est possible de retrouver un certain nombre d'informations sur des individus, grâce à des méthodes synthétisées dans ce tableau:

Nature de l'information	Démarche	Pré-requis
État civil d'une personne	Listes électorales de la mairie	Disposer d'une carte électorale Connaître la mairie de résidence
Extrait d'acte de naissance (sans filiation)	Mairie du lieu de naissance	État civil complet de la personne (nom, date, lieu de naissance)
Impôt payé, revenu imposable, nombre de parts	DDFiP	Résider dans le ressort de la même DDFiP

Table 1 – Sources administratives d'informations personnelles et conditions d'accès

3.3 Outils OSINT

Les premiers outils OSINT disponibles par des recherches rapides sur le net concernent des outils de recherche inverse où, connaissant l'adresse mail / le pseudonyme d'un individu sur un compte, on peut obtenir la liste des sites où des comptes avec cette adresse ou ce pseudonyme correspondent à un utilisateur en base de données. *Blackbird* est un de ces outils, tout comme *Epieos*. Toutefois, nous avons également remarqué l'importance des réseaux sociaux en tant que sources OSINT.

Par exemple, grâce à des recherches menées sur Vkontakte, réseau social extrêmement populaire en Russie, un certain nombre d'informations ont pu être retrouvées sur les propriétaires d'un hébergeur qui fournissait une infrastructure facilitant les activités cybercriminelles [6].

Cet exemple nous montre bien que l'OSINT est un outil extrêmement contextuel. Si l'étude n'était pas en Russie, Vkontakte n'aurait pas été une source vers laquelle les enquêteurs se seraient naturellement penchés.

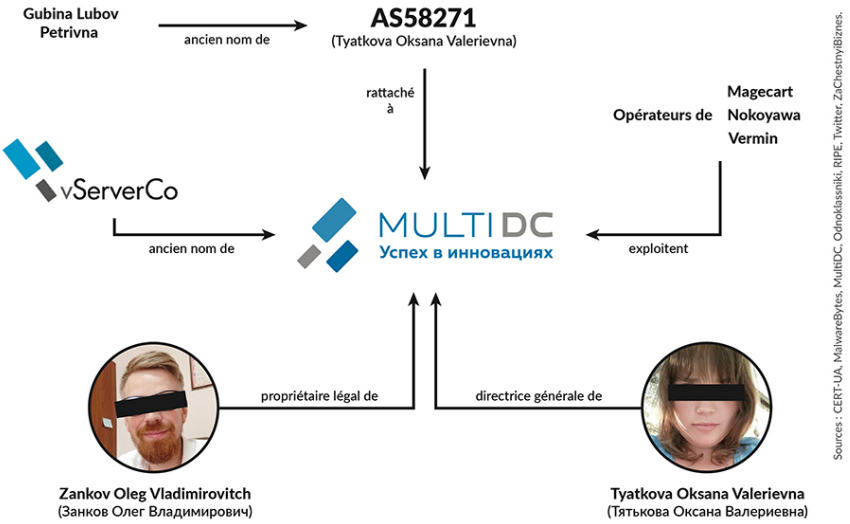

Figure 3 – Cas où de nombreux liens ont pu être établis sur des organismes et individus dans le cadre d'une enquête menée à l'aide de l'OSINT

Dans nos mises en pratique, l'usage de LinkedIn a semblé être le plus puissant en terme de rapport résultat / prédispositions. En effet, les recherches liées à l'administration publiques sont plus contextuelles et moins faciles d'accès car demandent parfois (comme pour la DDFiP) un peu de chance et d'audace. Concernant les outils de recherche à partir de pseudonymes ou d'adresses mail, elles demandent une connaissance préalable et restent plus limitantes que LinkedIn dans le cas des jeux de données, plateforme où l'on peut retrouver beaucoup d'identifiants externes élargis. Nous verrons en 5.2 un exemple d'application où l'usage de LinkedIn s'avère très pertinent.

Au terme de nos recherches sur l'OSINT, nous avons récapitulé dans un tableau les différentes méthodes d'OSINT que nous avons eu l'occasion d'explorer et de tester pour la plupart, hors déplacements dans les bâtiments de l'administration publique.

Outil	Usage	Prédispositions nécessaires
Blackbird	Retrouver des comptes en ligne liés à un certain nom d'utilisateur	Connaître le nom d'utilisateur qu'on veut croiser
LinkedIn	Retrouver des informations quant à la vie professionnelle de personnes physiques	Éventuellement l'adresse électronique ou le numéro de téléphone si on ne connaît pas le nom de la personne
Listes électorales de la mairie	Retrouver l'état civil d'une personne physique	Disposer d'une carte électorale
Mairie du lieu de naissance	Se procurer un extrait d'acte de naissance (sans filiation)	Connaître l'état civil complet de la personne (nom, date, lieu de naissance)
DDFiP	Retrouver l'impôt payé, le revenu imposable, le nombre de parts DDFiP	Résider dans le ressort de la même DDFiP et avoir de la chance
INSEE	Liste des personnes décédées en France depuis 1970	Aucune
GENEANET	Informations sur la généalogie d'une personne	Aucune
ozint.eu	S'entraîner à maîtriser les techniques d'OSINT classiques et se documenter sur l'OSINT	Aucune
Epieos	Trouver des informations sur une personne en ligne avec une recherche inversée par email ou numéro de téléphone	Aucune

Table 2 – Outils et techniques associés à l'OSINT

4
Notion de risque de réidentification

Sommaire

4.1 Définition et pertinence

La notion de risque de réidentification occupe une place centrale dans les débats autour de la protection des données personnelles, notamment dans le cadre de l'anonymisation. En France, la CNIL ne propose pas de méthodologie systématique pour quantifier ce risque, mais elle en reconnaît l'existence implicite à travers ses recommandations sur les techniques d'anonymisation. Le risque de réidentification peut être défini comme la probabilité qu'un individu soit identifié à partir de données censées être anonymes, par recoupement avec d'autres sources d'information, publiques ou privées. Cette définition met en évidence l'importance du contexte : l'anonymat n'est jamais absolu, mais dépend fortement des capacités de l'attaquant, de la nature des données en circulation et de la facilité d'accès aux bases auxiliaires.

Il est donc pertinent d'associer un niveau de risque à chaque jeu de données anonymisé. En effet, une anonymisation considérée comme suffisante dans un contexte donné peut s'avérer fragile si les données sont croisées avec des sources extérieures. Cette évaluation du risque permet de formuler des hypothèses d'attaques, appelées chemins d'attaque, qui simulent les actions d'un adversaire tentant de réidentifier une personne. Ces chemins incluent, par exemple, l'appariement avec des données publiques (listes électorales, réseaux sociaux), l'inférence statistique, ou encore l'utilisation de techniques d'intelligence artificielle pour reconstruire des informations manquantes.

