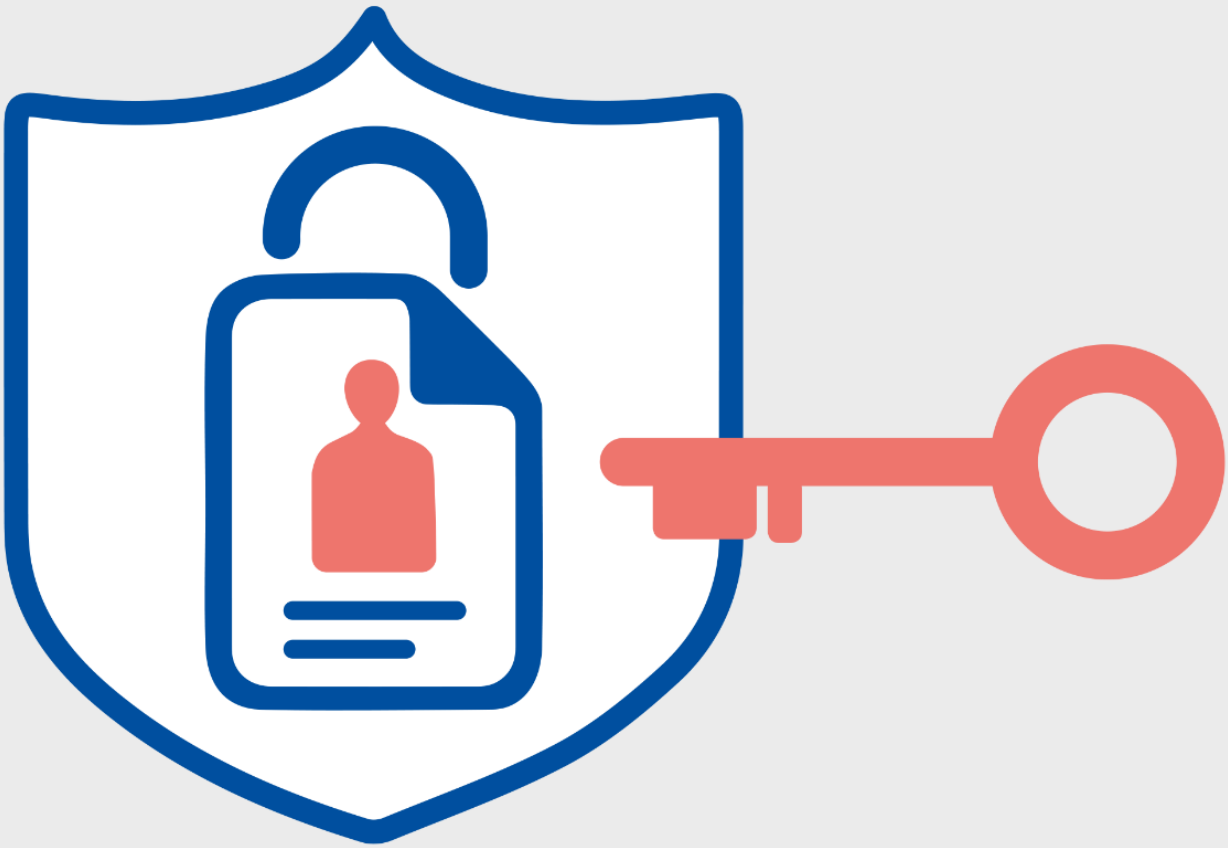
EUROPEAN UNION AGENCY
FOR CYBERSECURITY



TECHNIQUES ET MEILLEURES PRATIQUES DE PSEUDONYMISATION

Recommandations sur l'usage des technologies conformément aux dispositions en matière de protection des données et de respect de la vie privée

NOVEMBRE 2019

À PROPOS DE L'ENISA

Créée en 2004, l'Agence de l'Union européenne pour la cybersécurité (ENISA) a pour mission de garantir la cybersécurité au sein de l'Europe. L'ENISA travaille en collaboration avec l'Union européenne, ses États membres, le secteur privé et les citoyens européens afin d'établir des conseils et des recommandations sur les bonnes pratiques à appliquer en matière de sécurité de l'information. L'agence aide les États membres de l'UE à appliquer la législation européenne en vigueur, et s'efforce de renforcer la protection des infrastructures et réseaux d'information critiques en Europe. L'ENISA vise à améliorer les compétences existantes au sein des États membres de l'UE en favorisant le développement de communautés transfrontalières ayant pour vocation d'optimiser la sécurité des réseaux et de l'information à travers l'UE. Depuis 2019, elle met en place des programmes de certification en matière de cybersécurité. Pour plus d'informations sur l'ENISA et ses travaux, consultez le site https://www.enisa.europa.eu/media/enisa-en-francais/.

CONTACT

Pour contacter les auteurs de ce rapport, veuillez utiliser l'adresse isdp@enisa.europa.eu.
Pour les demandes de renseignements des médias concernant le présent document, veuillez utiliser l'adresse press@enisa.europa.eu.

CONTRIBUTEURS

Meiko Jensen (Université de Kiel), Cédric Lauradoux (INRIA), Konstantinos Limniotis (HDPA)

ÉDITEURS

Athena Bourka (ENISA), Prokopios Drogkaris (ENISA), Ioannis Agrafiotis (ENISA)

REMERCIEMENTS

Nous adressons nos remerciements à Giuseppe D'Acquisto (Garante), Nils Gruschka (Université d'Oslo) et Simone Fischer-Hübner (Université de Karlstad) pour la révision de ce rapport et pour leurs précieux commentaires.

MENTION LÉGALE

DÉCLARATION CONCERNANT LES DROITS D'AUTEUR

ISBN 978-92-9204-307-0, DOI 10.2824/247711

TABLE DES MATIÈRES

RÉSUMÉ

Dans le contexte du Règlement général sur la protection des données (RGPD)[1], la mise en œuvre appropriée d'une pseudonymisation des données à caractère personnel est en passe de devenir un sujet hautement débattu au sein de nombreuses communautés, qu'il s'agisse du monde des sciences, du monde académique, de la justice ou des organismes en charge de l'application des lois, ainsi que dans la gestion de la conformité pour diverses organisations européennes. Fondé sur les précédents travaux de l'ENISA dans ce domaine[2], le présent rapport explore de manière approfondie les notions de base de la pseudonymisation, ainsi que les solutions techniques à même d'en assurer la mise en œuvre pratique.

En particulier, sur la base de différents scénarios de pseudonymisation, le rapport commence par déterminer les principaux acteurs impliqués dans le processus de pseudonymisation, ainsi que leurs possibles rôles. Il analyse ensuite les différents modèles d'adversaires et les techniques d'attaque contre la pseudonymisation, telles que les attaques par force brute, par dictionnaire et par maximum de vraisemblance. Il présente également les principales techniques de pseudonymisation (par ex. les compteurs, le générateur de nombres aléatoires, la fonction de hachage cryptographique, le code d'authentification de message et le chiffrement) ainsi que les stratégies en matière de pseudonymisation (par ex. la pseudonymisation déterministe, par randomisation de documents ou entièrement aléatoire) disponibles aujourd'hui. Il traite en particulier des paramètres qui peuvent influencer en pratique le choix d'une technique ou d'une politique de pseudonymisation, notamment la protection des données, la fonctionnalité, l'évolutivité et les besoins de restauration. Certaines techniques de pseudonymisation plus avancées sont également abordées brièvement. Sur la base des descriptions précitées, le rapport présente deux cas d'utilisation de la pseudonymisation : 1) sur des adresses IP et 2) sur des adresses électroniques. Les particularités de chaque type d'identifiant sont analysées. Un cas d'utilisation plus complexe est également exposé sur la pseudonymisation d'enregistrements de données multiples, en discutant des possibilités de ré-identification.

L'une des principales conclusions de ce rapport est qu'il n'existe aucune solution de pseudonymisation simple et universelle qui fonctionnerait dans tous les cas de figures. Au contraire, un niveau élevé de compétences est nécessaire pour mettre en œuvre un processus de pseudonymisation robuste, ayant la capacité potentielle à réduire le risque de discrimination et les attaques de ré-identification, tout en offrant le degré de fonctionnalité nécessaire pour assurer le traitement des données pseudonymisées.

À cette fin, le présent rapport propose des conclusions et des recommandations destinées à l'ensemble des parties prenantes, en vue de l'adoption et de la mise en œuvre d'une solution de pseudonymisation des données.

[1] Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (Règlement général sur la protection des données) https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32016R0679

[2] Recommandations de l'ENISA sur l'usage des technologies conformément au RGPD: une introduction à la pseudonymisation des données, https://www.enisa.europa.eu/publications/recommendations-on-shaping-technology-according-to-gdpr-provisions

ADOPTER UNE APPROCHE BASEE SUR LES RISQUES EN MATIERE DE PSEUDONYMISATION

Bien que chaque technique de pseudonymisation connue présente ses propres propriétés intrinsèques et bien établies, le choix de la technique appropriée n'en est pas pour autant plus facile. Une approche basée sur les risques doit donc être adoptée, de manière à évaluer le degré de protection requis tout en définissant les besoins en matière de fonctionnalité et d'évolutivité.

Il est de la responsabilité des sous-traitants et des responsables du traitement des données d'étudier une approche basée sur les risques pour la mise en œuvre d'une solution de pseudonymisation, en prenant en compte la finalité et le contexte global du processus de traitement des données à caractère personnel, ainsi que les niveaux de fonctionnalité et d'évolutivité souhaités.

Les fournisseurs de produits, de services et d'applications doivent transmettre aux sous-traitants et aux responsables du traitement des informations adéquates concernant leur utilisation des techniques de pseudonymisation, ainsi que les niveaux de protection des données et de sécurité que celles-ci offrent.

Les organismes de régulation (par ex. les autorités de protection des données et le Comité européen de la protection des données) doivent fournir aux responsables du traitement des données et aux sous-traitants des consignes pratiques concernant l'évaluation des risques, tout en assurant la promotion des meilleures pratiques en matière de pseudonymisation.

DEFINIR L'ETAT DES CONNAISSANCES

Pour pouvoir adopter une approche basée sur les risques dans le domaine de la pseudonymisation, il est essentiel de définir l'état des connaissances en la matière. À cette fin, il est important de travailler sur des exemples et des cas d'utilisation précis, pour bénéficier d'un maximum de détails et d'options possibles sur l'implémentation technique.

Il est de la responsabilité de la Commission européenne et des institutions pertinentes au sein de l'Union européenne de favoriser la définition et la diffusion de l'état des connaissances en matière de pseudonymisation, en collaboration avec la communauté des chercheurs et l'industrie dans ce domaine.

Les organismes de régulation (par ex. les autorités de protection des données et le Comité européen de la protection des données) ont pour mission de promouvoir la publication des meilleures pratiques en matière de pseudonymisation.

FAIRE PROGRESSER L'ETAT DES CONNAISSANCES

Bien que ce rapport concerne les techniques de base de la pseudonymisation actuellement disponibles, l'utilisation de techniques plus avancées (et plus robustes), telles que celles issues des pratiques d'anonymisation, est très importante si l'on souhaite prendre en charge les scénarios de plus en plus complexes qui se présentent dans la pratique.

Le milieu de la recherche doit travailler à développer les techniques de pseudonymisation actuelles pour obtenir des solutions plus sophistiquées, prenant en charge avec efficacité les problématiques spécifiques associées à l'ère du Big Data. La Commission européenne et les institutions concernées au sein de l'UE doivent quant à elles soutenir et généraliser ces efforts.

1. INTRODUCTION

1.1 CONTEXTE

La pseudonymisation est un processus de désidentification reconnu qui suscite l'intérêt depuis l'entrée en vigueur du RGPD, dans lequel il est présenté autant comme un mécanisme intrinsèque de sécurité que de protection des données. En outre, dans le contexte du RGPD, la pseudonymisation peut justifier l'assouplissement, dans une certaine mesure et si elle est correctement employée, des obligations légales des responsables du traitement des données.

Au vu de l'importance croissante de ce processus pour les responsables du traitement des données et pour les personnes concernées, l'ENISA a publié en 2018 [1] un document de présentation du concept de pseudonymisation et des principales techniques en la matière, en lien avec son rôle au titre du RGPD. Commençant par une définition de la pseudonymisation (ainsi qu'une explication de ce qui la distingue des autres technologies, telles que l'anonymisation et le chiffrement), ce rapport aborde ensuite les avantages de la pseudonymisation en termes de protection des données essentielles. Suite à cette analyse, il présente différentes techniques pouvant être utilisées pour effectuer la pseudonymisation des données, notamment le hachage, le hachage avec clé ou avec salage, le chiffrement, la tokénisation, ainsi que d'autres approches pertinentes. Enfin, le rapport discute de certaines applications de pseudonymisation, en s'intéressant particulièrement au domaine des applications mobiles.

Bien que le travail de l'ENISA précité aborde quelques-unes des problématiques clés de la pseudonymisation, des études et des analyses complémentaires sont encore nécessaires, aussi bien pour renforcer le concept de pseudonymisation en tant que mesure de sécurité (art. 32 du RGPD) que pour façonner son rôle d'instrument de protection des données dès la conception (art. 25 du RGPD). En effet, comme le souligne également le rapport de l'ENISA, il est aujourd'hui particulièrement nécessaire de promouvoir les meilleures pratiques en matière de pseudonymisation et d'établir des exemples de cas d'utilisation à même de permettre la définition de l'état des connaissances dans ce domaine.

Dans un tel contexte, l'ENISA a intégré dans son programme de travail 2019 l'étude des applications pratiques de la pseudonymisation des données[3].

1.2 PORTEE ET OBJECTIFS

La mission globale du présent rapport est de fournir des consignes et des meilleures pratiques pour la mise en œuvre technique de la pseudonymisation des données.

Plus spécifiquement, ses objectifs sont les suivants:

- Discuter des différents scénarios de pseudonymisation et des acteurs pertinents impliqués.
- Présenter les techniques de pseudonymisation possibles en corrélation avec les modèles d'attaque et d'adversaire existants.

[3]L'ENISA ayant pour rôle d'établir des consignes sur les différents aspects des politiques de sécurité des réseaux et de l'information au sein de l'UE, il est logique que la prise en charge de certains domaines pertinents, tels que la protection des données et de la vie privée, soit une extension raisonnable de son mandat, en réponse aux besoins des parties prenantes. L'analyse de la mise en œuvre pratique de la pseudonymisation est un élément important pour garantir la sécurité des données à caractère personnel, tel qu'exposé à l'article 32 du RGPD.

- Analyser l'application de la pseudonymisation à des types d'identificateurs spécifiques, en particulier les adresses IP, les adresses électroniques et les autres types de jeux de données structurés (cas d'utilisation).
- Tirer les conclusions qui s'imposent et proposer des recommandations pour approfondir le travail sur le terrain.

Il est à noter que le choix des cas d'utilisation présentés se fondait sur le fait que les types d'identificateurs concernés (adresses IP, adresses électroniques, identificateurs de jeux de données structurés) correspondent à des scénarios courants dans le monde réel. Les cas d'utilisation choisis reflètent en outre les différentes exigences en matière de pseudonymisation, c'est-à-dire celles découlant du format strict des adresses IP et de la structure plus flexible des adresses électroniques, ou de la nature imprévisible des jeux de données volumineux.

Le public ciblé pour ce rapport est composé des responsables du traitement des données, des sous-traitants et des producteurs (produits, services et applications), des autorités de protection des données (APD), ainsi que des autres parties intéressées par la pseudonymisation des données.

La compréhension du présent document implique de connaître les principes de base de la protection des données à caractère personnel, ainsi que le rôle/le processus de pseudonymisation. Pour obtenir une présentation globale de la pseudonymisation des données au titre du RGPD, veuillez vous référer aux précédents travaux de l'ENISA en la matière [1].

La discussion et les exemples présentés dans ce rapport sont centrés sur les solutions techniques à même de promouvoir la protection des données et de la vie privée; il ne s'agit en aucun cas d'un avis juridique sur les cas concernés.

1.3 SOMMAIRE

Le rapport est structuré comme suit:

- Le chapitre 2 présente la terminologie utilisée dans le rapport, accompagnée d'une explication ou de remarques le cas échéant.
- Le chapitre 3 présente les scénarios les plus courants de pseudonymisation auxquels on peut s'attendre dans la pratique.
- Le chapitre 4 décrit les modèles d'attaque et d'adversaire possibles vis-à-vis de la pseudonymisation (ainsi que des scénarios précédemment présentés).
- Le chapitre 5 présente les principales techniques et stratégies de pseudonymisation disponibles aujourd'hui.
- Les chapitres 6, 7 et 9 analysent l'application des différentes techniques de pseudonymisation aux adresses IP, aux adresses électroniques et aux jeux de données plus complexes (cas d'utilisation).
- Le chapitre 0 offre une synthèse des discussions précédentes, ainsi que les principales conclusions et recommandations destinées à toutes les parties prenantes concernées.

Le présent rapport fait partie intégrante du travail de l'ENISA dans le domaine de la protection des données et de la vie privée[4], travail qui se concentre sur l'analyse des solutions techniques pour l'application du RGPD, le respect de la vie privée dès la conception, ainsi que la sécurité du traitement des données à caractère personnel.

[4] https://www.enisa.europa.eu/topics/data-protection

2. TERMINOLOGIE

Ce chapitre présente les termes qui seront utilisés tout au long de ce rapport et qui sont essentiels à sa compréhension par le lecteur. Certains de ces termes sont issus du RGPD, tandis que d'autres correspondent à des termes techniques standard ou sont explicitement définis pour les besoins du présent rapport.

Les principaux termes utilisés dans le rapport sont les suivants:

Le terme **données à caractère personnel** désigne toute information se rapportant à une personne physique identifiée ou identifiable (la **personne concernée**); est réputée être une «personne physique identifiable» une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale [RGPD, Article 4(1)].

Le **responsable du traitement des données** ou **responsable du traitement** est la personne physique ou morale, l'autorité publique, le service ou un autre organisme qui, seul ou conjointement avec d'autres, détermine les finalités et les moyens du traitement [RGPD, Article 4(7)].

Le terme **sous-traitant des données** ou **sous-traitant** est la personne physique ou morale, l'autorité publique, le service ou un autre organisme qui traite des données à caractère personnel pour le compte du responsable du traitement [RGPD, Article 4(8)].

La **pseudonymisation** est définie comme 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. [RGPD, Article 4(5)][5].