Le risque de réidentification est donc étroitement lié à l'irréversibilité de l'anonymisation : plus une anonymisation est réversible, plus elle expose les individus. L'échec d'une anonymisation a des conséquences concrètes : violation de la vie privée, stigmatisation, discrimination, et dans certains cas, préjudices économiques ou sociaux importants. C'est pourquoi l'analyse de risque doit être intégrée en amont de tout processus d'anonymisation, et non considérée comme une

étape secondaire. Elle permet de choisir des méthodes adaptées, d'évaluer leur robustesse, et de garantir que les données publiées ne puissent pas nuire aux personnes concernées.

4.2 Modélisation mathématique du risque

Le risque de réidentification peut être modélisé à l'aide de la formule générique utilisée dans l'analyse des risques :

$$R = S \times L$$

où R représente le niveau de risque, S la gravité (*Severity*) de l'impact en cas de réidentification, et L la vraisemblance (*Likelihood*) de survenue de cette réidentification. Cette approche permet de dépasser les simples mesures statistiques du type k-anonymat, en intégrant des dimensions contextuelles et humaines. Par exemple, la gravité dépendra de la sensibilité des données en question : des données de santé, des opinions politiques ou une orientation sexuelle auront un impact bien plus important qu'un code postal ou une tranche d'âge.

La vraisemblance, quant à elle, est influencée par plusieurs facteurs : la présence de quasi-identifiants dans les données, la rareté de certaines combinaisons d'attributs, l'accès potentiel à des bases auxiliaires, ainsi que les compétences techniques et les motivations des attaquants. Il s'agit donc d'une estimation contextuelle, qui peut être affinée à l'aide de simulations, d'analyses empiriques ou encore d'approches bayésiennes.

La formalisation par $R = S \times L$ présente l'avantage de s'adapter à des cas d'usage concrets. Par exemple, dans un contexte de recherche médicale, la vraisemblance de réidentification peut être relativement faible, mais la gravité est très élevée si les données concernent des maladies rares. À l'inverse, un jeu de données comportant des caractéristiques très distinctives (sexe, âge, ville, profession) peut avoir une vraisemblance élevée, mais une gravité modérée si les attributs sont peu sensibles.

Cette approche permet aussi une gestion dynamique du risque : en modifiant les techniques d'anonymisation (suppression d'attributs, généralisation, perturbation), on peut faire varier L sans altérer nécessairement la valeur informative des données. Cela aide à atteindre un équilibre entre la protection de la vie privée et l'utilité des données, objectif fondamental du traitement de données à caractère personnel.

4.3 Clever Identity pour l'analyse de risque

Un des deux encadrants de notre étude, Louis Philippe Sondeck, nous a permis d'accéder à un outil d'évaluation du risque de réidentification : *Clever Identity App*. Cet outil, développé par la société Clever Identity (dont Monsieur Sondeck est le CEO), nous a offert l'opportunité d'évaluer de manière concrète le risque de réidentification sur des jeux de données réels, en appliquant une méthodologie rigoureuse et reproductible.

4

Clever Identity App se distingue des approches purement théoriques par son orientation opérationnelle : l'outil prend en entrée un jeu de données pseudonymisé ou anonymisé, et identifie automatiquement les attributs pouvant servir de quasi-identifiants. Il évalue ensuite le degré d'unicité des individus en fonction de différentes combinaisons d'attributs, permettant d'estimer à quel point certaines personnes sont potentiellement réidentifiables. L'outil fournit un score de risque synthétique, mais également des visualisations détaillées sur la distribution de ces risques au sein du jeu de données.

Ce que nous avons particulièrement apprécié lors de l'utilisation de *Clever Identity App*, c'est la clarté de ses résultats. L'outil ne se limite pas à un diagnostic binaire (réidentifiable ou non), mais propose une cartographie des zones à risque au sein du dataset, permettant d'orienter les stratégies d'anonymisation (généralisation, suppression, bruit, etc.). Il facilite également la prise de décision en matière de diffusion ou de conservation des données, en mettant en évidence les compromis entre utilité et confidentialité.

Nous avons eu l'occasion de mettre en avant une utilisation concrète de cet outil sur un jeu de données synthétique comprenant des données de santé dont voici la structure :

Nom	Exposition
Antibiogramme	1- Interne restreint
Concentration minimale inhibitrice	1- Interne restreint
Espèce bactérienne	1- Interne restreint
Type de prélèvement	1- Interne restreint
Date de prélèvement	2- Interne élargi
Spectre MALDI-TOF	1- Interne restreint
Date de naissance	4- Externe élargi
Sexe	4- Externe élargi

Figure 4 – Structure du jeu de données et caractérisation des attributs

Il est possible de préciser les risques liés à chaque attribut :

Variable	Inférence	Individualisation	Corrélation	Exploitabilité	Personnes à risque
Antibiogramme	2- Modérée	1- Faible	1- Faible	1- Très Difficile	-
Concentration minimale inhibitrice	4- Très élevée	4- Très élevée	1- Faible	2- Difficile	8
Type de prélèvement	2- Modérée	3- Elevée	1- Faible	1- Très Difficile	-
Date de prélèvement	4- Très élevée	4- Très élevée	2- Modérée	3- Facile	220
Spectre MALDI-TOF	4- Très élevée	4- Très élevée	1- Faible	2- Difficile	102
Date de naissance	4- Très élevée	4- Très élevée	4- Très élevée	4- Très Facile	19
Sexe	2- Modérée	1- Faible	2- Modérée	2- Difficile	-

Figure 5 – Risques liés à chaque attribut

On peut ensuite proposer des contre-mesures pour chaque risque :

Référence	Libellé	Vulnérabilités
CM01	Generalize or randomize variables «Extended Internals»: Date de prélèvement	VULD01
CM02	Generalize or randomize variables «Restricted Internals»: Antibiogramme, Concentration minimale inhibitrice, Type de prélèvement, Spectre MALDI-TOF	VULD03
CM03	Generalize or randomize variables «Extended Externals»: Date de naissance, Sexe	VULD02

Référence	Libellé	Evénements redoutés
CM04	Modify or diversity «Severity 2, 3 and 4 feared events» values	FE01

Figure 6 – Contre-mesures proposées

Enfin, pour chaque risque, on peut mesurer la réduction du risque à l'aide des contre-mesures précédemment identifiées.

Référence	Intitulé	Vulnérabilités	Initial	Contre-mesures applicables	Cible
R01	An hacker with access to the data can re-identify people based on «Extended Internals»: Date de prélèvement	VULD01	4- Critique	CM01 + CM04	2- Significatif
R02	An hacker with access to the data can re-identify people based on «Restricted Internals»: Antibiogramme, Concentration minimale inhibitrice, Type de prélèvement, Spectre MALDI-TOF	VULD03	2- Significatif	CM02 + CM04	2- Significatif
R03	An hacker with access to the data can re-identify people based on «Extended Externals» variables: Date de naissance, Sexe	VULD02	4- Critique	CM03 + CM04	2- Significatif

Figure 7 – Risque final en prenant en compte les contre-mesures

L'outil est un très bon moyen de quantifier les risques pour des jeux de données anonymisés, avec un fort potentiel d'adaptabilité qui rend l'analyse de risques moins contextuelle qu'initialement.

5

Cas pratiques : panorama et analyse critique des différents outils

Sommaire

5.1 Exemple d'inférence concernant l'opinion politique

Notre premier cas pratique concerne un jeu de données contenant des informations sur l'opinion politique en France. Il contient les résultats définitifs du premier tour des élections présidentielles de 2022 commune par commune. Il est construit comme suit :

Nom de colonne	Description
`Code de la commune`	Code INSEE de la commune (à trois chiffres)
`Libellé de la commune`	Nom complet de la commune
`Etat saisie`	État d'avancement de la saisie des résultats (ex. `Complet`)
`Inscrits`	Nombre total d'électeurs inscrits sur les listes électorales
`Abstentions`	Nombre d'électeurs inscrits n'ayant pas voté
`% Abs/Ins`	Pourcentage d'abstention parmi les inscrits
`Votants`	Nombre total de votants
`% Vot/Ins`	Pourcentage de votants parmi les inscrits
`Blancs`	Nombre de bulletins blancs
`% Blancs/Ins`	Pourcentage de bulletins blancs par rapport aux inscrits
`% Blancs/Vot`	Pourcentage de bulletins blancs par rapport aux votants
`Nuls`	Nombre de bulletins nuls
`% Nuls/Ins`	Pourcentage de bulletins nuls par rapport aux inscrits
`% Nuls/Vot`	Pourcentage de bulletins nuls par rapport aux votants
`Exprimés`	Nombre de suffrages exprimés (hors blancs et nuls)
`% Exp/Ins`	Pourcentage des exprimés par rapport aux inscrits
`% Exp/Vot`	Pourcentage des exprimés par rapport aux votants
`Données candidats`	Données répétées pour chaque candidat : - `N°Panneau` : Numéro de panneau officiel du candidat - `Sexe` : Sexe du candidat (`M` ou `F`) - `Nom` : Nom de famille du candidat - `Prénom` : Prénom du candidat - `Voix` : Nombre de voix obtenues - `% Voix/Ins` : Pourcentage des voix rapporté aux inscrits - `% Voix/Exp` : Pourcentage des voix rapporté aux exprimés

Table 3 – Structure détaillée des résultats électoraux par commune avec résultats par candidat (résumé des colonnes répétées)

Le jeu de données, produit par le Ministère de l'Intérieur, provient de la plateforme française des données publiques data.gouv.fr.

Notre objectif est d'essayer d'inférer sur les résultats présents dans ce jeu de données pour savoir si certaines communes présentent des risques vis-à-vis du lien entre une information sensible et un groupe assez restreint d'individus.

Notre démarche s'est organisée en deux temps :

— La première approche a été d'isoler les communes présentant le moins d'habitants, puisque nous pensions que ces communes allaient présenter la plus grande exposition à l'inférence

— La deuxième approche a été d'identifier, dans tout le jeu de données, les villes où le pourcentage de suffrages exprimés pour un certain courant politique était le plus élevé, et de voie à combien ce pourcents on pouvait inférer.

La première approche est la suivante :

```
import pandas as pd

# Charger le fichier Excel
df2 = pd.read_excel("elections/resultats-par-niveau-subcom-t1-france-entiere.xlsx")

# Trier par nombre d'inscrits (ordre croissant) sur tout le dataframe
df_sorted = df2.sort_values(by="Inscrits", ascending=True)

# Sélectionner les 10 communes avec le moins d'inscrits
df_top10 = df_sorted.head(10)

# Colonnes à afficher dans la console
colonnes_afficher = [
    "Code du département",
    "Libellé du département",
    "Code de la commune",
    "Libellé de la commune",
    "Inscrits"
]

# Affichage console
print(df_top10[colonnes_afficher])

# Sauvegarder dans un fichier Excel
df_top10.to_excel("elections/10_communes_moins_inscrits.xlsx", index=False)
```

Le résultat du code est un tableau comprenant les dix villes présentant le moins d'inscrits, et leurs différents attributs. Notre intuition était que moins il y avait d'habitants dans une commune, plus il y avait de chances de pouvoir isoler des individus comme appartenant avec une forte probabilité à un courant politique.

Voici un extrait du tableau obtenu :

	Coc	Libellé du départeme	Code	Libellé de la commune	Etat saisie	Inscrits
2	54	Meurthe-et-Moselle	310	Leménil-Mitry	Complet	4
3	31	Haute-Garonne	127	Caubous	Complet	7
4	51	Marne	470	Rouvroy-Ripont	Complet	8
5	26	Drôme	274	Rochefourchat	Complet	8
6	11	Aude	82	Caunette-sur-Lauquet	Complet	9
7	21	Côte-d'Or	303	Les Goulles	Complet	9
8	80	Somme	270	Epécamps	Complet	10
9	55	Meuse	394	Ornes	Complet	11
10	52	Haute-Marne	109	Charmes-en-l'Angle	Complet	11
11	52	Haute-Marne	4	Aingoulaincourt	Complet	12

Figure 8 – Extrait du tableau des dix villes contenant le moins d'inscrits en 2022 pour les élections présidentielles

Parmi ces villes, aucune n'est réellement intéressante sur le plan de l'inférence, avec aucune ville où on arrive au moins à 60% à associer une commune à un courant politique. Il est alors nécessaire de reconsidérer notre approche pour essayer d'obtenir des résultats.

La deuxième approche est la suivante :

— L'objectif est d'obtenir trois tableaux : un contenant les dix communes votant le plus à gauche (là ou la somme des pourcentages des candidats Mélenchon, Arthaud, Poutou, Roussel est maximale), un autre contenant les dix communes votant le plus à droite (là où la somme des pourcentages des candidats Le Pen et Zemmour est maximale), et un dernier contenant les dix communes votant le plus pour le parti présidentiel (là ou le pourcentage du candidat Macron est maximal).