L'**anonymisation** est un processus par lequel les **données à caractère personnel** sont modifiées de façon irréversible de telle façon que la personne concernée ne puisse plus être identifiée, directement ou indirectement, que ce soit par le responsable du traitement seul ou en collaboration avec d'autres tiers (ISO/TS 25237:2017)[6].

L'**identificateur** ou **identifiant** est une valeur qui identifie un élément au sein d'un schéma d'identification[7]. Un identificateur unique n'est associé qu'à un seul élément. Il est généralement entendu dans le présent rapport que l'on parle d'identificateurs uniques, habituellement associés aux données à caractère personnel.

Le terme **pseudonyme**, parfois remplacé par **cryptonyme**, est un élément d'information associé à l'identificateur d'une personne ou à tout autre type de données à caractère personnel

[5] Voir également les définitions techniques pertinentes de la pseudonymisation dans [1].
[6] Voir l'analyse approfondie, notamment sur la différence entre la pseudonymisation et l'anonymisation dans [1].
[7] Le Groupe de travail «Article 29» sur la protection des données [31] définit l'identificateur comme étant un élément d'information possédant une relation étroite et particulièrement privilégiée avec un individu afin d'en permettre l'identification. La capacité suffisante d'un identificateur à garantir l'identification dépend du contexte de traitement spécifique des données à caractère personnel. Ainsi, un identificateur peut être un élément d'information simple (par ex. un nom, une adresse électronique, un numéro de sécurité sociale, etc.) mais aussi correspondre à des données plus complexes.

(par ex. des données de localisation). Les pseudonymes peuvent présenter des degrés d'associativité variés avec l'identificateur d'origine[8]. Le degré d'associativité des différents types de pseudonyme est un critère important à prendre en compte lors de l'évaluation de la force d'un pseudonyme, mais également dans la conception des systèmes pseudonymes où un certain degré de corrélation peut être nécessaire (par ex. lors de l'analyse des fichiers journaux pseudonymes ou pour les systèmes de réputation)[9].

La **fonction de pseudonymisation**, désignée par la lettre P, est une fonction chargée de substituer un identificateur Id par un pseudonyme $pseudo$.

Le **secret de pseudonymisation**, désigné par la lettre s, est un paramètre (facultatif) d'une fonction de pseudonymisation P. La fonction P ne peut être évaluée/calculée si s est inconnu.

La **fonction de récupération**, désignée par la lettre R, est une fonction chargée de substituer un pseudonyme $pseudo$ par l'identificateur Id à l'aide du secret de pseudonymisation s. Elle inverse l'action de la fonction de pseudonymisation P.

La **table d'association de pseudonymisation** est une représentation de l'action de la fonction de pseudonymisation. Elle associe chaque identificateur au pseudonyme correspondant. Suivant la fonction de pseudonymisation P utilisée, la table d'association peut correspondre au secret de pseudonymisation ou en faire partie.

L'**entité de pseudonymisation** est l'entité responsable de transformer les identificateurs en pseudonymes à l'aide de la fonction de pseudonymisation. Il peut s'agit d'un responsable du traitement des données, d'un sous-traitant (réalisant la pseudonymisation au nom d'un responsable de traitement), d'un tiers de confiance ou d'une personne concernée, suivant le scénario de pseudonymisation. Il est à noter que, selon cette définition, le rôle de l'entité de pseudonymisation se résume strictement à la mise en œuvre pratique de la pseudonymisation dans un cas de figure spécifique[10]. Cependant, dans le contexte du présent rapport, la responsabilité de l'ensemble du processus de pseudonymisation (et de l'ensemble de l'opération de traitement des données en général) incombe toujours au responsable du traitement.

Le terme **domaine d'identificateurs/domaine de pseudonymes** désigne les domaines desquels sont issus l'identificateur et le pseudonyme. Il peut s'agir de domaines différents ou des mêmes domaines. Il peut s'agir de domaines finis ou infinis.

Le terme **adversaire** désigne une entité cherchant à briser le processus de pseudonymisation et à remettre en lien un pseudonyme (ou un jeu de données pseudonymisé) avec son ou ses titulaires.

Une **attaque par ré-identification** est une attaque ciblant le processus de pseudonymisation, lancée par un adversaire dans le but de ré-identifier le titulaire d'un pseudonyme.

[8] À cette fin, le pseudonyme peut être considéré comme le «déguisement» de l'identificateur d'une personne, qui, suivant le contexte, peut rendre cette personne plus ou moins identifiable.
[9] Pour une discussion plus détaillée sur les degrés d'associativité des pseudonymes, cf. [4].
[10] Notez que, dans la définition de la pseudonymisation établie dans le RGPD (Article 4(5)), aucune mention n'est faite de l'entité qui détient les informations supplémentaires.

3. SCENARIOS DE PSEUDONYMISATION

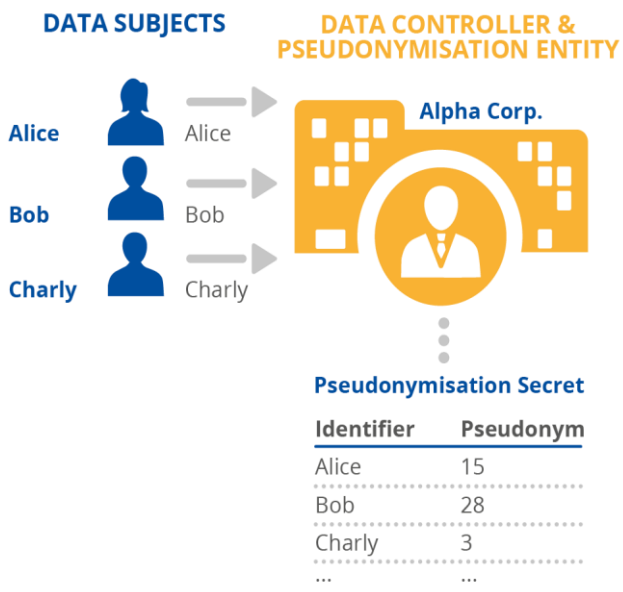
Comme discuté au chapitre [1], la pseudonymisation joue un rôle important au titre du RGPD en tant que mesure de sécurité (art. 32, RGPD), ainsi que dans le contexte de la protection des données dès la conception (art. 25, RGPD). Le bénéfice le plus évident de la pseudonymisation est qu'elle masque l'identité des personnes concernées aux yeux des tierces parties (autres que l'entité de pseudonymisation), dans le contexte d'une opération de traitement des données spécifiques. Le processus de pseudonymisation peut toutefois aller bien au-delà du simple masquage des identités réelles, en favorisant l'objectif de protection des données qu'est la non-retraçabilité [2], c'est-à-dire en réduisant le risque que des données relatives à la vie privée puissent être mise en relation par-delà des domaines de traitement des données différents. D'autre part, la pseudonymisation (étant en soi une technique de minimisation des données) peut contribuer à mettre en application le principe de minimisation des données au titre du RGPD, par exemple dans le cas où le responsable du traitement n'a pas besoin d'accéder à l'identité réelle des personnes concernées, mais uniquement à leurs pseudonymes. Enfin, un autre avantage significatif de la pseudonymisation, qui ne doit pas être sous-estimé, est celui de l'exactitude des données (pour une analyse plus détaillée du rôle de la pseudonymisation, cf. [1]).

Sur la base des bénéfices exposés ci-avant, ce chapitre présente différents scénarios de pseudonymisation communément retrouvés dans la pratique, les acteurs impliqués, ainsi que les objectifs de pseudonymisation spécifiques à chaque cas.

3.1 SCENARIO 1: PSEUDONYMISATION A USAGE INTERNE

L'un des scénarios de pseudonymisation des données courants est celui où les données sont recueillies directement auprès des personnes concernées et pseudonymisées par le responsable du traitement, pour être ensuite traitées en interne.

Graphique 1: Exemple de pseudonymisation - Scénario 1



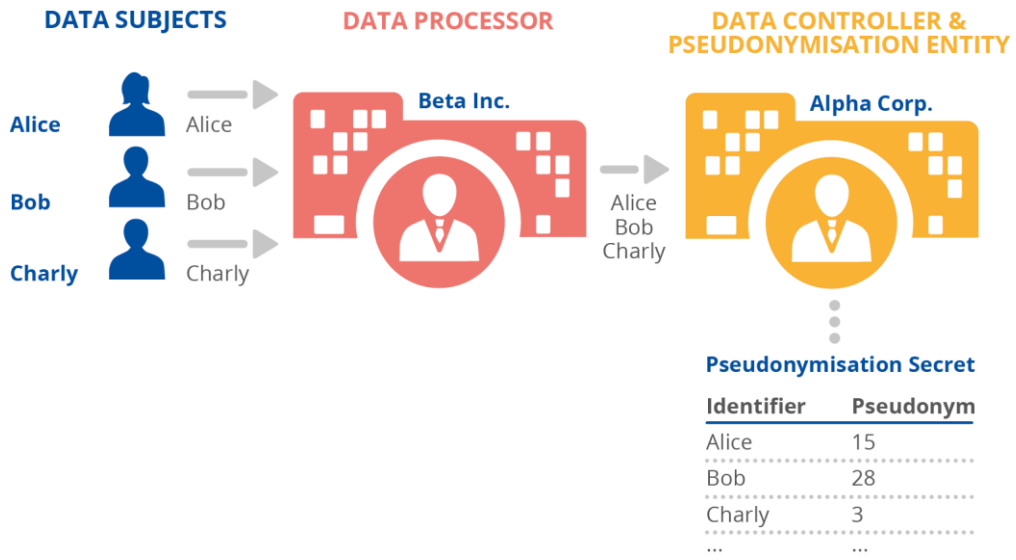
Dans Graphique 1, le responsable du traitement des données (Alpha Corp.) joue le rôle d'entité de pseudonymisation, car il réalise la sélection et l'affectation des pseudonymes pour les identificateurs. Notez que les personnes concernées ne connaissent pas ou ne sont pas nécessairement informées du pseudonyme qui leur est attribué, car le secret de pseudonymisation (dans cet exemple, la table d'association de pseudonymisation) est connu uniquement d'Alpha Corp. Le rôle de la pseudonymisation, dans ce scénario, est d'améliorer la sécurité des données à caractère personnel, soit pour leur usage en interne (par ex. le partage de données entre les différentes unités du responsable de traitement)[11], soit en cas d'incident de sécurité.

3.2 SCENARIO 2: PSEUDONYMISATION IMPLIQUANT UN SOUS-TRAITANT

Ce scénario est une variante du scénario 1, dans lequel un sous-traitant est également impliqué dans le processus, car chargé de recueillir les identificateurs auprès des personnes concernées (au nom du responsable du traitement). La pseudonymisation reste cependant effectuée par le responsable du traitement.

[11] Voir également l'Article (29) du RGPD concernant la notion d'«analyse générale» à usage interne.

Graphique 2: Exemple de pseudonymisation - Scénario 2



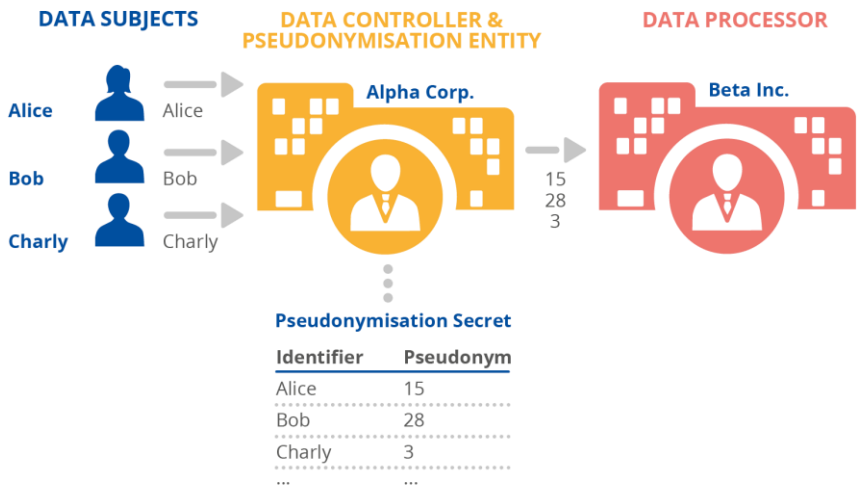
Dans le Graphique 2, un sous-traitant dédié (Beta Inc.) a pour tâche de recueillir les identificateurs auprès des personnes concernées et de transmettre ces informations à un responsable du traitement (Alpha Corp.), qui finalise le processus de pseudonymisation. Le responsable du traitement joue à nouveau le rôle d'entité de pseudonymisation. Ce scénario peut se produire dans le cas d'un fournisseur de services infonuagiques (de «cloud») qui héberge des services de collecte de données au nom du responsable du traitement. Il incombe toujours au responsable du traitement d'effectuer la pseudonymisation des données avant tout autre traitement. Les objectifs de la pseudonymisation sont les mêmes que dans le scénario 1 (à la différence que cette fois, un sous-traitant est également impliqué dans le processus).

3.3 SCENARIO 3: ENVOI DES DONNEES PSEUDONYMISEES A UN RESPONSABLE DU TRAITEMENT

Dans ce scénario, le responsable du traitement des données effectue la pseudonymisation mais le sous-traitant n'est pas impliqué dans le processus, il reçoit simplement les données pseudonymisées de la part du responsable du traitement.

Le graphique 3 illustre un responsable de traitement (Alpha Corp.) recueillant les données et réalisant la pseudonymisation (dans son rôle d'entité de pseudonymisation). La différence avec les scénarios précédents est que, dans ce cas, le responsable du traitement transfère les données pseudonymisées à un sous-traitant (Beta Inc.), par exemple à des fins d'analyse statistique ou de stockage permanent des données. Dans ce scénario, l'objectif de protection offert par la pseudonymisation des données peut se déployer: Beta Inc. n'a pas connaissance des identificateurs des personnes concernées, il est donc incapable de ré-identifier directement les personnes physiques associées aux données (en supposant qu'aucun autre attribut pouvant mener à une ré-identification ne soit mis à la disposition de Beta Inc.). La pseudonymisation permet ainsi de protéger la sécurité des données vis-à-vis du sous-traitant.

Graphique 3: Exemple de pseudonymisation - Scénario 3

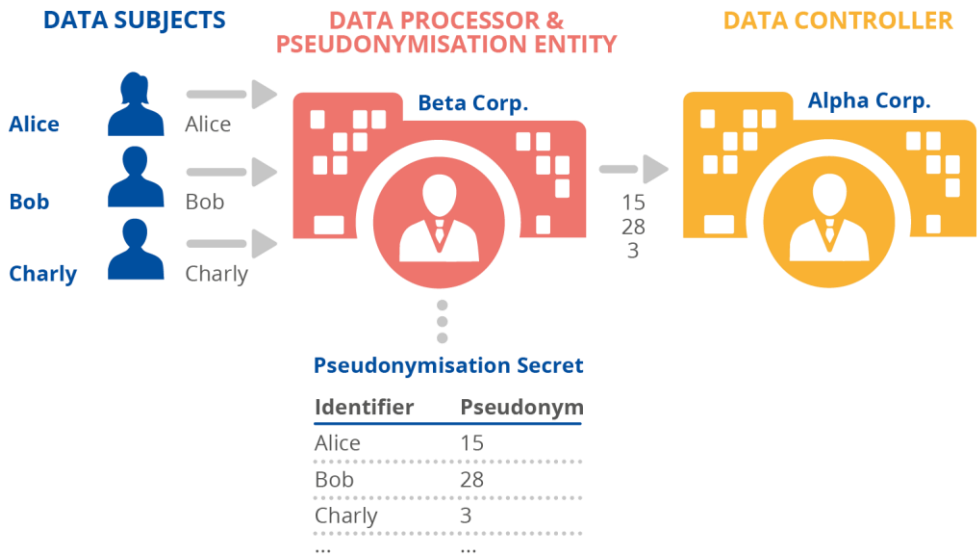


Dans une variante de ce scénario, les données pseudonymisées ne seraient pas transmises à un sous-traitant mais à un autre responsable de traitement (par ex. dans le contexte d'une obligation légale du responsable de traitement initial ou sur toute autre base juridique).

3.4 SCENARIO 4: UN SOUS-TRAITANT COMME ENTITE DE PSEUDONYMISATION

Dans un autre scénario possible, la responsabilité du processus de pseudonymisation est transférée par le responsable du traitement à un sous-traitant (par ex. le fournisseur de services Cloud chargé de gérer le secret de pseudonymisation et/ou d'organiser les différentes infrastructures techniques).

Graphique 4: Exemple de pseudonymisation - Scénario 4



Le Graphique 4 illustre un scénario dans lequel les données à caractère personnel sont envoyées par les personnes concernées à un sous-traitant (Beta Inc), qui se charge d'effectuer la pseudonymisation, jouant le rôle d'entité de pseudonymisation au nom du responsable du traitement (Alpha Corp). Les données pseudonymisées sont ensuite transmises au responsable du traitement. Dans ce scénario spécifique, seules les données pseudonymisées sont

conservées par le responsable du traitement. La sécurité du côté du responsable du traitement en est ainsi améliorée, grâce à une désidentification des données (par ex. en cas de violation de données du côté du responsable). Le responsable du traitement a dans tous les cas la possibilité de ré-identifier les données, par le biais du sous-traitant. La sécurité du côté du sous-traitant revêt dans ce cas une importance cruciale.

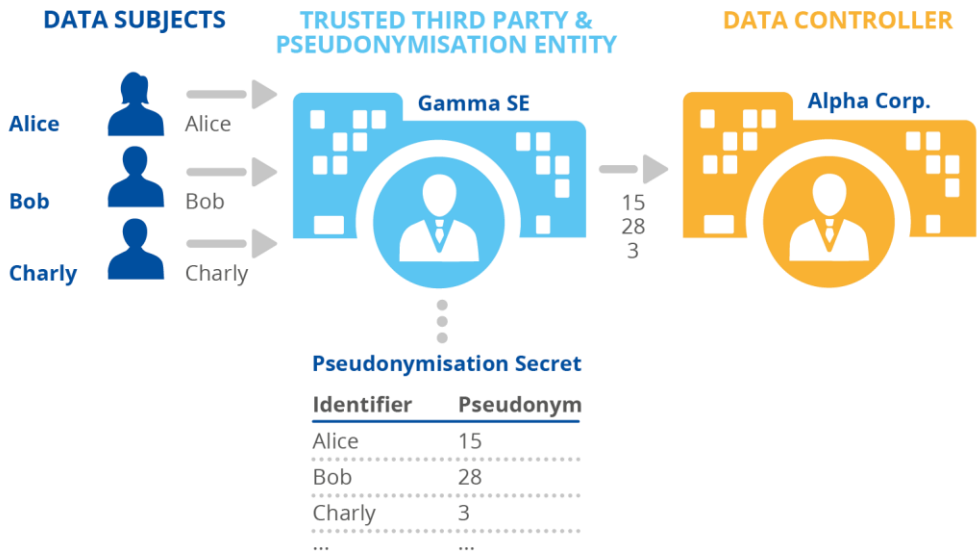
Dans une variante de ce scénario, plusieurs sous-traitants seraient impliqués dans le processus de pseudonymisation en jouant le rôle d'entités de pseudonymisation en série (chaîne de sous-traitants).

3.5 SCENARIO 5: UNE TIERCE PARTIE COMME ENTITE DE PSEUDONYMISATION

Dans ce scénario, la pseudonymisation est effectuée par une tierce partie (et non un sous-traitant) qui se charge ensuite de transférer les données au responsable du traitement. Contrairement au Scénario 4, le responsable du traitement n'a pas accès aux identificateurs des personnes concernées (car la tierce partie n'est pas sous son contrôle).

Le Graphique 5 illustre un scénario dans lequel les données à caractère personnel sont envoyées à une tierce partie (Gamma SE), qui se charge d'effectuer la pseudonymisation, jouant le rôle d'entité de pseudonymisation. Les données pseudonymisées sont ensuite transmises au responsable du traitement (Alpha Corp). Dans ce scénario, le responsable du traitement ne peut pas, directement ou indirectement, mettre en corrélation les enregistrements de données individuels et les personnes concernées. Le niveau de sécurité et de protection des données du côté du responsable du traitement s'en trouve ainsi accru, conformément au principe de minimisation des données. Un tel scénario peut se produire lorsque le responsable du traitement n'a pas besoin d'avoir accès à l'identité des personnes concernées (mais uniquement aux pseudonymes).

Graphique 5: Exemple de pseudonymisation - Scénario 5



Ce scénario peut être particulièrement pertinent dans les cas d'une responsabilité de traitement conjointe, lorsque l'un des responsables réalise la pseudonymisation (jouant le rôle de tiers de confiance, comme illustré au graphique 5) et que l'autre ne fait que recevoir les données pseudonymisées pour traitement ultérieur.

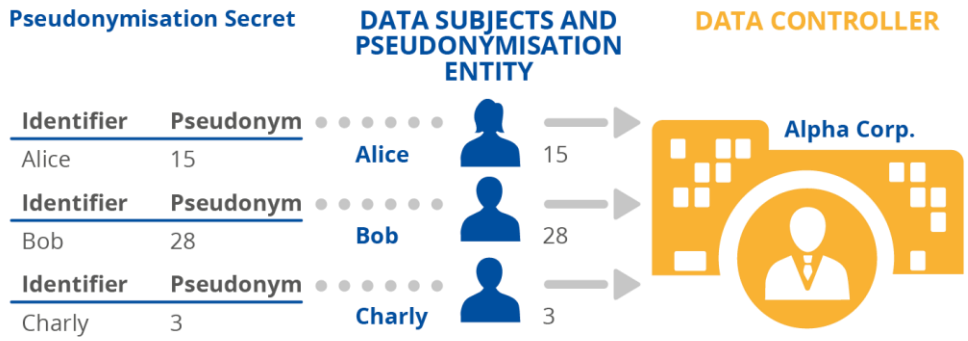
Dans une variante intéressante de ce scénario (qui nécessiterait une analyse plus approfondie), le rôle de tiers de confiance serait réparti entre plusieurs entités, qui ne pourraient créer et rétablir les pseudonymes que conjointement (ou éventuellement sur la base d'un plan de partage secret), afin que le pouvoir ne soit pas réuni dans les mains d'une seule entité.

3.6 SCENARIO 6: UNE PERSONNE CONCERNEE COMME ENTITE DE PSEUDONYMISATION

Dans ce cas particulier de pseudonymisation, les pseudonymes sont créés par les personnes concernées elles-mêmes, dans le cadre du processus global de pseudonymisation.

Comme illustré dans l'exemple du graphique 6, chaque individu génère son propre pseudonyme, puis utilise celui-ci pour transférer ses données[12].

Graphique 6: Exemple de pseudonymisation - Scénario 6



Ce type de modèle de pseudonymisation peut concerner l'utilisation de la clé publique d'une paire de clés dans des systèmes de «blockchain» (chaîne de blocs), de type Bitcoin, pour produire le pseudonyme. L'objectif de la pseudonymisation dans ce scénario est que le responsable du traitement ne connaisse pas[13] les identificateurs des personnes concernées, et que seules ces dernières contrôlent le processus de pseudonymisation. Bien entendu, la responsabilité du processus global de pseudonymisation incombe toujours au responsable du traitement[14]. Ce scénario est également conforme au principe de minimisation des données et peut être appliqué aux situations dans lesquelles le responsable du traitement n'a pas besoin d'avoir accès aux identificateurs originaux (c'est-à-dire lorsque les pseudonymes sont suffisants pour exécuter l'opération de traitement des données souhaitée).

[12] Notez que le pseudonyme peut être le même ou varier entre les différents services/applications (cf. chapitre 5).
[13] Au sens où le responsable du traitement n'a accès à aucun secret de pseudonymisation qui lui permettrait d'effectuer une ré-identification directe.
[14] Voir également l'Article 11 du RGPD, qui peut s'appliquer à ce cas de figure.

4. MODELE D'ADVERSAIRE

Comme mentionné au chapitre 3, le principal objectif de la pseudonymisation est de limiter les possibilités de mise en relation entre un jeu de données pseudonymisé et le titulaire des pseudonymes, de manière à protéger l'identité des personnes concernées. Ce type de protection est généralement destiné à entraver les actions d'un «adversaire» cherchant à effectuer une attaque de ré-identification.

Ce chapitre présente les différents modèles d'adversaire possibles, ainsi que les différents types d'attaque par ré-identification qui touchent le processus de pseudonymisation. Pour ce faire, les notions d'adversaire interne (ou initié) et externe seront expliquées, ainsi que leur rôle potentiel dans les scénarios de pseudonymisation présentés plus haut. Il est essentiel de comprendre ces sujets pour pouvoir analyser l'utilisation des techniques de pseudonymisation dans les prochains chapitres.

4.1 ADVERSAIRES INTERNES

Selon sa définition courante dans la terminologie informatique, un initié, ou adversaire/attaquant interne, est une personne malveillante possédant des connaissances, des capacités ou des autorisations spécifiques (vis-à-vis de l'élément ciblé)[15]. Dans le contexte de la pseudonymisation, cela implique que l'adversaire est en capacité d'obtenir des informations sur le secret de pseudonymisation et/ou d'autres informations pertinentes.

Par exemple, dans le contexte des scénarios présentés au chapitre 3, la menace interne proviendrait du responsable du traitement (par ex. un employé du responsable du traitement), dans les scénarios 1, 2, 3 et 4. Cette menace pourrait également provenir du sous-traitant (par ex. un employé malveillant du sous-traitant) dans les scénarios 2 et 4. Enfin, dans le scénario 5, l'adversaire interne pourrait provenir du tiers de confiance (qui joue dans ce scénario le rôle de l'entité de pseudonymisation). Par défaut, les tierces parties ayant légitimement accès aux données à caractère personnel (par ex. une autorité répressive ou une autorité de contrôle) ne sont pas considérées comme des adversaires[16].

4.2 ADVERSAIRES EXTERNES

À l'inverse de l'adversaire interne, l'adversaire externe ne bénéficie pas d'un accès direct au secret de pseudonymisation ou à toute autre information pertinente. Cependant, ce type d'adversaire peut avoir accès à un jeu de données pseudonymisé et avoir la capacité à exécuter le processus de pseudonymisation en vue de saisir arbitrairement des valeurs de données qu'il aura lui-même choisies (par ex. en ayant accès à une implémentation en boîte noire de la fonction de pseudonymisation ou en étant capable de forcer l'entité de pseudonymisation à pseudonymiser des valeurs arbitraires). L'objectif de l'adversaire externe est d'accroître le volume d'informations qu'il possède sur les jeux de données pseudonymisés, par exemple en découvrant l'identité cachée derrière un pseudonyme donné (et en obtenant

[15] Selon le CERT Insider Threat Center, du Software Engineering Institute (SEI) de la Carnegie Mellon University, une menace interne se définit comme la capacité potentielle d'un individu, bénéficiant ou ayant bénéficié d'un accès autorisé aux ressources d'une organisation, à utiliser cet accès de façon malveillante ou involontaire pour agir d'une manière qui affecterait de façon négative cette organisation, https://insights.sei.cmu.edu/insider-threat/2017/03/cert-definition-of-insider-threat---updated.html

[16] Il est toutefois à noter que la légitimité d'un tel accès doit être questionnée dans les cas où le principe de minimisation des données n'est pas respecté (par ex. lorsqu'une autorité de contrôle a accès au secret de pseudonymisation au lieu de recevoir explicitement les données à caractère personnel qu'elle est en droit de recevoir). Ces cas de figure peuvent s'apparenter à un modèle d'adversaire interne, car la tierce partie bénéficie d'un accès d'initié légitime, tout comme l'entité de pseudonymisation.

davantage d'informations sur cette identité à partir des données supplémentaires issues du jeu de données de ce pseudonyme).

Dans le contexte des scénarios présentés au chapitre 3, doit être par définition considéré comme un adversaire externe tout acteur agissant de façon malveillante dans l'ensemble des scénarios et ne faisant pas partie de l'entité de pseudonymisation ou n'agissant pas au nom de celle-ci. Un responsable du traitement des données (malveillant) peut être un adversaire externe dans les scénarios 5 et 6. Un sous-traitant (malveillant) peut également être un adversaire externe dans le scénario 3.

4.3 OBJECTIFS D'UNE ATTAQUE CONTRE LA PSEUDONYMISATION

Suivant le contexte et la méthode de pseudonymisation appliquée, l'adversaire peut avoir plusieurs objectifs vis-à-vis des données pseudonymisées, par exemple la récupération du secret de pseudonymisation, la ré-identification complète ou la discrimination. Alors que la plupart des exemples présentés dans les paragraphes suivants illustrent des adversaires cherchant à découvrir l'identité réelle des personnes concernées, il convient de noter qu'une attaque réussie ne se résume pas à une ingénierie inverse ou rétro-ingénierie, mais inclut aussi la capacité à cibler ou isoler un individu spécifique dans un groupe (même si son identité réelle n'est pas révélée).

4.3.1 Secret de pseudonymisation

Dans ce cas de figure, l'adversaire cherche à découvrir le secret de pseudonymisation (lorsqu'il en existe un). Cette attaque est le plus dangereuse, car lorsqu'il est en possession du secret de pseudonymisation, l'adversaire est à même de ré-identifier tout pseudonyme du jeu de données (ré-identification ou discrimination), ainsi que de réaliser d'autres processus de pseudonymisation sur le jeu de données.

4.3.2 Ré-identification complète

Lorsque le but de l'attaque est une ré-identification complète, l'adversaire souhaite rétablir le lien entre un ou plusieurs pseudonymes et l'identité de leur titulaire. Ce type de menace a déjà été largement discuté dans les publications spécialisées (cf. ex. [3] [4] [5]).

L'attaque par ré-identification la plus grave concerne la ré-identification de l'ensemble des pseudonymes. L'adversaire peut appliquer deux stratégies pour atteindre cet objectif: récupérer individuellement chaque identificateur à l'aide du pseudonyme correspondant; ou récupérer le secret de pseudonymisation (cf. 4.3.1). La forme la moins grave d'une attaque de ré-identification complète a lieu lorsque l'adversaire ne peut ré-identifier qu'un sous-ensemble des pseudonymes du jeu de données. Par exemple, si l'attaque porte sur un jeu de données pseudonymisées contenant les notes des étudiants d'une université. Chaque entrée du jeu de données contient un premier pseudonyme correspondant à l'identité d'un étudiant (prénom et nom de famille) et un second pseudonyme spécifiant le sexe de l'étudiant (par ex. un nombre impair pour les hommes et un nombre pair pour les femmes). L'attaque par ré-identification complète sera un succès si l'adversaire parvient à récupérer le prénom, le nom et le sexe de l'étudiant.

4.3.3 Discrimination

L'objectif de l'attaque par discrimination est d'identifier les propriétés (au minimum une) du titulaire d'un pseudonyme. Même si ces propriétés ne permettent pas directement de révéler l'identité du titulaire du pseudonyme, elles peuvent suffire à le discriminer, c'est-à-dire à le cibler ou le singulariser, d'une manière ou d'une autre.

Si l'on reprend l'exemple des notes des étudiants, le jeu de données peut contenir comme pseudonymes deux chiffres pairs parmi de nombreux chiffres impairs. Les chiffres pairs correspondent aux étudiantes et les chiffres impairs aux étudiants (ce fait est connu de l'attaquant). Les deux étudiantes correspondant aux chiffres pairs ont eu un résultat de 100 % à l'examen final. Imaginons ensuite qu'aucun autre étudiant n'ait obtenu un résultat de 100 % dans le jeu de données pseudonymisé. Si l'adversaire obtient des informations supplémentaires, par exemple le fait qu'un étudiant a obtenu un résultat de 100 % dans cette matière, il saura immédiatement qu'il s'agit d'une femme. À l'inverse, si l'adversaire sait qu'une étudiante suit ce cours, il saura immédiatement qu'elle a obtenu un résultat de 100 % à l'examen. Il est important de comprendre que l'adversaire ne découvre pas ici l'identité du titulaire d'un pseudonyme, mais seulement qu'il prend connaissance de certaines propriétés le concernant (par ex. son sexe ou sa note d'examen). Étant donné que plusieurs étudiants partagent la même combinaison de propriétés, l'adversaire n'est pas capable de relier un enregistrement de données individuel au titulaire d'un pseudonyme particulier. Toutefois, ces informations supplémentaires peuvent suffire à effectuer la discrimination recherchée par l'adversaire ou bien celui-ci peut les utiliser dans une attaque ultérieure fondée sur les connaissances contextuelles et visant à découvrir l'identité cachée derrière un pseudonyme.

4.4 PRINCIPALES TECHNIQUES D'ATTAQUE

Il existe trois techniques d'attaque principales destinées à «casser» la fonction de pseudonymisation: les attaques par force brute (recherche exhaustive), les attaques par dictionnaire et les attaques par conjecture[17]. L'efficacité de ces attaques dépend de plusieurs paramètres, notamment:

- La quantité de données que contient le pseudonyme concernant son titulaire (personne concernée).
- Les connaissances supplémentaires que possède l'adversaire.
- La taille du domaine d'identificateurs.
- La taille du domaine de pseudonymes.
- Le choix et la configuration de la fonction de pseudonymisation utilisée (ce qui inclut la taille du secret de pseudonymisation).

Ces techniques d'attaque sont présentées brièvement ci-après.

4.4.1 Attaque par force brute

L'applicabilité de cette technique d'attaque est conditionnée par la capacité de l'adversaire à traiter (calculer) la fonction de pseudonymisation (lorsqu'il n'existe pas de secret de pseudonymisation) ou par sa capacité d'accéder à une implémentation «en boîte noire» de la fonction de pseudonymisation. Suivant l'objectif de l'attaque, d'autres conditions peuvent s'appliquer. Si l'attaque par force brute a pour but d'effectuer une ré-identification complète (restauration de l'identité originale), le domaine d'identificateurs doit être de type fini et de taille relativement petite. Pour chaque pseudonyme qu'il rencontre, l'adversaire peut tenter de récupérer l'identificateur d'origine en appliquant la fonction de pseudonymisation sur chaque valeur du domaine d'identificateurs, jusqu'à obtenir une correspondance.

[17] Il est à noter que, comme mentionné précédemment dans ce rapport, d'autres attributs (à l'exception du pseudonyme et des données pseudonymisées) peuvent être utilisés pour identifier un individu. Reportez-vous au chapitre 8 pour plus d'informations à ce sujet.

Tableau 1: Pseudonymisation du mois de naissance

Mois de naissance	Pseudonyme	Mois de naissance	Pseudonyme
Janvier	281	**Juillet**	299
Février	269	**Août**	285
Mars	288	**Septembre**	296
Avril	291	**Octobre**	294
Mai	295	**Novembre**	307
Juin	301	**Décembre**	268

Prenons par exemple la pseudonymisation du mois de naissance dans un jeu de données. La taille du domaine d'identificateurs est de 12, ce qui permet à l'adversaire de passer rapidement en revue l'ensemble des possibilités. Les pseudonymes associés à chaque mois sont analysés dans cet exemple sous la forme de la somme du code ASCII des trois premières lettres du nom du mois (la première étant une majuscule). Imaginons que l'adversaire ait trouvé le pseudonyme 301. Il peut alors appliquer, pour chaque mois de naissance, la fonction de pseudonymisation jusqu'à trouver le mois auquel correspond la valeur 301. Le Tableau 1 illustre les calculs effectués par l'adversaire pour ré-identifier le pseudonyme 301, avec pour résultat la table d'association de la fonction de pseudonymisation.

Bien évidemment, la taille du domaine d'identificateurs est essentielle pour le succès de ce type d'attaque. Dans le cas de domaines d'identificateurs de petite taille, comme celui de l'exemple ci-dessus, réussir une attaque par force brute est un jeu d'enfant. Toutefois, lorsque le domaine d'identificateurs est de taille infinie, l'attaque par force brute devient irréalisable. Si la taille du domaine d'identificateurs est trop importante, il est extrêmement difficile pour un adversaire d'effectuer une ré-identification complète, ce qui lui laisse la possibilité d'une attaque par discrimination.

Dans un tel cas, l'adversaire peut envisager de viser un sous-domaine du domaine d'identificateurs, pour lequel il sera capable de traiter tous les pseudonymes. Revenons à l'exemple du tableau 1, sur un domaine supposé de petite taille. Imaginons que l'adversaire souhaite discriminer les individus dont le mois de naissance commence par la lettre J, par rapport à ceux dont le mois de naissance commence par une autre lettre. Ce sous-domaine contient les mois de janvier, juin et juillet. L'adversaire peut réaliser une recherche exhaustive sur ce sous-domaine en analysant les pseudonymes correspondant à Janvier, Juin et Juillet. S'il trouve un pseudonyme différent de 281, 301 et 299, il sait alors que le mois de naissance commence par une lettre autre que le J.