— On parcourt donc le tableau et on somme à chaque fois selon l'information recherchée, et on construit trois tableaux supplémentaires, avec un rajout de colonne qui permet d'évaluer le pourcentage de la population du village inscrite sur liste électorale votant pour un certain courant politique.

Voici le code correspondant :

```
import pandas as pd

df = pd.read_csv("elections/resultats-par-niveau-subcom-t1-france-entiere.csv", low_memory=False)

nom_idx = df.columns.get_loc('Nom')
gauche_candidats = {'MÉLENCHON', 'ARTHAUD', 'POUTOU', 'ROUSSEL'}

def top10_communes_candidats(candidats_cibles):
    voix_par_commune = {}
    noms_communes = {}

    for idx, row in df.iterrows():
        total_pourc = 0
        code_dep = row['Code du département']
        code_com = row['Code de la commune']
        nom_com = row["Libellé de la commune"]
        noms_communes[(code_dep, code_com)] = nom_com

        i = nom_idx
        while i < len(df.columns):
            cell = row.iloc[i]
            if isinstance(cell, str) and cell.upper() in candidats_cibles:
                pct_col_index = i + 3
                if pct_col_index < len(df.columns):
                    try:
                        pct_val = float(row.iloc[pct_col_index])
                        total_pourc += pct_val
                    except:
                        pass
                i += 4
            else:
                i += 1

        voix_par_commune.setdefault((code_dep, code_com), []).append(total_pourc)

    moyenne_par_commune = {commune: sum(vals)/len(vals) for commune, vals in voix_par_commune.items()
    ↪ }
    sorted_communes = sorted(moyenne_par_commune.items(), key=lambda x: x[1], reverse=True)
    return sorted_communes[:10], noms_communes
```

```
def filtre_et_ajoute(df, top10, noms_communes, description_col):
    communes_set = set(commune for commune, _ in top10)
    pourc_dict = dict(top10)
    filtered_rows = []

    for idx, row in df.iterrows():
        code_dep = row['Code du département']
        code_com = row['Code de la commune']
        if (code_dep, code_com) in communes_set:
            row[description_col] = pourc_dict[(code_dep, code_com)]
            filtered_rows.append(row)

    return pd.DataFrame(filtered_rows)

# Extrême droite
top10_lpz, noms_communes = top10_communes_candidats({'LE PEN', 'ZEMMOUR'})
df_lepen_zemmour = filtre_et_ajoute(df, top10_lpz, noms_communes, '% Inference Le Pen + Zemmour')
df_lepen_zemmour.to_excel("elections/top10_communes_lepen_zemmour_lignes.xlsx", index=False)

# Macron
top10_macron, _ = top10_communes_candidats({'MACRON'})
df_macron = filtre_et_ajoute(df, top10_macron, noms_communes, '% Inference Macron')
df_macron.to_excel("elections/top10_communes_macron_lignes.xlsx", index=False)

# Gauche / Extrême gauche
top10_gauche, _ = top10_communes_candidats(gauche_candidats)
df_gauche = filtre_et_ajoute(df, top10_gauche, noms_communes, '% Inference Gauche')
df_gauche.to_excel("elections/top10_communes_gauche_lignes.xlsx", index=False)
```

Ici, les résultats sont bien meilleurs. Par exemple, pour les candidats de gauche / extrême gauche principaux :

1	Code	Code c	Nom commune	Moyenne % Gauche
2	09	172	Loubaut	76
3	05	47	Eourres	70.59
4	26	359	Vachères-en-Quint	67.65
5	11	306	Quirbajou	66.66
6	09	12	Appy	63.65
7	26	36	Beaumont-en-Diois	62.07
8	48	98	Molezon	61.63
9	22	365	Trémargat	61.53
10	26	122	Espenel	60.43
11	34	72	Celles	60

Figure 9 – Dix communes ayant le plus voté pour les candidats de gauche / extrême gauche principaux en 2022

Le meilleur résultat ici concerne la commune de Loubaut, votant, selon la base de données, à 76% pour des candidats de gauche. Si on regarde en détails dans le jeu de données originel, on a, pour une commune comptant à l'heure actuelle 28 habitants, dont 25 inscrits sur liste électorale, 18 d'entre eux qui votent pour le candidat Mélenchon, et 1 pour la candidate Arthaud. Ainsi, on peut imaginer qu'à l'aide des boîtes aux lettres présentes dans le village, que certains habitants peuvent voir la confidentialité de leur opinion politique menacée par la mise à disposition de ce jeu de données en Open Source. C'est une forme d'inférence à l'aide de l'OSINT où l'on peut, à 76%, affirmer qu'un habitant de Loubaut vote pour un certain courant politique.

De même, pour les candidats d'extrême droite:

1	Code départe	Code commu	Nom commur	Moyenne % Le Pen + Zemmour
2	72	214	Nauvay	75
3	02	87	Bieuxy	66.67
4	10	47	Blignicourt	65.85
5	10	105	Courcelles-su	65.39
6	55	434	Rigny-Saint-M	64.45
7	52	201	Flammerécour	61.41
8	80	418	Hardecourt-au	60.87
9	51	157	Coizard-Joche	60.72
10	12	9	Arnac-sur-Dou	60.50
11	62	396	Guinecourt	60

Figure 10 – Dix communes ayant le plus voté pour les candidats d'extrême droite principaux en 2022

Ici le meilleur résultat concerne Nauvay, où les habitants inscrits sur les listes électorales (11 sur 12 habitants dans la commune) votent à 75% pour les candidats d'extrême droite, dont 8 pour la candidate Le Pen, et 7 pour le candidat Zemmour.

On voit que globalement, notre intuition sur les petites villes n'était pas mauvaise, puisqu'on retrouve les pourcentages les plus élevés dans des communes de moins de 25 habitants. De manière générale, pour la gauche / extrême gauche et l'extrême droite, les 10 communes présentes dans le tableau final contiennent moins de 100 inscrits, restant globalement de très petites communes, néanmoins pas les plus petites présentes dans le jeu de données.

Pour le candidat Macron du parti présidentiel, aucune commune n'excède 60% en pourcentage maximal. Ceci peut être lié à plusieurs facteurs, comme le fait que l'on combine plusieurs candidats pour identifier des courants politiques dans les cas précédents, ou le contexte politique global en France.