Lorsqu'un secret de pseudonymisation est utilisé, même le plus petit des domaines d'identificateurs ne permet pas de réaliser une telle attaque, car l'attaquant est incapable d'exécuter la fonction de pseudonymisation, dans la mesure où il n'existe pas d'accès à une implémentation en «boîte noire» pour cette fonction. Dans un tel cas, il est possible d'appliquer une attaque par force brute sur l'espace global des secrets de pseudonymisation, c'est-à-dire que l'attaquant vérifie de façon exhaustive tous les secrets possibles et, pour chacun d'eux, calcule la fonction de récupération. Cette attaque peut réussir si l'adversaire conjecture de façon correcte le secret de pseudonymisation, et ce quelle que soit la taille du domaine d'identificateurs. Par conséquent, pour contrecarrer ce type d'attaque, le nombre de secrets de pseudonymisation possibles doit être suffisamment élevé pour rendre l'attaque quasiment impossible.

4.4.2 Attaque par dictionnaire

L'attaque par dictionnaire est une variante optimisée de l'attaque par force brute, qui permet de réduire les coûts de calcul. L'adversaire doit faire face à un nombre important de pseudonymes pour son attaque par ré-identification complète ou discrimination. Il pré-traite un (énorme) ensemble de pseudonymes et enregistre les résultats dans un dictionnaire. Chaque entrée de ce dictionnaire contient un pseudonyme et l'identificateur ou les informations associé(es). Ensuite, à chaque fois qu'il doit ré-identifier un pseudonyme, l'adversaire effectue une recherche dans ce dictionnaire. Cette recherche présente le coût de pré-calcul d'une recherche exhaustive et stocke le résultat dans une mémoire de gros volume. La ré-identification d'un pseudonyme n'a qu'un coût de calcul équivalent à celui d'une recherche dans le dictionnaire. L'attaque par dictionnaire consiste donc principalement à calculer et stocker la table de mappage. Des compromis temps-mémoire sont même possibles en utilisant des tables Hellman [6] ou des «rainbow tables» (tables arc-en-ciel) [7] afin d'élargir encore la portée. Cependant, il existe des variantes spécifiques de cette attaque qui exigent une connaissance approfondie du comportement de la fonction de pseudonymisation. De telles attaques peuvent réussir, même sur des domaines infinis.

4.4.3 Attaque par maximum de vraissemblance

Ce type d'attaque utilise des connaissances supplémentaires (telles que la distribution de probabilité ou toute autre information collatérale) que l'adversaire peut avoir sur certains des titulaires de pseudonyme (ou tous), la fonction de pseudonymisation ou le jeu de données. Les attaques par recherche exhaustive et par dictionnaire supposent implicitement que tous les identificateurs ont la même probabilité ou la même fréquence d'occurrence. Cependant, certains identificateurs peuvent être plus fréquents que d'autres. Exploiter les caractéristiques statistiques des identificateurs est une pratique appelé «conjecture» [8] [9] [10], qui est largement répandue dans la communauté de piratage des mots de passe. Il est important de comprendre que l'attaque par conjecture peut être appliquée même lorsque le domaine d'identificateurs est de très grande taille. L'adversaire n'a pas nécessairement à accéder à la fonction de pseudonymisation, car une discrimination est possible en effectuant simplement une analyse de fréquence sur les pseudonymes observés.

Prenons pour exemple un cas de figure dans lequel les pseudonymes correspondent à des prénoms. Le domaine «prénoms» est difficile à explorer dans sa totalité. Cependant, l'adversaire sait quels sont les prénoms les plus populaires (Tableau 2). Il peut donc lancer sur le domaine une recherche exhaustive ou par dictionnaire sur la base des prénoms les plus populaires, et effectuer sa discrimination.

Tableau 2: Liste des prénoms les plus populaires

Prénoms les plus populaires					
Bob	Alice	Charlie	Eve	Robert	Marie

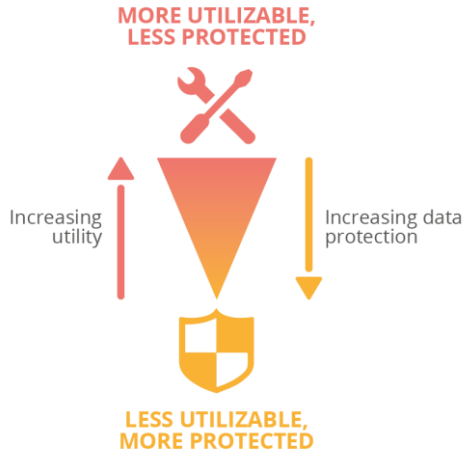
Imaginons un cas similaire, mais dans lequel le domaine d'identificateurs est de taille infinie. Il est possible de définir un sous-domaine des identificateurs compris dans le jeu de données. S'il est capable d'identifier par conjecture ce sous-domaine, l'adversaire peut alors lancer une recherche exhaustive (cf. chapitre 6 pour des cas d'utilisation pertinents sur la pseudonymisation des adresses électroniques). Suivant la quantité d'informations contextuelles ou de métadonnées que l'adversaire a en sa possession, ainsi que la quantité d'informations corrélables trouvées dans le jeu de données pseudonymisées, ce type d'attaque peut permettre de découvrir l'identité d'un titulaire de pseudonyme individuel, de plusieurs titulaires ou de tous

les titulaires du jeu de données. Dans le cas des jeux de données de petite taille, tout particulièrement, ce type d'attaque peut être réalisable avec un taux de succès élevé.

4.5 FONCTIONNALITE ET PROTECTION DES DONNEES

Suivant la fonction de pseudonymisation choisie, un pseudonyme peut contenir certaines informations sur l'identificateur original. Par conséquent, il existe pour chacun de ces types de pseudonyme le risque d'être la cible d'une attaque de ré-identification, telle que celles décrites précédemment. Par exemple, un attaquant possédant une connaissance contextuelle suffisante peut être capable de mettre en corrélation le pseudonyme avec son identificateur, par conjecture.

Graphique 7 : Fonctionnalité et protection des données



Cependant, dans de nombreux cas, les informations supplémentaires sur l'identificateur original contenues dans le pseudonyme sont conservées à des fins de mise en corrélation au sein des pseudonymes eux-mêmes, pour une opération ultérieure qui sera réalisée par un responsable de traitement légitime. Par exemple, un pseudonyme peut afficher directement l'année de naissance de la personne concernée (par ex. «AAAA-1999»). Il est ainsi possible de classer les pseudonymes sur la base de l'année de naissance, par exemple pour faire référence à l'âge, au statut légal (enfant ou adulte), aux conditions de vie (enfant scolarisé /actif /retraité), etc. Il peut s'agir d'une fonctionnalité volontaire de la fonction de pseudonymisation utilisée, qui permet aux responsables du traitement d'effectuer une forme de classification y compris sur des données pseudonymisées.

Le choix de la fonction de pseudonymisation peut donc clairement apporter une fonctionnalité (une utilité) aux pseudonymes créés, en tenant toutefois compte de la perte de protection potentielle qu'une telle approche de pseudonymisation entraîne. Il est donc important de trouver un compromis entre la fonctionnalité et la protection des données (cf. graphique 7). Lors de l'application de la pseudonymisation en conditions réelles, il est essentiel d'étudier soigneusement ce compromis, de manière à optimiser la fonctionnalité au vu des objectifs souhaités, tout en gardant un niveau de protection des titulaires de pseudonymes (personnes concernée) aussi élevé que possible.

5. TECHNIQUES DE PSEUDONYMISATION

À présent que les modèles d'adversaire et les types d'attaques ont été définis, le présent chapitre propose une brève présentation des stratégies et des techniques de pseudonymisation les plus courantes aujourd'hui. Pour une analyse plus détaillée des primitives cryptographiques, cf. [1].

Sur le principe, une fonction de pseudonymisation a pour mission d'associer (mapper) des identificateurs et des pseudonymes. Il existe une exigence fondamentale pour une fonction de pseudonymisation. Prenons pour exemple deux identificateurs $Id1$ et $Id2$, ainsi que les pseudonymes correspondants, $pseudo1$ et $pseudo2$. La fonction de pseudonymisation doit vérifier que $pseudo1$ est différent de $pseudo2$. Dans le cas contraire, la récupération de l'identificateur serait ambiguë: l'entité de pseudonymisation ne pourrait déterminer si $pseudo1$ correspond à $Id1$ ou $Id2$. Cependant, un même identificateur Id peut être associé à plusieurs pseudonymes ($pseudo1$,$pseudo2$, …) tant qu'il est possible pour l'entité de pseudonymisation d'inverser l'opération. Dans tous les cas, selon la définition de la pseudonymisation (cf. chapitre 2), il existe des informations supplémentaires qui permettent d'associer les pseudonymes aux identificateurs originaux; il s'agit du secret de pseudonymisation. La forme la plus simple d'un secret de pseudonymisation est la table de mappage de pseudonymisation.

Nous vous proposons dans les sections suivantes une définition des principales options disponibles pour pseudonymiser un identificateur unique. Vous seront ensuite présentées les différentes stratégies de pseudonymisation, avec une comparaison de leurs caractéristiques d'implémentation, ainsi que les différents critères qu'un responsable de traitement peut utiliser pour choisir une technique de pseudonymisation appropriée. Enfin, nous discuterons des possibilités de récupération par l'entité de pseudonymisation.

5.1 PSEUDONYMISATION A IDENTIFICATEUR UNIQUE

Les paragraphes suivants présentent les différentes approches possibles pour la pseudonymisation d'un identificateur unique, ainsi que les avantages et contraintes associés.

5.1.1 Compteur

Le compteur est la fonction de pseudonymisation la plus simple qui existe. Chaque identificateur est substitué par un nombre, défini par un compteur unitone. Une valeur initiale s est tout d'abord définie sur 0 (par exemple), puis incrémentée. Il est essentiel que les valeurs produites par le compteur ne se répètent jamais, pour éviter toute ambiguïté.

Les avantages du compteur reposent sur sa simplicité, qui en font un excellent candidat pour les jeux de données de petite taille et sans complexité. En termes de protection des données, le compteur fournit des pseudonymes sans connexion avec l'identificateur original (bien que le caractère séquentiel de cette méthode puisse fournir des informations sur l'ordre des données au sein du jeu de données). Cette solution, cependant, peut présenter des problèmes en termes d'implémentation et d'évolutivité dans le cas de jeux de données de grande taille, plus sophistiqués, car il est nécessaire de stocker l'ensemble de table de mappage de pseudonymisation.

5.1.2 Générateur de nombres aléatoires (GNA)

Le générateur de nombres aléatoires est un mécanisme capable de produire, dans un jeu de données, des valeurs ayant une probabilité égale d'être sélectionnées au sein de la population totale de possibilités, offrant donc un comportement imprévisible[18]. Cette approche est similaire à celle du compteur, à la différence qu'un nombre aléatoire, et non séquentiel, est attribué à chaque identificateur. Deux options sont disponibles pour créer cette association (mappage): un générateur de nombres réellement aléatoires ou un générateur pseudo-aléatoire cryptographique (cf. [11] pour une définition exacte). il est à noter que, dans les deux cas, des collisions peuvent se produire sans une vigilance particulière[19]. Le terme de collision désigne l'association de deux identificateurs à un même pseudonyme. La probabilité qu'une telle collision se produise repose sur le célèbre «paradoxe des anniversaires» [12].

Le GNA offre une protection des données robuste, car, contrairement au compteur, un numéro aléatoire est utilisé pour créer chaque pseudonyme, ce qui rend plus difficile l'extraction d'informations relatives à l'identificateur initial, à moins que la table de mappage n'ait été compromise. Les collisions peuvent être un problème, comme précédemment mentionné, de même que l'évolutivité (le stockage de l'ensemble de la table de mappage est nécessaire), suivant le scénario d'implémentation concerné.

5.1.3 Fonction de hachage cryptographique

La fonction de hachage cryptographique utilise comme entrée des chaînes de longueur arbitraire et les «mappe» à des sorties de longueur fixe [13] [14]. Elle offre ainsi les propriétés suivantes:

- Fonctionnement unidirectionnel: il est techniquement impossible de trouver une entrée correspondant à une sortie prédéfinie.
- Absence de collision: il est techniquement impossible de trouver deux entrées distinctes correspondant à la même sortie.

La fonction hachage cryptographique est appliquée directement à l'identificateur, afin d'obtenir le pseudonyme associé: $Pseudo = H(Id)$. Le domaine des pseudonymes dépend de la longueur de l'empreinte produite par la fonction.

Comme nous l'avons mentionné au chapitre [1], bien qu'elle puisse grandement contribuer à protéger l'intégrité des données, la fonction de hachage est généralement considérée comme une technique de pseudonymisation faible car elle est sujette aux attaques par force brute et par dictionnaire. Des exemples spécifiques de cette faiblesse sont présentés dans aux chapitres 6 et 7 ci-dessous.

5.1.4 Code d'authentification de message (MAC)

Cette primitive peut être considérée comme une fonction de hachage par clé. Elle est très semblable à la solution précédente, excepté qu'une clé secrète est introduite pour générer le pseudonyme. Sans connaître cette clé, il est impossible de mapper les identificateurs et leurs pseudonymes. La fonction HMAC [15] [16] est de loin la méthode de code d'authentification de message la plus populaire pour les protocoles internet.

Comme mentionné au chapitre [1], la fonction MAC est généralement considérée comme une technique de pseudonymisation robuste du point de vue de la protection des données, car il est

[18] Notez qu'il est possible d'utiliser une séquence de caractères plutôt qu'un nombre, si on le souhaite.
[19] Le risque de collision peut être réduit à quantité négligeable si l'on génère des pseudo-nombres de grande taille (par ex. de 100 chiffres de longueur).

impossible de retrouver l'identificateur à partir du pseudonyme, tant que la clé n'a pas été compromise. Il existe différentes variantes de cette méthode, adaptées aux exigences de l'entité de pseudonymisation en termes de fonctionnalité et d'évolutivité (plus d'exemples spécifiques aux chapitres 6 et 7 ci-dessous).

5.1.5 Chiffrement

Le présent rapport s'intéresse principalement au chiffrement symétrique (déterministe), et en particulier au chiffrement par bloc, tel que l'AES (norme de chiffrement avancé) et ses modes d'opération [11]. Le chiffrement par bloc est utilisé pour chiffrer un identificateur à l'aide d'une clé secrète (clé privée), qui sert à la fois de secret de pseudonymisation et de secret de récupération. L'application d'un chiffrement par bloc en pseudonymisation exige de gérer la taille des blocs. Les identificateurs peuvent être de taille plus petite ou plus grande que le bloc d'entrée de chiffrement. Si, l'identificateur est plus petit, un remplissage [11] doit être envisagé. Si l'identificateur est plus grand que le bloc d'entrée, deux options sont possibles pour résoudre le problème: l'identificateur peut être compressé pour atteindre une taille inférieure à celle du bloc; si la compression n'est pas possible, un mode d'opération [de type CTR (chiffrement basé sur un compteur)] peut être appliqué. Cependant, cette dernière option implique de gérer un paramètre supplémentaire, le vecteur d'initialisation.

Comme mentionné au chapitre [1], le chiffrement peut également être une technique de pseudonymisation robuste, dont plusieurs propriétés sont similaires à celles du code d'authentification des messages (MAC). Des exemples spécifiques sont présentés aux chapitres 6 et 7.

Bien que ce rapport s'intéresse principalement aux méthodes de chiffrement déterministes, le chiffrement probabiliste est une autre possibilité, qui peut en particulier être utilisée dans les cas où plusieurs pseudonymes doivent être dérivés pour un même identificateur (cf. également la stratégie de pseudonyme entièrement aléatoire ci-dessous). Pour en savoir plus, reportez-vous au chapitre [1].

5.2 STRATEGIES DE PSEUDONYMISATION

Bien que le choix de la technique de pseudonymisation soit un critère essentiel, la stratégie (ou mode) de mise en œuvre de la pseudonymisation est tout aussi importante pour son application pratique.

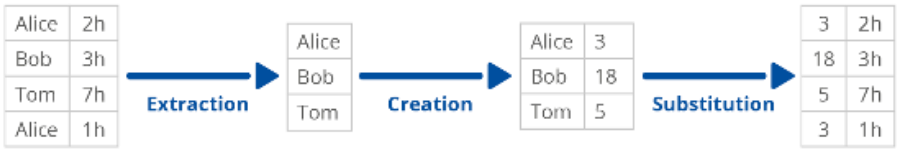
Cette section traite de la problématique générale de la pseudonymisation d'une base de données ou de tout document contenant des identificateurs k. Prenons par exemple un identificateur Id qui apparaît plusieurs fois dans deux jeux de données, A et B. Après la pseudonymisation, l'identificateur Id est substitué par le biais de l'une des stratégies suivantes: pseudonymisation déterministe, pseudonymisation par randomisation de documents et pseudonymisation entièrement aléatoire.

5.2.1 Pseudonymisation déterministe

Dans toutes les bases de données et à chaque fois qu'il apparaît, Id est toujours remplacé par le même pseudonyme, $pseudo$. Cette association est la même au sein d'une base de données unique et entre différentes bases de données. La première étape pour mettre en œuvre cette

stratégie consiste à extraire la liste des identificateurs uniques de la base de données. Cette liste est ensuite mappée avec les pseudonymes, puis les identificateurs originaux sont finalement substitués à ceux-ci dans la base de données (cf. Graphique 8).

Graphique 8: Pseudonymisation déterministe

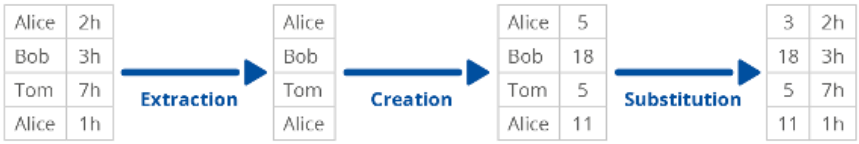


Toutes les techniques mentionnées au chapitre 5.1 peuvent être directement utilisées pour mettre en œuvre une pseudonymisation déterministe.

5.2.2 Pseudonymisation par randomisation de documents

À chaque fois que Id apparaît dans une base de données, il y est substitué par un pseudonyme différent, , $pseudo_2$,...). Cependant, Id est toujours mappé au même recueil de ($pseudo_1$, $pseudo_2$ dans les jeux de données A et B.

Graphique 9: Pseudonymisation par randomisation de documents



Dans le cas présent, la pseudonymisation est identique uniquement entre des bases de données différentes. Cette fois, la table de mappage est créée sur la base de tous les identificateurs contenus dans la base de données. Chaque occurrence d'un identificateur donné (par ex., Alice dans le Graphique 9) est traitée de façon indépendante.