5.2 Tests sur une base de données synthétique de données de santé mentale

Pour le deuxième exemple, nous avons cherché à tester des outils OSINT sur des jeux de données de santé, puisque ces jeux de données nous intéressaient particulièrement depuis le début de l'étude.

En effet, si on parle d'attribut sensible, la santé est probablement le domaine qui vient à l'esprit de la majorité en premier.

Comme expliqué en 2.1.1, les conséquences d'une fuite d'une donnée sensible de santé présentent des risques élevés pour la confidentialité des individus concernés.

La première étape, qui est donc la recherche d'un jeu de données adapté, est loin d'être la plus simple. Pour respecter la confidentialité de potentiels individus et par bonne protection des données de santé open source, nous avons travaillé sur un jeu de données synthétique issu de Kaggle, plateforme communautaire de partage de travaux sur les bases de données.

Le jeu de données a été créé par Shodolamu Opeyemi, data scientist, et présente des informations sur des profils pouvant s'apparenter à des étudiants indiens.

Le jeu de données se structure comme suit :

Nom de la colonne	Description
`id`	Identifiant unique de l'individu
`Gender`	Sexe de l'individu (Male / Female)
`Age`	Âge de l'individu (en années)
`City`	Ville de résidence de l'individu
`Profession`	Profession de l'individu (dans ce cas, toujours "Student")
`Academic Pressure`	Niveau de pression académique ressenti (échelle numérique)
`Work Pressure`	Niveau de pression liée au travail (échelle numérique)
`CGPA`	Moyenne académique cumulative (sur 10)
`Study Satisfaction`	Satisfaction par rapport aux études (échelle de satisfaction)
`Job Satisfaction`	Satisfaction par rapport à un emploi (souvent 0 pour étudiants sans emploi)
`Sleep Duration`	Durée moyenne de sommeil par nuit (catégorielle : `Less than 5 hours`, `5-6 hours`, etc.)
`Dietary Habits`	Habitudes alimentaires (ex. : `Healthy`, `Moderate`)
`Degree`	Diplôme poursuivi ou obtenu (ex. : BSc, PhD, M.Tech...)
`Have you ever had suicidal thoughts ?`	Réponse à la question : "Avez-vous déjà eu des pensées suicidaires ?" (Yes/No)
`Work/Study Hours`	Nombre moyen d'heures consacrées quotidiennement au travail ou aux études
`Financial Stress`	Niveau de stress financier perçu (échelle numérique)
`Family History of Mental Illness`	Présence d'antécédents familiaux de troubles mentaux (Yes/No)
`Depression`	Présence ou non de dépression (1 = Oui, 0 = Non)

Table 4 – Description des colonnes du jeu de données sur la santé mentale des étudiants

Notre objectif ici est d'identifier quels profils dans le jeu de données sont uniques en terme d'attributs externes élargis. En effet, ce sont ces attributs que l'on est le plus susceptibles de retrouver grâce à de l'OSINT.

On filtre d'abord sur ceux ayant répondu "Yes" à "Have you ever had suicidal thoughts ?", ce qui isole les profils sensibles.

On sélectionne ensuite, parmi les profils filtrés ceux qui sont uniques en termes de combinaison Âge + Ville + Profession + Diplôme + Genre, qui vont être les mieux réidentifiables.

Nous avons décidé d'implémenter un calcul de risque dans notre modélisation. Ici, on a choisi de définir un score de risque inversement proportionnel au nombre de fois qu'une combinaison apparaît, par $R = \frac{1}{N}$, où N est le nombre d'occurrences d'une combinaison d'externes élargis dans le jeu de données. Par exemple, si une combinaison apparaît une seule fois, on aura $R = 1/1 = 1$. Si une combinaison apparaît 2 fois, on aura $R = 1/2 = 0,5$.

On associe chaque profil à une recherche Google menant à des LinkedIn pouvant correspondre aux individus à risque dans le jeu de données.

On extrait enfin la proportion de profils uniques dans le jeu de données, à titre indicatif.

Notre code est le suivant:

```
import pandas as pd
import webbrowser
import urllib.parse

file_path = "depression/Student Depression Dataset.csv"
df = pd.read_csv(file_path)

quasi_identifiants = ["Gender", "Age", "City", "Profession", "Degree"]



filtered_df = df[df["Have you ever had suicidal thoughts ?"] == "Yes"]

rare_combinations = filtered_df.groupby(quasi_identifiants).size().reset_index(name='count')
rare_combinations['risk_score'] = rare_combinations['count'].apply(lambda x: 1 / x)

unique_profiles = rare_combinations[rare_combinations['risk_score'] == 1.0]
```

```
# Création de la page HTML pour afficher les profils
html_content = """
<html>
<head>
    <meta charset="utf-8">
    <title>Profils à risque de ré-identification</title>
    <style>
        body { font-family: Arial, sans-serif; margin: 20px; }
        table { border-collapse: collapse; width: 100%; }
        th, td { border: 1px solid #ccc; padding: 8px; text-align: left; }
        th { background-color: #f2f2f2; }
        tr:hover { background-color: #f5f5f5; }
        a { text-decoration: none; color: #337ab7; }
    </style>
</head>
<body>
    <h2>Profils uniques susceptibles d'être ré-identifiés</h2>
    <table>
        <tr>
            <th>Age</th>
            <th>City</th>
            <th>Profession</th>
            <th>Degree</th>
            <th>Gender</th>
            <th>Recherche Google</th>
        </tr>
"""

# Générer une ligne HTML pour chaque profil
for index, row in unique_profiles.iterrows():
    query = f"{row['Age']} ans {row['City']} {row['Profession']} {row['Degree']} {row['Gender']}
    ↪ LinkedIn"
    query_encoded = urllib.parse.quote(query)
    google_url = f"https://www.google.com/search?q={query_encoded}"

    html_content += f"""
        <tr>
            <td>{row['Age']}</td>
            <td>{row['City']}</td>
            <td>{row['Profession']}</td>
            <td>{row['Degree']}</td>
```

```
            <td>{row['Gender']}</td>
            <td><a href="{google_url}" target="_blank">Recherche</a></td>
        </tr>
    """

html_content += """
    </table>
</body>
</html>
"""

# Sauvegarder le fichier HTML
with open("depression/result.html", "w", encoding="utf-8") as f:
    f.write(html_content)

# Ouvrir la page HTML dans le navigateur
webbrowser.open("depression/result.html")

# Calculer la part des profils uniques dans la base filtrée
total_cases = len(filtered_df)
unique_cases = len(unique_profiles)

if total_cases > 0:
    proportion_unique = (unique_cases / total_cases) * 100
    print(f"Proportion de la base avec un score de risque de 1.0 : {proportion_unique:.2f}%")
```

Voici un extrait du résultat :

Profils uniques susceptibles d'être ré-identifiés

Age	City	Profession	Degree	Gender	Recherche Google
18.0	Agra	Student	BA	Female	Recherche
18.0	Kolkata	Student	MSc	Female	Recherche
19.0	Ahmedabad	Student	M.Pharm	Female	Recherche
19.0	Bangalore	Student	BHM	Female	Recherche
19.0	Indore	Student	MBBS	Female	Recherche
19.0	Kanpur	Student	MSc	Female	Recherche
19.0	Meerut	Student	M.Com	Female	Recherche
20.0	Bangalore	Student	B.Com	Female	Recherche
20.0	Bangalore	Student	BA	Female	Recherche
20.0	Ghaziabad	Student	M.Tech	Female	Recherche
20.0	Kanpur	Student	B.Com	Female	Recherche
20.0	Lucknow	Student	BHM	Female	Recherche
20.0	Meerut	Student	MBA	Female	Recherche

Figure 11 – Extrait du résultat du code identifiant les profils uniques dans le jeu de données et y associant une recherche Google menant à des profils LinkedIn potentiellement correspondant.

Les profils uniques dans le jeu de données correspondent à $31,02\%$ des profils présents dans le jeu de données.

Sur les profils présents dans le résultat, on peut tester l'outil de recherche Google associé. En voici un exemple :

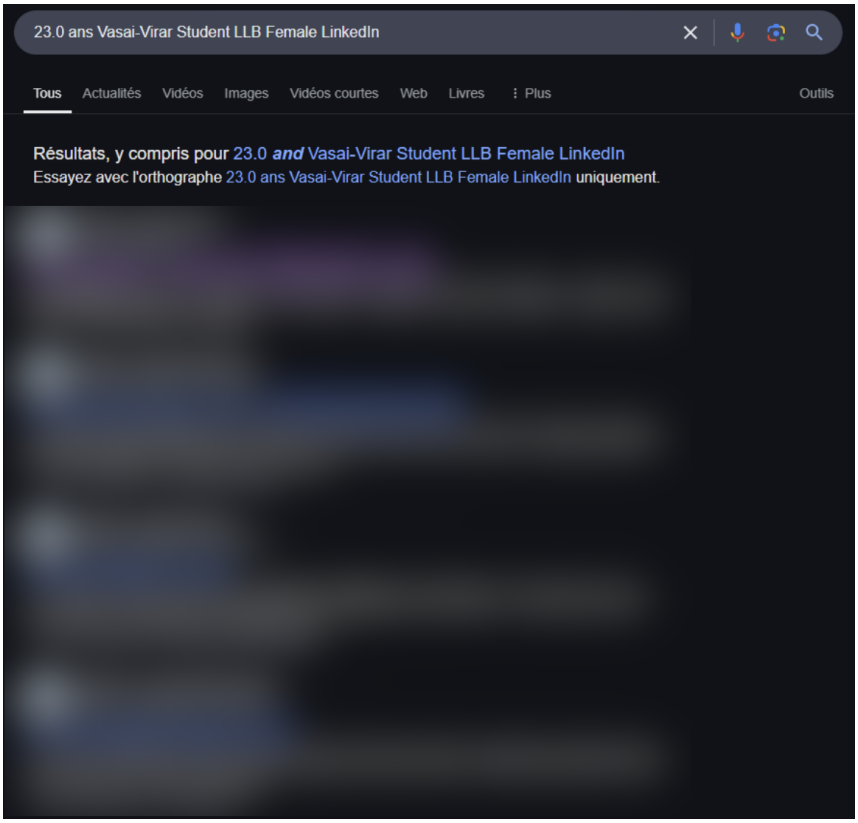

Figure 12 – Exemple de recherche Google associée à un profil unique

Pour ce profil en particulier, en explorant les profils LinkedIn proposés par la recherche Google, on en trouve un unique qui correspond à toutes les informations dans la base de données, ce qui laisse à penser qu'en cas d'accès à un vrai jeu de données présentant ce type de structure, un potentiel attaquant pourrait réidentifier sans trop de difficultés un profil. Ici, le jeu de données étant synthétique, les profils identifiés ne correspondent pas à des personnes physiques, mais on pourrait imaginer que cela soit le cas pour un vrai jeu de données, et que ce chemin d'attaque est vraisemblablement raisonnable.

Cet autre exemple met aussi en valeur l'importance de la bonne protection de la localisation des individus dans une base de données. En effet, plus une ville était petite, moins on trouvait de profils potentiels, et le risque de réidentification était donc plus important.

6
Complications et limites techniques et juridiques

Sommaire

6.1 Règlementations et législations

Malgré tout ce que nous venons de voir, il apparaît que les jeux de données sont de nos jours, fort heureusement, bien anonymisés pour la plupart.

Concernant les données de santé (qui nous intéressaient plus que d'autres, notamment à cause de la gravité qu'elles représentent en cas de réidentification), elles sont sécurisées grâce à des techniques telles que l'accès à distance et l'utilisation de fonctions de hachage comme la méthode FOIN (Fonction d'Occultation des Informations Nominatives), qui est un système de pseudonymisation sécurisé développé en 1996 par la Caisse Nationale de l'Assurance Maladie des Travailleurs Salariés (CNAMTS). Son objectif principal est de permettre le suivi longitudinal des trajectoires de soins des patients, ou plus généralement, de chaîner les informations relatives aux prises en charge successives d'un même patient dans les bases de données de santé, telles que le PMSI (Programme de Médicalisation des Systèmes d'Information) et le SNIIRAM (Système National d'Information InterRégimes de l'Assurance Maladie) pour garantir la confidentialité des informations tout en facilitant leur exploitation [7].

C'est tout le compromis à trouver : on veut que notre jeu de données soit correctement anonymisé, mais qu'il reste utilisable par ceux qui en ont besoin. C'est pourquoi le cadre légal autour de l'utilisation et de l'anonymisation des données de santé a été fortement débattu.

Il faut toutefois garder en tête que la méthode FOIN protège contre la réversion du pseudonyme mais pas contre l'inférence d'identité ou d'informations sensibles à partir de la combinaison de quasi-identifiants présents dans la base pseudonymisée, potentiellement enrichis par des informations

externes. D'autres techniques d'anonymisation (comme la généralisation, la suppression locale, l'agrégation, ou le bruitage des quasi-identifiants) ou des dispositifs d'accès sécurisé sont nécessaires pour limiter ce risque, et d'autant plus si les données sont destinées à de l'open source. Des techniques d'anonymisation supplémentaires (généralisation, suppression locale, échantillonnage, perturbation) sont nécessaires pour réduire le risque de ré-identification en rendant les individus moins distinguables.