5.2.3 Pseudonymisation entièrement aléatoire

Enfin, pour toutes les occurrences de Id au sein d'une base de données A ou B, Id est remplacé par un pseudonyme différent ($pseudo_1$, $pseudo_2$.On parle alors de pseudonymisation entièrement aléatoire. Cette stratégie peut être considérée comme une variante de la pseudonymisation par randomisation de documents. Ces deux stratégies ont en réalité le même comportement lorsqu'elle sont appliquées à un document unique. Toutefois, si le même document est pseudonymisé deux fois par une pseudonymisation complètement aléatoire, deux résultats différents peuvent être obtenus. Avec la pseudonymisation par randomisation de documents, le même résultat aurait été obtenu deux fois. En d'autres termes, avec une pseudonymisation par randomisation de documents, la fonction aléatoire est sélective (par ex. uniquement pour Alice), alors qu'avec une pseudonymisation entièrement aléatoire, la fonction aléatoire est globale (elle s'applique à tous les enregistrements).

5.3 CHOIX D'UNE TECHNIQUE ET D'UNE STRATEGIE DE PSEUDONYMISATION

Le choix d'une technique et d'une stratégie de pseudonymisation repose sur différents paramètres, principalement le degré de protection des données et la fonctionnalité du jeu de données pseudonymisé souhaités par l'entité de pseudonymisation. En termes de protection, tel que discuté dans les sections précédentes, le générateur de nombres aléatoires, le code d'authentification de message et le chiffrement sont les techniques les plus sûres, car elles contrecarrent par conception les attaques par recherche exhaustive, par dictionnaire et par conjecture. Toutefois, les exigences de l'entité de pseudonymisation en matière de fonctionnalité peuvent l'amener à adopter une combinaison de plusieurs approches différentes ou des variantes de l'approche sélectionnée. De même, en ce qui concerne les stratégies de pseudonymisation, la pseudonymisation complètement aléatoire offre le meilleur niveau de

protection, mais empêche toute comparaison entre les bases de données. Les fonctions déterministes et par randomisation de documents offrent une bonne fonctionnalité, mais permettent une mise en corrélation entre les enregistrements. Des solutions spécifiques peuvent être définies, suivant les identificateurs à pseudonymiser (cf. chapitres 6 et 7 pour des exemples précis).

De plus, l'une des préoccupations de l'entité de pseudonymisation est la complexité associée à certaines approches en termes d'implémentation et d'évolutivité: la pseudonymisation des identificateurs peut-être être réalisée simplement et a-t-elle une incidence sur la taille de la base de données?

Tableau 3: Comparaison entre les différentes techniques en termes flexibilité (format d'identificateur) et de taille de pseudonyme

Méthode	Taille de l'identificateur	taille du pseudonyme *m* en bits
Compteur	Toutes	$m = log_2$k
Générateur de nombres aléatoires	Toutes	$m \gg 2log_2 k$
Fonction de hachage	Toutes	Fixe ou$m \gg 2log_2 k$
Code d'auth. de message	Toutes	Fixe ou$m \gg 2log_2 k$
Chiffrement	Fixe[20]	Fixe ou identique à celle de l'identificateur

La plupart des solutions peuvent être appliquées sur des identificateurs de taille variable, à l'exception de certaines méthodes de chiffrement. La taille du pseudonyme dépend de k, le nombre d'identificateurs contenu dans la base de données. Dans le cas du générateur de nombres aléatoires, de la fonction de hachage et du code d'authentification de message, il existe un risque de collision: la taille du pseudonyme doit donc être choisie avec soin (cf. «paradoxe des anniversaires»). La fonction de hachage et le code d'authentification de message sont des méthodes adaptées lorsque l'on souhaite s'assurer que la taille de l'empreinte empêche tout risque de collision. Enfin, la taille des pseudonymes produits par une méthode de chiffrement peut être fixe ou égale à celle de l'identificateur original. Le Tableau 3 présente l'évolutivité des approches susmentionnées vis-à-vis de la fonction de récupération.

5.4 RECUPERATION

L'utilisation d'informations supplémentaires se trouvant, par définition, au cœur de la pseudonymisation, l'entité de pseudonymisation doit mettre en œuvre un mécanisme de récupération. Ce mécanisme peut être plus ou moins complexe suivant la fonction de pseudonymisation utilisée. En général, il consiste à utiliser un pseudonyme $pseudo$ et un secret de pseudonymisation S pour récupérer (rétablir) l'identificateur correspondant Id. Cette situation peut notamment se produire lorsque l'entité de pseudonymisation a détecté une anomalie dans son système et souhaite contacter les entités concernées. Si cette anomalie est, par exemple, une violation de données, l'entité de pseudonymisation doit en informer les personnes concernées conformément au RGPD. De plus, le mécanisme de récupération peut être

[20] Le chiffrement par bloc fonctionne sur une entrée de taille fixe. Cependant, certains modes d'opération (tels que le compteur CTR) permettent de travailler sur toutes les tailles d'entrée.

nécessaire pour permettre l'exercice des droits des personnes concernées (au titre des articles 12 à 21 du RGPD).

Tableau 4: Comparaison entre les différentes techniques en termes de mécanisme de récupération

Méthode	Récupération selon pseudonyme
Compteur	Table d''association
Générateur de nombres aléatoires	Table de mappaged'association
Fonction de hachage	Table d'association
Code d'auth. de message	Table de mappaged'association
Chiffrement	Déchiffrement

La plupart des méthodes décrites précédemment imposent à l'entité de pseudonymisation de tenir une table d'association contenant les identificateurs et les pseudonymes afin de pouvoir récupérer un identificateur, à l'exception du chiffrement (Tableau 4). Le mécanisme de déchiffrement peut être appliqué directement à l'identificateur.

5.5 PROTECTION DU SECRET DE PSEUDONYMISATION

Pour que le processus de pseudonymisation soit efficace, l'entité de pseudonymisation doit toujours protéger le secret de pseudonymisation par des mesures techniques et organisationnelles adéquates. Celles-ci dépendent évidemment du scénario de pseudonymisation concerné (cf. chapitre 3).

Premièrement, le secret de pseudonymisation doit être isolé du jeu de données, c'est-à-dire que le secret de pseudonymisation et le jeu de données ne doivent jamais être conservés dans le même fichier (auquel cas il sera trop facile à un adversaire de déchiffrer les identificateurs). Deuxièmement, le secret de pseudonymisation doit être supprimé de façon sécurisée de tout support de stockage à risque (systèmes et mémoire de stockage). Troisièmement, des règles de contrôle d'accès rigoureuses doivent être mises en place pour garantir que seules les entités autorisées ont accès au secret. Un système de journalisation sécurisé doit conserver une trace de toutes les demandes d'accès relatives au secret. Quatrièmement, le secret de pseudonymisation doit être chiffré s'il est conservé sur un ordinateur, avec un stockage et une gestion des clés adaptés au type de chiffrement utilisé.

5.6 TECHNIQUES DE PSEUDONYMISATION AVANCEES

Il existe de multiples techniques de pseudonymisation autres que celles présentées dans les chapitres ci-dessus, des techniques plus avancées, adaptées à de nombreux contextes différents. Bien que décrire chacune de ces techniques en détail dépasse le cadre du présent rapport, quelques-unes de ces techniques sont présentées ci-dessous, pour les lecteurs intéressés.

Au-delà d'un hachage simple des données, des structures plus avancées telles que les arbres de Merkle [17, 18] utilisent des hashs d'ensembles de hashs, par ex. $h_3=\text{hash}(h_1,h_2)$, pour obtenir des pseudonymes structurés ne pouvant être dévoilés que uniquement, et pas entièrement. De même, les chaînes de hachage [19] reposent sur le hachage répété des valeurs de hachage de valeurs de hachage, par ex. $h_4=h_3(h_2(h_1(x)))$, afin de produire une valeur exigeant de nombreuses inversions de hachage pour ré-identifier les données d'origine

d'un pseudonyme donné. La chaîne de pseudonymisation est un exemple de cette technique de hachage. Elle implique plusieurs entités de pseudonymisation, chacune utilisant les pseudonymes créés par l'entité précédente pour créer de nouveaux pseudonymes en séquence (par ex. en appliquant une couche de hachage supplémentaire). Une telle chaîne reste fiable, même si un adversaire parvient à découvrir toutes les pseudonymisations appliquées dans la chaîne, sauf une, ce qui en fait une technique de pseudonymisation très robuste. C'est une pratique courante, notamment dans les essais cliniques.

Si le domaine d'entrée couvre plusieurs dimensions (cf. chapitre 8 pour un exemple), il est possible d'utiliser des filtres de Bloom [20] qui, sauf lorsqu'ils sont utilisés comme technique d'anonymisation, peuvent opérer une pseudonymisation efficace et réalisable d'un point de vue technique sur toutes les combinaisons de valeurs d'entrée possibles sur les différents domaines, malgré le problème de l'explosion d'état.

Les pseudonymes de transaction corrélables et/ou l'associativité de pseudonyme contrôlée avec l'option d'une ré-identification progressive peuvent également constituer une approche intéressante [21].

Enfin, toutes les techniques pouvant être utilisées pour renforcer l'anonymisation peuvent également être utiles dans un processus de pseudonymisation, notamment les techniques courantes de k-anonymat [3, 22, 23] ou de confidentialité différentielle [24], et au-delà [25]. Pour obtenir une description, cf. [2]. La preuve à divulgation nulle de connaissance [26] et le domaine plus large des authentifiants basés sur attribut peut également offrir des solutions intéressantes [2].

6. PSEUDONYMISATION DES ADRESSES IP

Sur la base des techniques et des informations présentées dans les chapitres précédents, le présent chapitre propose une étude de cas spécifique sur la pseudonymisation des adresses IP.

L'adresse IP est utilisée pour identifier de façon unique un périphérique présent sur un réseau IP. Il existe deux types d'adresses IP: les adresses IPv4 [27] et les adresses IPv6 [28]. Dans ce cas d'utilisation, le rapport s'intéresse aux adresses Ipv4, qui sont aujourd'hui les plus couramment utilisées. Il serait trop complexe d'étendre les concepts décrits précédemment aux adresses Ipv6 et dépasserait le cadre du présent document. Une adresse IPv4 se compose de 32 bits (128 bits pour les adresses IPv6) divisés entre un préfixe de réseau (multiplets les plus significatifs) et l'identificateur de l'hôte (multiplets les moins significatifs) avec l'aide d'un masque de sous-réseau. L'adresse est souvent présentée en notation décimale séparée par des points, qui se compose de quatre nombres décimaux compris entre 0 et 255, séparés par des points, par ex. «127.0.0.1». La taille du préfixe de réseau et de l'identificateur d'hôte dépend de la taille du bloc CIDR (Classless Inter-Domain Routing, routage interdomaine sans classe [29]). Certaines adresses IP sont spéciales, telles que 127.0.0.1 (hôte local) ou 224.0.0.1 (multidiffusion). Ces adresses spéciales sont toutes définies dans [30] et catégorisées dans 15 classes.

L'Internet Assigned Numbers Authority (IANA) gère l'ensemble de l'espace d'adressage IP avec l'aide de cinq registres internet régionaux (RIR). Ils affectent des sous-ensembles d'adresses IP aux organisations locales, tels que les fournisseurs de service internet, qui à leur tour affectent ces adresses aux périphériques des utilisateurs finaux. Chaque affectation d'adresse IP est documentée par le RIR correspondant, dans une base de données appelée WHOIS[21]. L'affectation peut être statique ou dynamique (à l'aide du protocole DHCP - Dynamic Host Configuration Protocol - par exemple).

D'un point de vue légal, le statut des adresses IP a été discuté par la Cour de justice de l'Union européenne lors du dossier C-582/14 Breyer contre Bundesrepublik Deutschland[22]. Les adresses IP, qu'elles soient statiques ou dynamiques, sont considérées comme des données à caractère personnel. Cette décision a été confirmée par l'Avis 4/2007 du groupe de travail «Article 29» sur la protection des données, sur le concept de données à caractère personnel [31]. Par conséquent, les bases de données ou les traces réseau contenant des adresses IP doivent être protégées et la pseudonymisation est dans ce cas une technique logique, qui permet d'utiliser des adresses IP tout en évitant qu'elles ne puissent être associées à des personnes spécifiques. Pour choisir une technique de pseudonymisation adaptée aux adresses IP, il convient de trouver un bon compromis entre fonctionnalité et protection des données. En effet, le responsable du traitement des données peut avoir besoin de calculer des statistiques ou de détecter des tendances (configuration incorrecte d'un périphérique ou gestion de la qualité des services) dans la base de données pseudonymisée. Les critères de fonctionnalité et de protection des données ne peuvent être traités de façon séparée dans la

[21] Pour en savoir plus, cf.: https://whois.icann.org
[22] Pour plus d'informations, cf.: https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A62014CJ0582

pratique, toutefois, nous les aborderons ici de façon indépendante pour une meilleure compréhension.

6.1 PSEUDONYMISATION ET NIVEAU DE PROTECTION DES DONNEES

La principale problématique en matière de pseudonymisation des adresses IP est la taille de l'espace d'entrée (le domaine d'identificateurs): il n'y a que 2^{32} adresses IP possibles. Cette taille modérée permet aux adversaires d'effectuer des attaques par recherche exhaustive et par dictionnaire pour obtenir une ré-identification complète ou une discrimination, si la fonction de pseudonymisation n'a pas été correctement choisie.

Au vu de cette problématique, les fonctions de hachage cryptographique sont particulièrement vulnérables. Nous prendrons pour exemple une adresse IP pseudonymisée à l'aide de la fonction de hachage SHA-256. Un adversaire possédant un pseudonyme/une empreinte peut utiliser les outils existants[23] pour effectuer une recherche exhaustive. Le Tableau 5 indique la durée de cette recherche, effectuée sur un ordinateur portable standard équipé d'un processeur Intel(R) Core(TM) i7-8650U de 1,90 GHz (8 cœurs), ainsi que la taille du dictionnaire. Dans le pire des cas, il ne faut à l'attaquant que deux minutes pour récupérer l'adresse IP associée à un pseudonyme donné.

Tableau 5: Coûts pratiques des attaques contre une pseudonymisation par fonction de hachage

Classe IP	Nombre d'IP possibles	Durée de la recherche exhaustive	Taille du dictionnaire
145.254.160.X	256	200 ms	8 Ko
145.254.X.X	65 536	200 ms	2 Mo
145.X.X.X	16777216	2 s	512 Mo
X.X.X.X	4294967296	2 min 16 s	128 Go

Supposons ensuite que l'adversaire souhaite déterminer si un pseudonyme correspond à une adresse donnée [30]. Il n'est pas nécessaire d'appliquer cette attaque par discrimination sur les 2^{32} d'adresses IP possible, mais seulement sur les 588 518 401 adresses spéciales possibles.

Le cas d'utilisation simple présenté ci-avant démontre que la pseudonymisation des adresses IP effectuée uniquement à l'aide de fonctions de hachage cryptographique est un échec. Si l'on souhaite assurer une protection des données adéquate, il est donc nécessaire de choisir d'autres fonctions de pseudonymisation, telles que le code d'authentification de message (MAC), le chiffrement avec clé secrète ad hoc ou le générateur de nombres aléatoires (GNA). Comme discuté précédemment dans ce rapport, un adversaire ne pourra lancer ce type d'attaque car ces méthodes utilisent une clé secrète (MAC et chiffrement) ou une source aléatoire (GNA). Il est également possible d'utiliser un compteur, en usant toutefois de précaution concernant les prédictions possibles (découlant de la nature séquentielle du compteur).

6.2 PSEUDONYMISATION ET NIVEAU DE FONCTIONNALITE

Comme mentionné précédemment, dans le cas des adresses IP, la fonctionnalité (l'utilité) peut être une exigence cruciale de l'entité de pseudonymisation, par exemple pour le calcul de

[23] Tel qu'un logiciel de piratage de mot de passe de type «John The Ripper», ou autre.

statistiques ou pour garantir la sécurité du réseau. L'approche choisie (indépendamment de la technique choisie) doit donc offrir une protection efficace tout en préservant certaines informations utiles de base (découlant des adresses IP). Dans cette section, nous présenterons deux dimensions différentes de cette problématique: tout d'abord, la possibilité de minimiser le niveau/le champ de la pseudonymisation des adresses IP; puis le choix de la stratégie de pseudonymisation (mode d'opération).

6.2.1 Niveau de pseudonymisation

Dans la section précédente, l'hypothèse de départ était que la pseudonymisation était appliquée à l'ensemble de l'adresse IP (32 bits). Cependant, si l'on souhaite améliorer la fonctionnalité, il est possible de l'appliquer uniquement aux bits les moins significatifs de l'adresse (identificateur d'hôte) pour préserver le préfixe réseau. Cette technique est appelée pseudonymisation avec conservation du préfixe [32]. Elle permet d'identifier l'origine globale d'un paquet (réseau) sans avoir à connaître exactement quel périphérique au sein du réseau l'a envoyé. Il est essentiel de savoir combien il existe de périphériques pour un préfixe donné. Le Tableau 5 présente différentes tailles de préfixe. Cette technique est déjà utilisée par plusieurs fournisseurs de services pour pseudonymiser les adresses IP (cf. exemple [33]).

6.2.2 Choix du mode de pseudonymisation

Le choix du mode de pseudonymisation a un impact fort sur le niveau de fonctionnalité et de protection des données, indépendamment du choix de la technique de pseudonymisation adoptée. Dans cette section, nous explorerons cette relation avec un exemple précis.

Prenons par exemple la pseudonymisation des adresses IP source et de destination d'une trace réseau. Le Tableau 6 indique les adresses source et de destination des premiers paquets d'une requête HTTP entre un client (145.254.160.237) et un serveur (65.208.228.223).

Tableau 6: Source et destination d'une requête HTTP

	Source	Destination
Paquet 1	145.254.160.237	65.208.228.223
Paquet 2	65.208.228.223	145.254.160.237
Paquet 3	145.254.160.237	65.208.228.223
Paquet 4	145.254.160.237	65.208.228.223
Paquet 5	65.208.228.223	145.254.160.237

Dans l'exemple ci-dessus, appliquons une pseudonymisation déterministe par générateur de nombres aléatoires. Chaque adresse IP est ainsi associée à un pseudonyme unique. La table d'association obtenue dans ce cas est présentée dans le Tableau 7. Après application de la pseudonymisation déterministe, on obtient les résultats du

Tableau 8.

Tableau 7: Table d'association pour une pseudonymisation déterministe

Adresse IP	Pseudonyme
145.254.160.237	238

65.208.228.223	47

Tableau 8: Adresses source et de destination transformées par pseudonymisation déterministe

Numéro de paquet	Source	Destination
Paquet 1	238	47
Paquet 2	47	238
Paquet 3	238	47
Paquet 4	238	47
Paquet 5	47	238

Comparons à présent les informations qu'il est possible d'extraire de la trace réseau d'origine (Tableau 6) et les résultats du

Tableau 8. Comme l'indique cette comparaison, il est possible, à partir des deux traces (originale et pseudonymisée), de déduire le nombre total d'adresses IP impliquées et combien de paquets ont été envoyés par chaque adresse durant la communication. Par conséquent, bien que les adresses IP du tableau 8 soient pseudonymisées, il est possible d'obtenir le même niveau d'analyse statistique (et donc de fonctionnalité) à partir de ces adresses.

Prenons à présent la cas d'une pseudonymisation par randomisation de documents utilisant un générateur de nombres aléatoires (GNA). À chaque fois qu'une adresse IP est trouvée, celle-ci est transformée en un pseudonyme différent. Par exemple, l'adresse IP 145.254.160.237 est associée à cinq pseudonymes: 39, 71, 48, 136 et 120 (Tableau 9). Après application de la pseudonymisation par randomisation de documents, on obtient les résultats du Tableau 10.

Tableau 9: Table d'association pour une pseudonymisation par randomisation de documents

Adresse IP	Pseudonyme
145.254.160.237	39, 71, 48, 136, 120
65.208.228.223	23, 30, 60, 160, 231

Tableau 10: Adresses source et de destination transformées par pseudonymisation par randomisation de documents

Numéro de paquet	Source	Destination
Paquet 1	39	23
Paquet 2	30	71
Paquet 3	48	60
Paquet 4	136	160
Paquet 5	231	120

Comme indiqué au tableau 10, alors qu'il était possible dans le Tableau 6 et le

Tableau 8 de compter deux adresses IP, cela n'est plus le cas dans le Tableau 10 qui implique virtuellement dix adresses IP. Le niveau de fonctionnalité est donc réduit (contrairement au niveau de protection, qui a augmenté). Bien entendu, l'application d'une pseudonymisation complètement aléatoire aura un impact encore plus fort sur le niveau de fonctionnalité. Le Tableau 11 compare les différents modes de pseudonymisation d'adresse IP, dans cette optique.

Tableau 11: Mode et pseudonymisation et fonctionnalité

Mode de pseudonymisation			
Fonctionnalité	**Déterministe**	**Randomisation de documents**	**Complètement aléatoire**
Statistiques (compte, ...)	OUI	NON	NON
Sémantique de protocole	OUI	NON	NON
Comparaison entre différentes traces	OUI	OUI	NON

Il est clair qu'il n'existe aucune solution unique et universelle à ce problème, et que le choix final repose sur les besoins spécifiques de l'entité de pseudonymisation en matière de fonctionnalité et de protection des données.

7. PSEUDONYMISATION DES ADRESSES ELECTRONIQUES

Dans ce chapitre, nous aborderons le sujet de la pseudonymisation des adresses électroniques comme cas d'utilisation spécifique des techniques présentées précédemment.

L'adresse électronique (e-mail) est l'identificateur standard d'un individu. Elle se présente sous la forme local@domaine, où la partie locale correspond à l'utilisateur titulaire de l'adresse et le domaine au fournisseur des services de messagerie électronique. Les adresses électroniques sont généralement utilisées dans diverses applications; par exemple, elles peuvent servir d'identifiant principal lorsque l'utilisateur s'inscrit à un service électronique. Les adresses électroniques sont également présentes dans de nombreuses bases de données, qui peuvent aussi contenir d'autres identificateurs, notamment le nom de l'utilisateur.

Les utilisateurs tendent à utiliser la même adresse électronique dans des applications différentes, à la partager avec diverses organisations, notamment lorsqu'ils créent un compte sur un site Web. De plus, les adresses électroniques sont souvent publiées en ligne, et il est prouvé qu'elles peuvent être facilement trouvées ou devinées[24]. En raison de ces caractéristiques particulières, la protection est particulièrement cruciale lorsqu'une adresse électronique est utilisée comme identificateur.

Dans le présent cas d'utilisation, les adresses électrique sont considérées comme des identificateurs (par ex. dans une base de données ou un service en ligne) et nous analyserons l'application de différentes techniques de pseudonymisation. Nous partirons de l'hypothèse que le processus de pseudonymisation est effectué par l'entité de pseudonymisation (par ex. le responsable du traitement des données) dans le cadre de la prestation/l'exécution d'un service.

7.1 COMPTEUR ET GENERATEUR DE NOMBRES ALEATOIRES

Conformément aux descriptions présentées au chapitre 5, le compteur, tout comme le générateur de nombres aléatoires, peuvent être utilisés pour effectuer la pseudonymisation des adresses électroniques, à l'aide d'une table d'association telle que celle illustrée au Tableau 12. Il est évident que la pseudonymisation sera aussi robuste que la table d'association sera sécurisée, et stockée séparément des données pseudonymisées.

Tableau 12: Exemple de pseudonymisation d'adresses électroniques par GNA ou compteur (pseudonymisation totale)

Adresse électronique	Pseudonyme (générateur de nombres aléatoires)	Pseudonyme (compteur)
alice@abc.eu	328	10
bob@wxyz.com	105	11
eve@abc.eu	209	12
john@qed.edu	83	13

[24] Il a en effet été prouvé que, en récupérant simplement une information de base comme un nom d'utilisateur sur les réseaux sociaux, il était possible de recueillir efficacement des millions d'adresses électroniques [38].

alice@wxyz.com	512	14
mary@clm.eu	289	15

Au tableau 12, la méthode par compteur et la méthode par GNA génèrent toutes deux des pseudonymes ne révélant aucune information sur l'identificateur initial (l'adresse électronique) et ne permettant pas d'analyse complémentaire (ex. analyse statistique) sur les pseudonymes. Pour améliorer le niveau de fonctionnalité, il est possible d'appliquer la pseudonymisation sur une partie seulement de l'adresse électronique, par exemple la partie locale (sans toucher au domaine, cf. Tableau 13).

Tableau 13: Exemple de pseudonymisation d'adresses électroniques par GNA ou compteur (pseudonymisation de la partie locale de l'adresse uniquement)

Adresse électronique	Pseudonyme (générateur de nombres aléatoires)	Pseudonyme (compteur)
alice@abc.eu	328@abc.eu	10@abc.eu
bob@wxyz.com	**105@wxyz.com**	**11@wxyz.com**
eve@abc.eu	209@abc.eu	12@abc.eu
john@qed.edu	**83@qed.edu**	**13@qed.edu**
alice@wxyz.com	512@wxyz.com	14@wxyz.com
mary@clm.eu	**289@clm.eu**	**15@clm.eu**

Comme indiqué dans le Tableau 13, il est encore possible de connaître le domaine même une fois les adresses pseudonymisées, et donc d'effectuer une analyse le cas échéant (par ex. sur le nombre d'utilisateurs provenant du même domaine). Comme discuté précédemment dans ce document, la méthode par compteur peut être plus faible en termes de protection car elle permet d'effectuer des prédictions en raison de sa nature séquentielle (par ex. dans les cas où les adresses électroniques sont issues du même domaine, la méthode par compteur peut révéler des informations sur l'ordre d'apparition des différents utilisateurs dans la base de données).

Sur la base de ce cas de figure simple, et suivant le niveau de protection des données et de fonctionnalité que l'entité de pseudonymisation souhaite mettre en œuvre, différentes variantes sont possibles en conservant des degrés d'information variés sur les pseudonymes (par ex. sur les domaines identiques, la partie locale de l'adresse, etc.).

Tableau 14: Exemples de pseudonymisation d'adresses électroniques par GNA – Niveaux de fonctionnalité variés

Adresse électronique	Pseudonyme (GNA) conservant les informations sur les domaines identiques	Pseudonyme (GNA) conservant les informations sur les extensions/pays identiques	Pseudonyme (GNA) conservant les informations sur les parties locales et les domaines identiques	Pseudonyme (GNA) conservant les informations sur les extensions/pays, les domaines et les parties locales identiques
alice@abc.eu	328@1051	328@1051.3	328@1051	328@1051.3
bob@wxyz.com	105@833	105@833.7	105@833	105@833.7
eve@abc.eu	209@1051	209@1051.3	209@1051	209@1051.3

john@qed.edu	83@420	83@420.8	83@420	83@420.8
alice@wxyz.com	512@833	512@833.7	328@833	328@833.7
mary@clm.eu	289@2105	289@2105.3	289@2105	289@2105.3

Le principal inconvénient de la méthode par compteur et de la méthode par GNA est leur évolutivité dans le cas de jeux de données de grande taille, tout particulièrement lorsque l'on souhaite que le même pseudonyme soit toujours affecté à la même adresse (comme dans le cas d'une pseudonymisation déterministe, cf. Tableau 12). En effet, dans un tel cas, l'entité de pseudonymisation doit effectuer un contrôle croisé sur l'ensemble de la table de pseudonymisation, dès qu'une nouvelle entrée doit être pseudonymisée. La complexité s'accroit dans les scénarios de mise en œuvre les plus sophistiqués, comme illustré au tableau 14, par exemple lorsque l'entité de pseudonymisation souhaite classifier les adresses électroniques d'un même domaine ou pays sans révéler celui-ci.

7.2 FONCTION DE HACHAGE CRYPTOGRAPHIQUE

Comme indiqué en [34], le nombre total de comptes de messagerie électronique au niveau mondial est estimé à 4,7 milliards ≈ 2^{32} (étant donné que, malgré un espace théorique quasiment infini pour les adresses électroniques valides, les adresses existantes occupent en réalité un espace bien plus restreint). Ce fait, également mentionné dans les chapitres précédents, a pour conséquence que les adresses électroniques peuvent être facilement trouvées ou devinées[25], ce qui explique la faiblesse des fonctions de hachage cryptographique en matière de pseudonymisation [34]. Il est en effet facile, tant pour un adversaire interne qu'externe, d'accéder à une liste pseudonymisée d'adresses électroniques, pour effectuer une attaque par dictionnaire (Graphique 10). Cette observation concerne tous les scénarios de pseudonymisation présentés au chapitre 3 (que l'entité de pseudonymisation soit le responsable du traitement des données, le sous-traitant ou un tiers de confiance).

Graphique 10: Rétablissement d'une adresse électronique à partir de sa valeur de hachage

Malgré les inconvénients précédemment mentionnés pour les fonctions de hachage cryptographique, il convient de noter que, comme indiqué en [35], les fournisseurs de services partagent souvent les adresses électroniques avec des tiers, par le simple fait d'effectuer un hachage. Un exemple concret de ce phénomène est ce que l'on appelle les listes d'audience personnalisées, qui donnent aux entreprises la possibilité de comparer les valeurs de hachage des adresses électroniques de leurs clients, pour la création de listes de clients communes[26].

[25] En théorie, si toutes les adresses possibles étaient à disposition d'un adversaire, même une attaque par force brute serait réalisable; dans tous les cas, l'espace (relativement) restreint des adresses électroniques explique qu'une attaque par conjecture aléatoire pourrait réussir. Plus inquiétant encore, dans cette ère du Big Data, il n'est parfois même pas nécessaire d'effectuer des conjectures aléatoires, étant donné que les adresses électroniques valides sont souvent disponibles au public ou facilement dérivées dans des contextes spécifiques (par ex. si le domaine et le format d'une organisation spécifique sont connus).
[26] cf. exemple: https://www.facebook.com/business/help/112061095610075?helpref=faq_content

Malgré les importants risques en matière de protection des données exposés précédemment, les valeurs de hachage cryptographique peuvent malgré tout être utiles dans certaines conditions, par exemple pour le codage interne des adresses électroniques (par ex. dans le contexte d'activités de recherche) et comme mécanisme de validation/d'intégrité, pour un responsable du traitement des données (cf. également [1]). Les fonctions de hachage peuvent également être utilisées pour pseudonymiser des éléments individuels d'une adresse électronique (par ex. le domaine uniquement), offrant une fonctionnalité accrue pour les pseudonymes dérivés, dans la mesure où la partie restante est pseudonymisée par une méthode plus robuste (par ex. MAC). Le risque que l'adresse électronique dans sa totalité soit rétablie est alors grandement réduit.

7.3 CODE D'AUTHENTIFICATION DE MESSAGE (MAC)

Par rapport au hachage simple, la méthode par code d'authentification de message (MAC) offre des avantages clairs en termes de protection des données pour la pseudonymisation des adresses électroniques, dans la mesure où la clé secrète est stockée de façon sécurisée. D'autre part, l'entité de pseudonymisation peut utiliser différentes clés secrètes, par exemple pour générer des pseudonymes différents pour chaque secteur avec les mêmes adresses électroniques. La fonction MAC peut également être utilisée pour empêcher le responsable du traitement des données d'accéder aux adresses électroniques, dans les cas où il lui suffit d'accéder aux pseudonymes pour réaliser l'opération prévue (comme dans les scénarios 5 et 6 du chapitre 3). Une telle situation pourrait se produire dans le cas d'un affichage publicitaire basé sur les centres d'intérêt, pour lequel les annonceurs doivent associer un pseudonyme unique à chaque individu, sans pouvoir révéler l'identité réelle de l'utilisateur [36].

Comme avec les techniques précédentes, plusieurs scénarios de mise en œuvre pratiques peuvent être envisagés pour accroître l'utilité des pseudonymes. Par exemple, l'une des approches possibles consiste à appliquer la fonction MAC séparément aux différentes parties de l'adresse électronique (par ex. la partie locale et le domaine), en utilisant la même clé secrète. Un exemple caractéristique de cette approche est présenté dans la Graphique 11: l'utilisation de la même clé pour chaque fonction MAC génère les mêmes sous-pseudonymes pour les domaines correspondants (en vert), lorsque ceux-ci sont identiques. Toutefois, étant donné que le résultat généré par la fonction MAC a une taille fixe, qui est généralement bien supérieure à la taille de l'adresse électronique initiale[27], les pseudonymes obtenus peuvent être d'une taille relativement conséquente (qui augmente encore si les différentes parties de l'adresse sont pseudonymisées séparément).

Graphique 11: Utilisation de la fonction MAC pour générer des adresses électroniques pseudonymisées présentant un bon niveau de fonctionnalité

[27] La taille standard du résultat d'une fonction de hachage (avec ou sans clé) est de 256 bits, soit 32 caractères.

L'un des aspects importants de la mise en œuvre pratique de la fonction MAC est la récupération. Il convient de souligner que même l'entité de pseudonymisation des données, qui a accès à la clé secrète, n'est pas capable de rétablir directement les adresses à partir des pseudonymes; une telle inversion est possible uniquement de façon indirecte, en re-générant les pseudonymes pour chaque adresse électronique connue puis en les mettant en corrélation avec ceux de la liste pseudonymisée. Lorsqu'une table d'association de pseudonymisation est disponible, inverser les pseudonymes devient une opération facile; toutefois, dans un tel cas, les exigences de sécurité du stockage augmentent en conséquence. Pour toutes ces raisons, la fonction MAC n'est probablement pas la technique de pseudonymisation la plus pratique dans les cas où le responsable du traitement des données a besoin de facilement mettre en corrélation les pseudonymes avec les adresses électroniques correspondantes (comme dans les scénarios présentés aux chapitres 3.1 et 3.2).

7.4 CHIFFREMENT

Une alternative au code d'authentification de message est le chiffrement, tout particulièrement lorsqu'il est appliqué de façon déterministe, c'est-à-dire en utilisant une clé secrète afin de générer un pseudonyme pour chaque adresse électronique (chiffrement symétrique). Le déploiement en est facilité, car il n'est pas nécessaire de fournir une table d' association de pseudonymisation: la récupération s'effectue directement par un processus de déchiffrement [37].

Notez que, bien qu'il soit possible d'appliquer certains algorithmes cryptographiques asymétriques (à clé publique) de façon déterministe[28], ceux-ci ne sont pas recommandés pour la pseudonymisation des adresses électroniques (ou pour les autres types de données, cf. [1]). Imaginons par exemple que l'entité de pseudonymisation souhaite générer, pour chaque adresse électronique, différents pseudonymes en fonction des différents utilisateurs/destinataires concernés (internes ou externes), en partant de l'hypothèse que chaque destinataire sera à même de ré-identifier ses propres données, mais pas les données pseudonymisées des autres destinataires. Une méthode pour obtenir ce résultat serait de chiffrer les adresses électroniques avec la clé publique de chaque destinataire, permettant uniquement à celui-ci de réaliser le déchiffrement. Cependant, si l'on part du principe que les clés publiques sont disponibles à tout un chacun, un adversaire pourrait lancer une attaque par dictionnaire sur les adresses électroniques connues (ou devinées) (comme indiqué dans la Graphique 10, où un chiffrement par clé publique, avec une clé publique connue, a été utilisé au lieu de la fonction de hachage).

La nature du chiffrement empêche par défaut toute utilité pour les données pseudonymisées. Chiffrer séparément les différentes parties de l'adresse électronique peut suffire pour résoudre ce problème, de même qu'avec les codes d'authentification de message (MAC) (cf. Graphique 11), où la fonction MAC peut être remplacé par un algorithme de chiffrement. Généralement, pour permettre aux pseudonymes de contenir quelques informations utiles, il est nécessaire d'appliquer des techniques cryptographiques spécifiques. L'exemple suivant illustre ce processus, avec le chiffrement avec conservation du format.

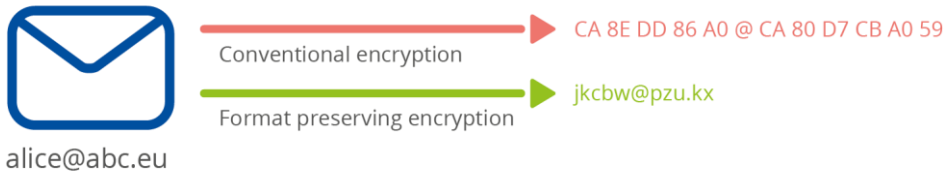
CHIFFREMENT AVEC CONSERVATION DU FORMAT (FPE)

Le schéma d'une base de données suppose parfois un type particulier de données, dans des champs particuliers. Par exemple, il est supposé qu'une adresse électronique se compose d'une partie locale, suivie du symbole @, lui-même suivi d'un nom de domaine. Si le responsable du traitement des données n'a pas besoin de conserver les adresses électroniques

[28] Malgré le fait que, pour des raisons de sécurité, un algorithme de clé publique soit probabiliste sur le principe [1].

initiales, mais uniquement une liste pseudonymisée reproduisant la structure de la base de données, le chiffrement avec conservation du format (FPE, Format Preserving Encryption) est une méthode adaptée. Il existe plusieurs méthodes de mise en œuvre pour le chiffrement FPE, fondées sur des schémas de chiffrement connus[29]. Dans tous les cas, une substitution (pseudo-)aléatoire de caractères[30] par d'autres caractères du même alphabet (c'est-à-dire l'ensemble de caractères alphanumériques enrichi par des caractères spéciaux utilisé dans la partie locale des adresses électroniques) suffit à garantir que le pseudonyme dérivé présente le format souhaité. La différence entre le chiffrement FPE et la cryptographie conventionnelle est illustré dans la Graphique 12.