La loi dite « Informatique et libertés » de janvier 1978 a posé la définition suivante, qui reconnaît les possibilités d'identification directe et indirecte : "Sont réputées nominatives les informations qui permettent, sous quelque forme que ce soit, directement ou non, l'identification des personnes physiques auxquelles elles s'appliquent".

Cependant, le Parlement et le Conseil européen ont voulu affiner encore cette définition en précisant, dans le considérant n°26 d'une directive européenne de 1995, les éléments à prendre en considération pour déterminer si une personne est, ou non, identifiable. Pour cela, dit la directive, il faut "considérer l'ensemble des moyens susceptibles d'être raisonnablement mis en œuvre, soit par le responsable du traitement, soit par une autre personne, pour identifier ladite personne" [7]. On en revient au souci de l'expertise nécessaire pour réidentifier. Si des moyens trop importants doivent être mis en œuvre pour réidentifier, cela revient à une reconsidération du risque potentiel.

Le projet de loi présenté à l'époque cherche un équilibre raisonné entre ouverture et protection des données de santé. D'un côté, certains plaident pour une ouverture maximale de ces données, vues comme un "trésor" pour la santé, la démocratie et l'économie, considérant que les risques de ré-identification sont surestimés ou acceptables au regard des bénéfices. De l'autre, d'autres mettent en garde contre les risques de ré-identification et de mésusage au regard du principe constitutionnel de protection de la vie privée, affirmant qu'il est impossible de rendre des jeux de données individuelles totalement anonymes pour un accès libre à tous.

6.2 Fuites de base de données, rôle du darknet

Le darknet constitue aujourd'hui un vecteur majeur de diffusion de bases de données issues de fuites, ou de piratages. Accessible via des réseaux/navigateurs comme Tor, il héberge des forums, marketplaces et dépôts anonymes, où circulent des informations sensibles : identifiants, adresses email, données de santé, historiques de navigation, etc. Pour des attaquants cherchant à réidentifier

des individus à partir de données pseudonymisées, le darknet offre une source précieuse d'informations auxiliaires, à croiser avec des jeux de données accessibles publiquement. Nous avons pu nous-mêmes expérimenter à quel point il est possible d'accéder avec une facilité déconcertante à des fuites disponibles sur le darknet. En moins d'une heure, il est possible (en suivant des tutoriels disponibles sur Internet, donc sans compétences particulières !) de se retrouver sur des sites proposant d'acheter des fuites pour quelques dizaines de dollars. Toutefois, comme nous le verrons en 6.3, ceci sort du cadre de l'OSINT.

L'un des cas les plus connus de fuite sur le darknet est celui de LinkedIn en 2021, où les données de 700 millions d'utilisateurs ont été mises en vente sur un forum du darknet. Ces informations ont été largement réutilisées pour du phishing, du ciblage publicitaire ou des tentatives de réidentification [8].

Cependant, l'accès au darknet soulève aussi des questions juridiques. En soi, utiliser Tor ou accéder à certains sites anonymes n'est pas illégal. En revanche, consulter, télécharger ou utiliser des bases de données volées constitue une infraction, notamment au regard du droit pénal (recel de données issues d'un délit, atteinte à la vie privée, etc.). Pour un analyste OSINT ou un chercheur en cybersécurité, cela impose de strictes précautions : documentation rigoureuse, encadrement légal, et limitation à des contextes légitimes (ex. : audit, recherche encadrée, etc.).

Le darknet représente donc une menace concrète dans toute stratégie de réidentification. Il alimente la vraisemblance d'attaque en mettant à disposition des données toujours plus riches et accessibles — mais il constitue également une zone grise sur le plan légal, qu'il ne faut jamais sous-estimer.

6.3 Le problème de l'Open Source et de l'OSINT

Le cadre légal autour des données de santé s'applique aussi pour l'open source, et donc pour l'OSINT, qui n'est qu'une utilisation de l'Open Source. Nos recherches nous ont bien montré qu'il était très difficile de trouver de réels jeux de données présentant des attributs sensibles et des quasi-identifiants. La facilité modérée de réidentification et le peu d'expertise nécessaire pour réidentifier dans de tels cas permettent de comprendre ces difficultés.

En effet, les données en open data (données ouvertes) sont expressément désignées comme libres de copie et de réutilisation, à titre professionnel ou personnel, notamment celles du secteur public. Ces données en open data, encadrées notamment par le règlement UE Data Governance Act (n°2022/868), doivent cependant avoir été préalablement traitées pour ne plus contenir de données à caractère personnel (obligation d'anonymisation) ni porter atteinte à un secret d'affaires ou un droit d'auteur [9].

L'usage des données OSINT légalement copiées dépend de leur origine. Les données en open data peuvent être réutilisées "à des fins commerciales ou non commerciales", de manière confidentielle ou publique. Les données d'une base privée légitimement consultée peuvent être réutilisées "à quelque fin que ce soit" si elles constituent des parties non substantielles. La réutilisation inclut le téléchargement ou l'indexation par lien http actif [9].

Publier publiquement des données même librement accessibles peut être problématique. Naviguer sur un site web ne donne pas le droit d'en copier le contenu (protection par le droit d'auteur, article L.335-2 du code de la propriété intellectuelle, 3 ans de prison et 300000€ d'amende). Publier des données à caractère personnel identifiées constitue un délit (traitement "frauduleux, déloyal ou illicite", article 226-18 du Code pénal, 5 ans de prison et 300000€ d'amende) [9].

Ces limites sur la technique envisagée, expliquées par le cadre légal autour de l'open source, invitent à reconsidérer les méthodes employées. L'OSINT permettrait à des potentiels attaquants de déblayer le terrain et de savoir où s'orienter, et peut dans certains cas permettre de réidentifier des personnes, quand il ne s'agit pas de jeux de données avec quasi-identifiants.

Toutefois, dès qu'on parle de jeux de données classiques contenant des attributs sensibles sur des personnes physiques, les méthodes employées pour obtenir des informations sensibles sont souvent différentes et illégales, comme l'intrusion de cybercriminels dans les systèmes d'informations des

organismes possédant les données (les hôpitaux par exemple si on parle de données de santé).

On peut citer par exemple la fuite de données provoquée par le cybergang *BianLian*, ayant mis à la vente sur le darknet une fuite de 300 Go de données de santé provenant du CHU de Rennes, contenant notamment des numéros de sécurité sociale [10].