Graphique 12: Comparaison entre chiffrement conventionnel et chiffrement avec conservation du format (FPE) pour dériver un pseudonyme à partir d'une adresse électronique



Notez que, au Graphique 12, un chiffrement de flux symétrique a été utilisé comme chiffrement conventionnel, afin de garantir que le pseudonyme dérivé présente la même longueur que l'adresse initiale (les caractères du pseudonyme dérivé ne sont pas alphanumériques et sont donc indiqués sous forme hexadécimale).

Il convient également de noter que, suivant le cas de figure, il peut être nécessaire de concevoir avec soin et de façon appropriée les implémentations FPE, afin d'éviter l'apparition de schémas engendrant une fuite d'informations sur l'identité des utilisateurs.

[29] Voir pour exemple https://csrc.nist.gov/publications/detail/sp/800-38g/rev-1/draft, un document préliminaire du National Institute of Standards and Technology (NIST) sur les méthodes de chiffrement avec conservation de format appropriées, présentant les vulnérabilités potentielles lorsque la taille de domaine est trop petite.
[30] La substitution de caractère est une forme de chiffrement spéciale (des problèmes de sécurité puissent se produire si la substitution n'est pas correctement effectuée).

8. LA PSEUDONYMISATION EN PRATIQUE: UN SCENARIO PLUS COMPLEXE

Comme nous avons pu l'observer dans les deux cas d'utilisation exposés aux chapitres 6 et 7, la pseudonymisation reste une tâche complexe et sujette aux erreurs, même dans le cas de types de données très simples comme les adresses IP ou les adresses électroniques. Dans les systèmes réels, cependant, c'est rarement le choix de la technique de pseudonymisation utilisée pour un ou deux identificateurs spécifiques qui cause le plus de problèmes, mais la corrélation implicite au sein d'un ensemble de pseudonymes et d'autres valeurs de données, qui sont réunis dans une structure de données plus complexe. L'exemple le plus fréquent de ce problème se produit lorsqu'un service en ligne crée des profils d'utilisateur au moment de l'inscription de celui-ci, puis enrichit ces profils avec des informations personnelles relatives à l'utilisateur dès que de nouvelles données sont disponibles. Dans ce cas, même si l'adresse électronique et l'ensemble des adresses IP trouvées dans les journaux d'accès de l'utilisateur sont rigoureusement pseudonymisées, tel que discuté précédemment, il existe encore un risque important de ré-identification ou de discrimination, parfois même sur la structure de données pseudonymisée elle-même. Dans cette section, nous aborderons des cas plus complexes de pseudonymisation des données.

8.1 EXEMPLE DE SCENARIO

Aux fins de discussion, imaginons un exemple de scénario très similaire à ce que l'on pourrait trouver dans le monde réel: un réseau social en ligne. L'opérateur imaginaire de ce réseau social, RéseauSocial Inc. (ci-après dénommé RS), joue le rôle de responsable du traitement des données et permet à ses utilisateurs (que nous supposerons humains uniquement) de créer un compte qui sera stocké dans son centre de données. Grâce à ce compte, les utilisateurs ont accès à différentes fonctionnalités, qui les mettent par exemple en lien avec d'autres utilisateurs, des entreprises ou des centres d'intérêt. Lors de l'inscription, les utilisateurs de RS doivent donner leur nom réel (prénom et nom de famille), un surnom, leur date de naissance et leur sexe, ainsi que différentes informations facultatives à caractère personnel (situation géographique, centres d'intérêts, caractères biométriques, etc.) et une adresse électronique valide. Lorsqu'un utilisateur accède à des services proposés par RS, cette interaction est consignée dans un journal et ajoutée à son profil d'utilisateur (avec l'horodatage et l'adresse IP d'accès).

Pour respecter les exigences du RGPD, la direction de RS a décidé de pseudonymiser les adresses IP dans les journaux d'accès, par l'une des techniques discutées au chapitre 6. Les autres informations restent en texte clair, car il est parfois nécessaire de les présenter à l'utilisateur sur les sites Web de RS ou pour effectuer des contrôles et des validations (par ex. la date de naissance est nécessaire pour calculer l'âge et vérifier que l'utilisateur a plus de 16 ans pour accéder à certains services spéciaux). La pseudonymisation de l'adresse électronique n'est pas une procédure réalisable dans ce cas de figure, car RS doit pouvoir envoyer aux utilisateurs des courriers électroniques avec notification (et d'autres contenus).

Imaginons une deuxième compagnie imaginaire, Services de Sécurité en Ligne SA (ci-après dénommée SSL), qui joue le rôle de sous-traitant pour le compte de RS, avec pour mission d'assurer les services de sécurité et de stockage pour certains éléments de la base de données utilisateur de RS. Dans cette position, SSL a accès aux fichiers journaux pseudonymisés, c'est-

à-dire aux adresses IP pseudonymisées avec horodatage de l'accès à l'ensemble des sites Web, mais pas aux adresses IP d'origine. SSL ne peut donc pas ré-identifier les utilisateurs associés à une adresse IP spécifique, car ces données sont stockées dans une base de données RS différente qui ne lui est pas accessible. En ce qui concerne la pseudonymisation, nous retrouvons ici le scénario du chapitre 3.3, avec RS jouant le rôle de responsable du traitement des données et SSL celui de sous-traitant.

8.2 INFORMATIONS INHERENTES AUX DONNEES

Au premier abord, SSL ne semble pas capable de briser le processus de pseudonymisation des adresses IP appliqué par RS, si l'on part de l'hypothèse que RS a utilisé une fonction de pseudonymisation suffisamment robuste. Toutefois, suivant la fonction de pseudonymisation choisie, et tout particulièrement la stratégie de pseudonymisation appliquée (cf. chapitre 5.2), SSL peut malgré tout parvenir à déduire si un certain pseudonyme apparaît fréquemment, rarement, une seule fois ou pas du tout dans la base de données. Cette information en soi ne suffit pas pour révéler une identité, mais elle peut déjà être utilisée pour identifier les utilisateurs présentant des accès fréquents. Si le journal d'accès contient de nombreux fois un pseudonyme, SSL peut en déduire que celui-ci correspond à un utilisateur fréquent du réseau RS. À l'inverse, si un pseudonyme apparaît pour la première fois dans le jeu de données, il est fort probable que cet utilisateur vienne de s'inscrire sur RS et qu'il ait accédé à son compte pour la première fois, ou encore que l'adresse IP d'un utilisateur déjà inscrit ait changé (ce qui est assez fréquent, toutes ces observations deviennent donc probabilistes).

Ce type d'informations inhérentes aux données peut être utile à SSL, notamment pour savoir combien d'utilisateurs du réseau RS sont des utilisateurs courants et combien ne se sont connectés qu'une seule fois, sans revenir (avec un degré d'erreur probabiliste découlant du changement d'adresse IP). Ces informations sont déjà critiques dans la relation professionnelle entre RS et SSL.

Au-delà des informations inhérentes aux données, le fait que SSL bénéficie d'un accès continu à la base de données de RS lui permet de recueillir également des informations, d'une autre manière: en surveillant en continu le jeu de données stocké pour RS, SSL est informé de toute modification qui lui est apportée. Cela inclut le nombre d'accès au site Web de RS, ce qui n'a aucune incidence particulière, mais peut également lui servir à compter le nombre de nouveaux utilisateurs ayant créé un compte (première création d'un pseudonyme) chaque jour ou chaque mois. Bien que de nature majoritairement statistique, ce type d'information peut déjà être utilisé pour planifier des attaques par discrimination (en prévoyant des impacts différents suivant le groupe d'utilisateurs): SSL est la première entité à être informée lorsqu'un pseudonyme de nouvel utilisateur est généré et quel jour, ce qui lui permet de surveiller le volume d'interactions de cet utilisateur spécifique avec RS. Cette information peut rapidement se transformer en problématique de protection des données, comme nous le découvrirons ci-après.

8.3 DONNEES CORRELEES

Dans notre exemple de scénario, les données accessibles à SSL lui fournissent davantage d'informations que les simples adresses IP: chaque entrée de journal contient en effet un horodatage de l'accès. Ainsi, au lieu de surveiller fréquemment les changements apportés à la base de données de RS, il suffit à SSL de consulter les horodatages associés à chaque pseudonyme pour effectuer le même type de discrimination. Ces horodatages sont conservés avec les adresses IP pseudonymisées, et donc directement et individuellement corrélées à cette information. Sur la base des données corrélées, SSL peut grandement accroître sa connaissance d'utilisateurs RS spécifiques: cet utilisateur accède-t-il à RS le matin, durant sa pause déjeuner ou le soir? Uniquement ou principalement le dimanche? Uniquement lors des

fêtes religieuses du calendrier orthodoxe? Uniquement durant les vacances scolaires au Danemark?

Chaque caractérisation supplémentaire permet à SSL de se rapprocher d'une rupture de la pseudonymisation, simplement sur la base des horodatages et de sa capacité à mettre en corrélation différents enregistrements de données avec des pseudonymes identiques. Cette caractérisation des utilisateurs du réseau RS par SSL peut commencer à être considérée comme des informations personnelles. Toutefois, la mise en corrélation exige davantage d'informations à lier avec les jeux de données structurés eux-mêmes, par exemple le calendrier religieux orthodoxe ou les dates des vacances scolaires danoises. Ce type d'association peut être considéré comme une attaque par connaissances contextuelles, tel que discuté au chapitre 4, avec toutefois une complexité variable concernant les informations contextuelles nécessaires. D'autre part, ce type d'informations extraites est de nature statistique, et donc pas fiables à 100 %, mais offrant un certain degré de probabilité. Dans le cas présent, plus le nombre d'entrées de la base de données est élevé, plus les hypothèses de corrélation seront fiables (ou falsifiables). Ainsi, plus le réseau social RS est grand, plus il est facile pour SSL de lancer des attaques par discrimination, voire même des attaques par ré-identification.

Cet exemple ne comportait pourtant qu'une adresse IP pseudonymisée et un horodatage. Il en serait de même, voire avec davantage de fiabilité, dans le cas d'une adresse électronique pseudonymisées au lieu d'une adresse IP, car celle-ci a tendance à changer moins fréquemment et constitue donc un identificateur unique pour un utilisateur humain.

8.4 MISE EN CORRELATION DE LA DISTRIBUTION DES OCCURRENCES

Les structures de données de l'exemple ci-dessus sont plus simplistes et de petite envergure: de simples adresses IP et un horodatage. Pourtant, elles peuvent suffire à lancer des attaques par discrimination, voire des attaques par ré-identification, si les informations contextuelles sont suffisantes. De plus, en conditions réelles, les entrées de données contiennent généralement plus d'informations que ces deux valeurs, et les enregistrements de données plus de détails pouvant être mis à profit pour dévoiler les pseudonymes.

Si l'on suppose que RS conserve davantage d'informations que l'horodatage et l'adresse IP pseudonymisée dans chaque enregistrement de données, par exemple le type et la version du navigateur[31] de l'utilisateur, la langue configurée ainsi que les préférences de langue de l'utilisateur (telles que définies dans les paramètres du navigateur), la version du système d'exploitation de l'ordinateur de l'utilisateur, etc. Comme l'a découvert la Electronic Frontier Foundation dans son projet Panopticlick[32], cette combinaison de paramètres de navigateur peut déjà être suffisante pour identifier de façon unique un navigateur en particulier, et donc son utilisateur, sur un site Web. Si RS enregistre désormais toutes ces informations pour chaque accès à son site Web, SSL pourra en avoir également connaissance.

Même si RS applique une certaine forme de pseudonymisation sur chacune de ces configurations (par ex. en enregistrant uniquement une valeur de hachage par clés pour la chaîne représentant la version du navigateur de l'utilisateur), SSL pourra malgré tout voir ces chaînes de version de navigateur pseudonymisées, calculer les statistiques sur la fréquence d'affichage de ces valeurs de hachage dans la base de données globale de RS, puis comparer cette distribution des valeurs existantes avec les statistiques publiquement disponibles recueillies sur le site Web Panopticlick pour dévoiler la chaîne de version de navigateur originale cachée derrière chaque valeur de hachage, et ce malgré l'application d'une fonction

[31] Il est à noter qu'il s'agit là d'un comportement par défaut du journal, par exemple sur un serveur Web Apache.
[32] https://panopticlick.eff.org/

pseudonymisation appropriée. Le simple fait que la distribution statistique des différents pseudonymes soit mise en corrélation avec la distribution statistique de leur équivalent en texte clair peut suffire pour dévoiler les pseudonymes, avec une forte probabilité de succès.

Bien entendu, le résultat dépend grandement l'approche de pseudonymisation choisie. Si une approche soigneusement conçue a été mise en place, l'ajout de métadonnées à l'argument de la fonction de pseudonymisation peut offrir une meilleure protection contre l'ingénierie inverse.

8.5 CONNAISSANCES SUPPLEMENTAIRES

Si SSL possède des informations supplémentaires sur les caractéristiques d'un utilisateur donné et tente de révéler les enregistrements de données de cet utilisateur à partir de la base de données pseudonymisée que RS met à sa disposition, chaque information supplémentaire peut être critique. Si SSL sait que l'utilisateur ciblé est un homme et qu'il utilise le navigateur Chrome sur un iPad, cette information seule réduit déjà de manière significative le champ des possibles parmi les profils d'utilisateur auquel il a accès. Chacune de ces valeurs de données, même pseudonymisée, réduit le champ des possibles, c'est-à-dire l'ensemble de profils d'utilisateur contenus dans la base de données RS et pouvant correspondre à l'utilisateur ciblé par SSL. Les informations relatives au navigateur peuvent être mises à profit dans les attaques par probabilité de distribution, décrites dans la Section 8.4, éliminant une large portion des profils d'utilisateur dont les pseudonymes de navigateur sont trop fréquents ou trop rares pour correspondre à la probabilité de configuration «navigateur Chrome sur iPad».

Dans les profils restants, une simple attaque par force brute ou distribution statistique révèle à SSL quel pseudonyme correspond à quel sexe d'utilisateur, éliminant environ la moitié de ces profils restants. Si ensuite tous les profils d'utilisateur restants ont en commun que leur premier accès au réseau RS a eu lieu entre les mois de mai et juillet 2018, SSL dispose donc d'une information supplémentaire sur l'utilisateur ciblé, à savoir qu'il s'est inscrit sur RS à cette période. Voilà une attaque par inférence menée avec succès. En analysant de façon plus approfondie les profils d'utilisateur restants, SSL peut identifier un schéma d'horodatage spécifique de l'utilisation du réseau RS observé chez deux de ces profils, et correspondant au schéma d'utilisation supposé de l'utilisateur ciblé (que SSL a pu observer à certaines occasions par le passé). La cible de recherche se résume alors à deux profils d'utilisateur.

Chaque information que ces deux profils ont en commun doit donc être considérée comme vraie pour l'utilisateur ciblé, ce qui en dit sans doute long à SSL sur la cible de recherche existante. Pour éliminer le faux candidat restant, il suffit à SSL de surveiller l'utilisation spécifique du réseau RS par ces deux profils et, lors de l'accès suivant, de confirmer si cet accès a pu ou non être initié par l'utilisateur ciblé (sur la base de la connaissance contextuelle qu'il a de lui). En fin de compte, SSL parvient à mettre en corrélation le profil d'utilisateur avec l'identité ciblée. SSL peut ainsi déchiffrer toutes les pseudonymisations effectuées sur les valeurs de données de cet utilisateur, et potentiellement dévoiler ou discriminer d'autres profils également.

Il convient pourtant de noter que le problème des informations supplémentaires disponibles est «orthogonal» à la pseudonymisation, tout en étant par essence une problématique de protection des données. Par conséquent, comme il en a été fait mention plus haut dans ce rapport, il peut être intéressant d'envisager, en plus de la pseudonymisation, l'injection de bruit dans les arguments de la fonction de pseudonymisation ou l'usage de généralisation, afin de rendre les attaques par force brute moins efficaces (cf. également chapitre 5.6). Ce degré de liberté est une façon de renforcer la pseudonymisation et de se protéger contre les attaques les plus probables.

8.6 CORRELATION ENTRE PLUSIEURS SOURCES DE DONNEES

Dans la suite du scénario illustré dans les sections précédentes, avec les sociétés RS et SSL, un scénario de pseudonymisation encore plus délicat et complexe apparaît lorsque l'on ne parle plus de deux entreprises seulement, mais d'un marché à grande échelle des données pseudonymisées. Dans ce type de scénario, de multiples organisations partagent des jeux de données pseudonymisés contenant des données à caractère personnel, avec une certaine exigence de fonctionnalité (par ex. créer des profils à des fins marketing) tout en protégeant l'identité même des personnes concernées. L'argument fréquent dans de tels scénarios est que la pseudonymisation empêche une ré-identification des personnes concernées, légitimant ainsi le partage des données pratiqué. Ce rapport n'a pas pour vocation d'argumenter contre ou en faveur de la légitimité d'un tel partage de jeux de données pseudonymisés, mais de discuter des problématiques associées à l'application efficace de la pseudonymisation dans un tel contexte.

Imaginons par exemple un groupe d'entreprises A à E, qui toutes recueillent des données à caractère personnel sur leurs utilisateurs, notamment les données recueillies par RS dans l'exemple précédent. Une mise en corrélation des profils d'utilisateur issus de différentes entreprises peut être effectuée en comparant les adresses électroniques utilisées par les utilisateurs respectifs. Si deux profils d'utilisateur provenant, par exemple, des entreprises B et D se sont inscrits avec exactement la même adresse électronique, il s'agit certainement de la même personne. L'adresse électronique elle-même étant, bien entendu, considérée comme donnée à caractère personnel, tel que discuté au chapitre 7. Il est donc obligatoire d'effectuer une pseudonymisation de l'adresse électronique dans les jeux de données B et D avant de les partager avec les entreprises du groupe (A, B, C, D et E).

Ici, la difficulté réside dans le fait que tous les participants souhaitent conserver un bon niveau de fonctionnalité sur les données pseudonymisées, pour mettre en corrélation les profils appartenant à une même personne, sans toutefois réduire le degré de protection de l'identité de cet utilisateur. Les cinq entreprises du groupe doivent donc appliquer exactement le même processus de pseudonymisation, avec la même fonction de pseudonymisation et le même secret de pseudonymisation, pour pouvoir comparer et mettre en corrélation les enregistrements de données issus des différents jeux de données. Il y a dans ce cas de figure un véritable décalage entre fonctionnalité (mise en corrélation des adresses électroniques pseudonymisées) et protection des données (adresses électroniques des utilisateurs). En d'autres termes, les entreprises B et D doivent être capables de et autorisées à savoir que des enregistrements de données spécifiques partagent la même adresse électronique, et appartiennent donc au même utilisateur, mais elles ne doivent pas être autorisées à savoir de quelle adresse électronique (et donc de quelle personne concernée) il s'agit.

Comme discuté au chapitre 7, dans certains cas de figure, l'utilisation de fonctions de pseudonymisation faibles (telles que le hachage simple) rend possible des attaques peu sophistiquées par force brute, conjecture ou distribution statistique. Enrichies par les données supplémentaires (non personnelles) contenues dans les enregistrements de données partagés, et parfois par des connaissances contextuelles complémentaires, ces attaques doivent être prises au sérieux car rencontrant le succès dans de nombreux cas de figure. Plus inquiétant encore, plus le nombre d'entreprises partageant des informations sur les attributs d'une personne concernée est élevé, plus la quantité d'informations disponibles pour un adversaire pour briser la pseudonymisation est importante, et plus la probabilité de succès de ce type d'attaque est forte.

Il existe des risques en matière de confidentialité, même dans les scénarios classiques où les entreprises appliquent différentes techniques de pseudonymisation (parfois même robustes)

aux identificateurs de leurs utilisateurs (adresse électronique ou adresse IP). Imaginons que les entreprises A à E précédemment mentionnées fournissent des données pseudonymisées de ce type à la société SSL, par exemple dans le cadre d'une prestation de services statistiques. Si les pseudonymes fournis sont accompagnés d'informations sur l'équipement matériel/le navigateur des utilisateurs, tel que décrit dans la Section 8.4 (paramètres de navigateur, système d'exploitation, etc.), et sachant que les informations relatives à ces équipements/logiciels sont uniques[33], SSL pourra facilement mettre en corrélation les pseudonymes fournis par les différentes entreprises et correspondant au même utilisateur.

8.7 CONTRE-MESURES

Comme discuté au chapitre 5, les techniques de pseudonymisation aléatoire (complètement aléatoire ou par randomisation de documents) permettent de réduire la corrélation entre les pseudonymes issus de différents jeux de données, réduisant en conséquence, voire éliminant totalement, les caractéristiques statistiques des bases de données pseudonymisées. Elles permettent en outre de limiter la capacité de mise en corrélation entre des enregistrements de données différents (potentiellement répartis sur plusieurs entreprises) et un même profil d'utilisateur. Ainsi, même lorsqu'une pseudonymisation aléatoire est appliquée, SSL peut être à même de lancer les attaques précédemment discutées, s'il est capable de déterminer si deux pseudonymes différents appartiennent au même identificateur. De même, les entreprises B et D peuvent ré-identifier avec succès la personne concernée masquée derrière des profils d'utilisateur partagés. Ici, le compromis entre protection et fonctionnalité est à nouveau évident.

La question est la suivante: comment se protéger efficacement contre ce type d'attaque ciblant la pseudonymisation?

D'après l'analyse effectuée dans le présent rapport, la meilleure approche en matière de pseudonymisation consiste à:

- Prendre en compte l'ensemble du jeu de données disponible
- Connaître la taille des valeurs de données individuelles sur le domaine d'entrée
- Appliquer la pseudonymisation sur l'ensemble des valeurs de données, de manière à rendre les attaques par force brute et par dictionnaire irréalisables
- Éliminer toutes les possibilités d'application des attaques par connaissances contextuelles ou distribution statistique
- Concevoir une fonction de pseudonymisation à grande échelle de façon à ce que le jeu de données pseudonymisé n'offre que la fonctionnalité strictement nécessaire aux fins de traitement requises, et éliminer toute autre fonctionnalité

Dans l'exemple de scénario présenté dans les chapitres précédents, la société RS utilise un schéma de pseudonymisation qui pseudonymise non seulement les adresses IP, mais également toutes les combinaisons adresse IP/horodatage possibles. Dans un tel cas, il devient impossible de mettre en corrélation un horodatage avec une source de données externes, car ces informations ne sont plus accessibles pour la société SSL. Pour réaliser avec succès une ré-identification, SSL devrait connaître (ou deviner par conjecture) la combinaison adresse IP/horodatage exacte. En effet, il est très peu probable qu'un adversaire parvienne à briser la pseudonymisation d'une combinaison d'entrées de données sans connaître (ou deviner) toutes les données d'entrée en texte clair. Dans une telle situation, ce type de

[33] Le terme de «device fingerprinting» (détermination de l'identité d'un appareil/logiciel) décrit ce risque de sécurité.

pseudonymisation permettrait de bloquer bien plus efficacement toutes les tentatives de SSL de récupérer un pseudonyme donné.

Des exemples de technique de base pour des fonctions de pseudonymisation robustes ont déjà été présentées au chapitre 5, avec une discussion approfondie sur leur résilience face aux attaques exposées au chapitre 4. Pour étendre ces techniques à des enregistrements de données structurés, il suffit souvent de prendre comme entrée l'ensemble de l'enregistrement de données et de lui appliquer une combinaison personnalisée de fonctions de hachage par clés et de techniques d'anonymisation courantes. Des techniques de pseudonymisation plus avancées ont été rapidement présentées au chapitre 5.6 ainsi que dans un précédent rapport de l'ENISA [2].

9. CONCLUSIONS ET RECOMMANDATIONS

Dans le contexte du Règlement général sur la protection des données (RGPD), la mise en œuvre appropriée d'une pseudonymisation des données à caractère personnel est en passe de devenir un sujet hautement débattu au sein de nombreuses communautés, qu'il s'agisse du monde des sciences, du monde académique, de la justice ou des organismes d'application des lois, ainsi que dans la gestion de la conformité pour diverses organisations européennes. Ce rapport a permis d'introduire quelques notions de base, ainsi que les définitions, les techniques, les attaques et les contre-mesures qui sous-tendent ce débat interdisciplinaire d'avenir.

Comme ce rapport l'indique, le domaine de la pseudonymisation des données au sein d'infrastructures de l'information complexes est une véritable gageure, qui dépend en grande partie du contexte, des entités impliquées, des informations contextuelles et des paramètres de mise en œuvre. Il n'existe en réalité aucune solution de pseudonymisation simple et universelle qui fonctionnerait dans tous les cas de figures. Au contraire, un haut niveau de compétences est nécessaire pour mettre en œuvre un processus de pseudonymisation robuste, ayant la capacité potentielle à réduire au mieux le risque de discrimination et les attaques de ré-identification, tout en offrant le degré de fonctionnalité nécessaire pour assurer le traitement des données pseudonymisées.

À cette fin, sur la base de l'analyse offerte par le présent rapport, des conclusions et des recommandations destinées à l'ensemble des parties prenantes se dessinent, en vue de l'adoption et de la mise en œuvre d'une solution de pseudonymisation des données.

ADOPTER UNE APPROCHE BASEE SUR LES RISQUES EN MATIERE DE PSEUDONYMISATION

Bien que chaque technique de pseudonymisation connue présente ses propres propriétés intrinsèques et bien établies, le choix de la technique appropriée n'en est pas pour autant plus facile. Il est en effet nécessaire d'effectuer un examen attentif du contexte dans laquelle la pseudonymisation sera effectuée, de prendre en compte tous les objectifs souhaités pour chaque processus de pseudonymisation spécifique (de qui doit-on protéger les identités, quelle est la fonctionnalité souhaitée pour les pseudonymes dérivés, etc.), ainsi que la facilité de mise en œuvre. Il est donc essentiel d'adopter une approche basée sur les risques dans le choix de la technique de pseudonymisation appliquée, afin de correctement évaluer et minimiser les risques pour la confidentialité des usagers. En effet, si protéger les données supplémentaires nécessaires à la ré-identification est une étape obligée, elle ne suffit en rien à éliminer tous les risques.

Il est de la responsabilité des sous-traitants et des responsables du traitement des données d'étudier une approche basée sur les risques pour la mise en œuvre d'une solution de pseudonymisation, en prenant en compte la finalité et le contexte global du processus de traitement des données à caractère personnel, ainsi que les niveaux de fonctionnalité et d'évolutivité souhaités.

Les fournisseurs de produits, de services et d'applications doivent transmettre aux sous-traitants et aux responsables du traitement des informations adéquates concernant leur utilisation des techniques de pseudonymisation, ainsi que les niveaux de protection des données et de sécurité que celles-ci offrent.

Les organismes de régulation (par ex. les autorités de protection des données et le Comité européen de la protection des données) doivent fournir aux responsables du traitement des données et aux sous-traitants des consignes pratiques concernant l'évaluation des risques, tout en assurant la promotion des meilleures pratiques en matière de pseudonymisation.

DEFINIR L'ETAT DES CONNAISSANCES

Pour pouvoir adopter une approche basée sur les risques dans le domaine de la pseudonymisation, il est essentiel de définir l'état des connaissances en la matière. En effet, bien qu'il existe diverses techniques de pseudonymisation comme illustré dans le présent rapport, l'application pratique de ces techniques peut varier, notamment suivant le type d'identificateur et de jeu de données concerné. À cette fin, il est important de travailler sur des exemples et des cas d'utilisation précis, pour bénéficier d'un maximum de détails et d'options possibles sur l'implémentation technique de la pseudonymisation.

Il est de la responsabilité de la Commission européenne et des institutions pertinentes au sein de l'Union européenne de favoriser la définition et la diffusion de l'état des connaissances en matière de pseudonymisation, en collaboration avec la communauté des chercheurs et l'industrie dans ce domaine.

Les organismes de régulation (par ex. les autorités de protection des données et le Comité européen de la protection des données) ont pour mission de promouvoir la publication des meilleures pratiques en matière de pseudonymisation.

FAIRE PROGRESSER L'ETAT DES CONNAISSANCES

Bien que ce rapport ait pour thème central les techniques de base de la pseudonymisation actuellement disponibles pour les responsables du traitement des données et les sous-traitants, il existe des cas de figure plus complexes (qui, comme l'illustre le rapport, sont assez fréquents dans la pratique) qui exigent l'utilisation de techniques plus avancées (et plus robustes), telles que celles issues des pratiques d'anonymisation. Plus encore, la notion même d'anonymisation doit aujourd'hui être revisitée, face à l'évolution des modèles d'adversaire, qui remettent en question l'efficacité des techniques existantes en la matière.

Le milieu de la recherche doit travailler à développer les techniques de pseudonymisation actuelles pour obtenir des solutions plus sophistiquées, prenant en charge avec efficacité les problématiques spécifiques associées à l'ère du Big Data. La Commission européenne et les institutions concernées au sein de l'UE doivent quant à elles soutenir et généraliser ces efforts.

Références

[1] M. J. Dworkin, «SHA-3 Standard: Permutation-Based Hash and Extendable-Output Functions», 2015.

[2] A. Pfitzmann et M. Hansen, «A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management», 2010.

[3] A. Machanavajjhala, J. Gehrke, D. Kifer et M. Venkitasubramaniam, «L-diversity: Privacy beyond k-anonymity», *22nd International Conference on Data Engineering (ICDE'06)*, 2006.

[4] M. Barbaro, T. Zeller et S. Hansell, *A face is exposed for aol searcher no. 4417749,* vol. 9, New York Times, 2006, p. 8.

[5] M. Hellman, «A cryptanalytic time-memory trade-off», *IEEE transactions on Information Theory,* vol. 26, no. 4, pp. 401-406, 1980.

[6] J. L. Massey, «Guessing and entropy», *Proceedings of 1994 IEEE International Symposium on Information Theory*, 1994.

[7] D. G. Malone et W. Sullivan, «Guesswork and entropy», *IEEE Transactions on Information Theory,* vol. 50, no. 3, pp. 525--526, 2004.

[8] H. C. Van Tilborg et S. Jajodia, Encyclopedia of cryptography and security, Springer Science & Business Media, 2014.

[9] J. Katz, A. J. Menezes, P. C. Van Oorschot et S. A. Vanstone, Handbook of applied cryptography, CRC Press, 1996.

[10] M. Bellare, R. Canetti et H. Krawczyk, «Keying hash functions for message authentication», in *Annual international cryptology conference*, 1996.

[11] L. Demir, A. Kumar, M. Cunche et C. Lauradoux, «The pitfalls of hashing for privacy», *IEEE Communications Surveys & Tutorials,* pp. 551-565, 2017.

[12] H. Krawczyk, R. Canetti et M. Bellare, «HMAC: Keyed-Hashing for Message Authentication», *RFC,* pp. 1-11, 1997.

[13] N. Li, T. Li et S. Venkatasubramanian, «t-closeness: Privacy beyond k-anonymity and l-diversity», *23rd International Conference on Data Engineering*, 2007.

[14] N. Li, T. Li et S. Venkatasubramanian, «t-closeness: Privacy beyond k-anonymity and l-diversity», *23rd International Conference on Data Engineering*, 2007.

[15] I. Polakis, G. Kontaxis, S. Antonatos, E. Gessiou, T. Petsas et E. P. Markatos, «Using social networks to harvest email addresses», *Proceedings of the 9th Annual ACM Workshop on Privacy in the Electronic Society*, 2010.

[16] T. Eastlake et D. Hansen, «US Secure Hash Algorithms (SHA and SHA-based HMAC and HKDF)», 2011.

[17] A. Narayanan et V. Shmatikov, «Robust De-anonymization of Large Sparse Datasets», *IEEE Symposium on Security and Privacy*, 2008.

[18] Recommandations de l'ENISA sur l'usage des technologies conformément au RGPD- Une introduction à la pseudonymisation des données, Athènes, 2018.

[19] WP29, «Groupe de travail "Article 29" sur la protection des données: avis 4/2007 sur le concept de données à caractère personnel», 2007

[20] IETF, «Internet Engineering Task Force: RFC8200, Internet Protocol, Version 6 (IPv6) Specification», STD 86, 2017.

[21] IETF, «Internet Engineering Task Force: RFC4632, Classless Inter-domain Routing (CIDR): The Internet Address Assignment and Aggregation Plan», BCP 122, 2006.

[22] IETF, «Internet Engineering Task Force: RFC 5735, Special Use IPv4 Addresses», 2010.

[23] R. C. Merkle, «A Digital Signature Based on a Conventional Encryption Function», *Advances in Cryptology — CRYPTO '87,* pp. 369-378, 1988.

[24] G. Becker, «Merkle Signature Schemes, Merkle Trees and Their Cryptanalysis», Bochum, 2008.

[25] L. Lamport, «Password authentication with insecure communication», *Communications of the ACM,* pp. 770-772, Novembre 1981.

[26] B. H. Bloom, «Space/time trade-offs in hash coding with allowable errors», *Communications of the ACM,* pp. 422-426, Juillet 1970.

[27] L. Sweeney, «K-anonymity: A model for protecting privacy», *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* vol. 10, no. 5, pp. 557-570, 2002.

[28] L. Sweeney, «Only You, Your Doctor, and Many Others May Know», *Technology Science,* vol. 2015092903, no. 9, p. 29, 2015.

[29] C. Dwork et A. Roth, «The Algorithmic Foundations of Differential Privacy», *Foundations and Trends in Theoretical Computer Science,* pp. 211-407, Août 2014.

[30] R. Noumeir, A. Lemay et J.-M. Lina, «Pseudonymization of radiology data for research purposes», *Journal of digital imaging,* vol. 20, no. 3, pp. 284-295, 2007.

[31] Y. Yona et S. Diggavi, «The effect of bias on the guesswork of hash functions», *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017.

[32] ENISA, «Privacy and data protection by design - from policy to engineering», 2014.

[33] Sommet numérique, Groupe de discussion sur la protection des données, «White Paper on Pseudonymization», 2017.

[34] P. Oechslin, «Making a Faster Cryptanalytic Time-Memory Trade-off», *CRYPTO 2003*, 2003.

[35] IETF, «Internet Engineering Task Force: RFC 791, Internet Protocol DARPA Internet Program Protocol Specification», 1981.

[36] IETF, «Internet Engineering Task Force: IPFIX Working Group, IP Flow Anonymization Support», 2011.

[37] «An Analysis of Google Logs Retention Policies», *Journal of Privacy and Confidentiality,* vol. 3, no. 1, 2011.

[38] S. Weber, «On Transaction Pseudonyms with Implicit Attributes», *Cryptology ePrint Archive: Report 2012/568, https://eprint.iacr.org/2012/568,* 2012.

[39] W. H. a. F. W. F, «A Survey of Noninteractive Zero Knowledge Proof System and Its Applications», *The Scientific World Journal,* 2014.

À PROPOS DE L'ENISA

Créée en 2004, l'Agence de l'Union européenne pour la cybersécurité (ENISA) a pour mission de garantir la cybersécurité au sein de l'Europe. L'ENISA travaille en collaboration avec l'Union européenne, ses États membres, le secteur privé et les citoyens européens afin d'établir des conseils et des recommandations sur les bonnes pratiques à appliquer en matière de sécurité de l'information. L'agence aide les États membres de l'UE à appliquer la législation européenne en vigueur, et s'efforce de renforcer la protection des infrastructures et réseaux d'information critiques en Europe. L'ENISA vise à améliorer les compétences existantes au sein des États membres de l'UE en favorisant le développement de communautés transfrontalières ayant pour vocation d'optimiser la sécurité des réseaux et de l'information à travers l'UE. Depuis 2019, elle met en place des programmes de certification en matière de cybersécurité. De plus amples informations sur l'ENISA et ses travaux peuvent être consultées à l'adresse: www.enisa.europa.eu.

ENISA
European Union Agency for Cybersecurity

Athens Office
1 Vasilissis Sofias Str
151 24 Marousi, Attiki, Greece

Heraklion office
95 Nikolaou Plastira
700 13 Vassilika Vouton, Heraklion, Greece

enisa.europa.eu

ISBN 978-92-9204-307-0
doi: 10.2824/247711