Accéder à cette fuite sur le darknet n'est pas de l'OSINT. En effet, on arrive ici aux limites entre open source et données illégales et télécharger le produit d'un vol (comme un "leak" disponible sur un site .onion) constitue du recel de vol, sanctionné par 5 ans de prison et 375000€ d'amende (article 321-1 du Code pénal). L'article 323-3 du Code pénal punit également l'extraction, la détention ou la reproduction de données d'un SI auquel on aurait accédé frauduleusement, même sans intention de nuire [9].

Conclusion

Synthèse des enseignements clés

La distinction cruciale entre l'anonymisation, qui doit rendre l'identification irréversible pour échapper au RGPD, et la pseudonymisation, qui est réversible, souligne bien qu'il ne suffit pas de supprimer les identifiants directs ; une évaluation du risque de réidentification par des moyens raisonnables est indispensable, en considérant les critères réglementaires d'individualisation, de corrélation et d'inférence, auxquels l'anonymisation doit résister. L'étude suggère que l'inférence, notamment, présente un potentiel de succès plus important en cas de bonne anonymisation, sans toutefois garantir des résultats satisfaisants.

L'OSINT a été testé comme un moyen "raisonnable" d'aide à la réidentification. Les cas pratiques, comme l'analyse des résultats électoraux dans de petites communes ou les tests sur une base de données synthétique de santé mentale, ont démontré que l'OSINT peut parfois augmenter le risque de réidentification, particulièrement par inférence, en croisant des quasi-identifiants externes élargis (comme l'âge, la ville, la profession, le diplôme, le genre) avec des sources publiques (résultats électoraux, LinkedIn). Le risque d'inférence s'est montré élevé dans des contextes spécifiques, par exemple jusqu'à 76% pour l'opinion politique dans de très petites communes.

Cependant, d'importantes limites techniques et juridiques à la réidentification via l'OSINT sont également selon nous à prendre en compte. Les jeux de données, notamment ceux contenant des données sensibles comme la santé, sont souvent protégés par des techniques d'anonymisation robustes (généralisation, suppression, etc.) et des dispositifs d'accès sécurisé. Le cadre légal, notamment le critère des moyens raisonnablement mis en œuvre pour l'identification, introduit une barrière à l'évaluation des potentielles méthodes d'attaque. De plus, le cadre réglementaire strict autour de l'Open Data exige une anonymisation préalable des données publiques. Les risques les plus importants pour la vie privée proviennent souvent d'activités cybercriminelles illégales, telles que les fuites de données sur le darknet (dont l'accès constitue un délit de recel), plutôt que de l'exploitation légale de l'OSINT. L'étude a ainsi permis d'affiner la compréhension des critères d'évaluation des risques de réidentification en intégrant l'approche OSINT dans le cadre de ces limitations concrètes.

L'OSINT apparaît comme un moyen extrêmement contextuel, sans possibilité de formaliser ou de

généraliser des chemins d'attaques, même s'ils semblent souvent s'orienter vers les quasi-identifiants des jeux de données.

Pistes d'avenir quand au sujet du projet

Il faudra à l'avenir être bien attentif aux avancées technologiques de l'OSINT, notamment les outils d'agrégation, l'intelligence artificielle pour l'analyse de contenu et les plateformes collaboratives, afin de comprendre leur potentiel impact sur le risque de réidentification de données anonymisées. Nous pouvons par exemple nous interroger sur la confidentialité de nos données, à l'heure où Meta (anciennement Facebook) a annoncé vouloir entraîner ses systèmes d'IA avec les données de tous les utilisateurs européens adultes dès fin mai 2025.

L'OSINT est un domaine extrêmement large et des outils plus variés peuvent débloquer de nouvelles pistes de réflexion. À titre d'exemple, le *web scraping*, qui consiste en la collecte automatisée et parfois massive d'informations sur le Web peut constituer une forme d'automatisation de l'OSINT, et peut donner des résultats efficaces. C'est ce que des outils déjà existants comme Blackbird implémentent sur les pseudonymes par exemple, et qu'il est possible de faire de manière personnalisée.

La question de la formalisation et de la généralisation des différents chemins d'attaques possibles sur les jeux de données se pose également, permettant d'avoir une bonne vue d'ensemble de ce qu'il faut couvrir pour améliorer les standards au niveau de la sécurité des jeux de données.

Bibliographie

[1] Ministère de l'Économie, des Finances et de la Souveraineté industrielle et numérique. *L'emprunt avec un risque aggravé de santé (dispositif AERAS)*. URL: https://www.economie.gouv.fr/particuliers/emprunt-convention-aeras-credit.

[2] Clever Identity. *Anonymisation vs pseudonymisation : la notion d'irréversibilité*. 2023. URL: https://cleveridentity.fr/anonymisation-vs-pseudonymisation-la-notion-dirreversibilite/.

[3] Maryline Laurent, Louis Philippe Sondeck. *Discrimination rate: an attribute-centric metric to measure privacy*. 2017.

[4] Louis Philippe Sondeck, Maryline Laurent. *Practical and Ready-to-Use Methodology to Assess the Re-identification Risk in Anonymized Datasets*. 2025. URL: https://arxiv.org/abs/2501.10841.

[5] Groupe de travail "Article 29" sur la protection des données. *Recommandations relatives à la pseudonymisation des données à caractère personnel*. 2021. URL: https://www.cnil.fr/sites/cnil/files/atoms/files/wp216_fr.pdf.

[6] Julien Guilet. *Méthodes d'OSINT sur le Ru.net au profit de la CTI stratégique*. MISC : Hors-série n°27. 2023.

[7] DREES. *Des données individuelles aux données localisées : quelles limites à la désagrégation géographique ?* 2015. URL: https://www.casd.eu/wp/wp-content/uploads/dss64-2.pdf.

[8] Stella Rosso. *LinkedIn : les données de 700 millions d'utilisateurs en vente sur internet*. Siècle Digital. 2021. URL: https://siecledigital.fr/2021/07/01/linkedin-donnees-700-millions-utilisateurs-en-vente/.

[9] OZINT. *Livre blanc : Le cadre légal de l'OSINT*. 2023. URL: https://ozint.eu/contributions/Livre%20blanc-Le%20cadre%20legal%20OSINT-2023.pdf.

[10] Le Monde Informatique. *Des pirates bombardent des développeurs avec des npm malveillants*. 2024. URL: https://www.lemondeinformatique.fr/actualites/lire-des-pirates-bombardent-des-developpeurs-avec-des-npm-malveillants-96992.html.