Anonymat, anonymisation, désanonymisation

Guillaume Piolle
guillaume.piolle@centralesupelec.fr
http://guillaume.piolle.fr/

CentraleSupélec, campus de Rennes

5 décembre 2016

Sécurité informatique et vie privée

La protection de la vie privée est-elle
une composante de la sécurité informatique ?

Dimensions classiques de la sécurité informatique

- Confidentialité ;
- Intégrité ;
- Disponibilité ;
- + authentification, non-répudiation, contrôle d'accès, contrôle de flux. . .

Sécurité informatique et vie privée

Dimensions de la protection des données personnelles

- Information ;
- Consentement ;
- Droit d'accès/rectification/suppression ;
- Finalité et proportionnalité ;
- Durée de rétention ;
- Transmission aux tiers.

Sécurité informatique et vie privée

- La protection de la vie privée (ou des données personnelles) peut être considérée **du point de vue de la sécurité informatique** ;
- Certaines exigences de la vie privée **peuvent être remplies** grâce aux outils classiques de la sécurité informatique ;
- Certaines exigences de la vie privée **ne peuvent pas être remplies** grâce aux outils classiques de la sécurité informatique ;
- Certaines exigences de vie privée sont **incompatibles** avec certaines exigences de la sécurité informatique.

Parfois présentée comme une **sous-discipline**, parfois comme une discipline **connexe** ou **transverse**, parfois comme une discipline **concurrente**.

Conflits entre sécurité et vie privée

Besoin d'auditabilité

Un impératif de la sécurité informatique : se donner les moyens de détecter les comportements malveillants ou erronés et de désigner des responsables.

Outil : conservation de **journaux** (logs) retraçant l'activité d'un système (logiciel, serveur web, etc.).

Exemple de journal d'authentification (/var/log/auth.log)

```
Sep 29 20:59:01 vpsxxx CRON[14089]: pam_unix(cron:session): session opened for user root
                                    by (uid=0)
Sep 29 20:59:01 vpsxxx CRON[14090]: pam_unix(cron:session): session opened for user root
                                    by (uid=0)
Sep 29 20:59:01 vpsxxx CRON[14090]: pam_unix(cron:session): session closed for user root
Sep 29 20:59:01 vpsxxx CRON[14089]: pam_unix(cron:session): session closed for user root
Sep 29 20:59:46 vpsxxx sshd[14140]: Did not receive identification string
                                    from 161.139.xxx.xxx
Sep 29 21:00:01 vpsxxx CRON[14141]: pam_unix(cron:session): session opened for user root
                                    by (uid=0)
Sep 30 11:53:18 vpsxxx sshd[6842]: Authentication tried for root with correct key
                                   but not from a permitted host
                                   (host=nat-profs.rennes.supelec.fr, ip=193.54.192.3).
Sep 30 11:53:18 vpsxxx sshd[6842]: Authentication tried for root with correct key
                                   but not from a permitted host
                                   (host=nat-profs.rennes.supelec.fr, ip=193.54.192.3).
Sep 30 11:53:21 vpsxxx sshd[6842]: Accepted password for root from 193.54.192.3
                                   port 55130 ssh2
Sep 30 11:53:21 vpsxxx sshd[6842]: pam_unix(sshd:session): session opened
                                   for user root by (uid=0)
```

Attention, dans les journaux les adresses MAC et IP sont complètes.

Exemple de journal de pare-feu (dans /var/log/messages)

```
Sep 26 10:33:55 vpsxxx kernel: netfilter-input IN=eth0 OUT= MAC=[masqué]
SRC=90.84.xxx.xxx DST=46.105.yyy.yyy LEN=60 TOS=0x00 PREC=0x00 TTL=48
ID=25922 DF PROTO=TCP SPT=59766 DPT=8080 WINDOW=5840 RES=0x00 SYN URGP=0

Sep 26 10:34:47 vpsxxx kernel: netfilter-input IN=eth0 OUT= MAC=[masqué]
SRC=90.84.xxx.xxx DST=46.105.yyy.yyy LEN=60 TOS=0x00 PREC=0x00 TTL=48
ID=20314 DF PROTO=TCP SPT=55315 DPT=8080 WINDOW=5840 RES=0x00 SYN URGP=0

Sep 26 10:34:48 vpsxxx kernel: netfilter-input IN=eth0 OUT= MAC=[masqué]
SRC=90.84.xxx.xxx DST=46.105.yyy.yyy LEN=60 TOS=0x00 PREC=0x00 TTL=48
ID=20315 DF PROTO=TCP SPT=55315 DPT=8080 WINDOW=5840 RES=0x00 SYN URGP=0
```

Exemple de journal Apache (/var/log/apache2/access.log)

```
212.113.aaa.aaa - [28/Sep/2011:15:53:07 +0200] "GET / HTTP/1.1" 200 460 "-" "Mozilla/5.0
(compatible; MJ12bot/v1.4.0; http://www.majestic12.co.uk/bot.php?+)"
188.165.bbb.bbb - [28/Sep/2011:20:07:05 +0200] "GET /phpmyadmin/main.php HTTP/1.0" 200 9773
"-" "-"
188.165.bbb.bbb - [28/Sep/2011:20:07:07 +0200] "GET /phpmyadmin/libraries/select_lang.lib.p
HTTP/1.0" 403 522 "-" "-"
188.40.ccc.ccc - [28/Sep/2011:20:21:55 +0200] "GET / HTTP/1.0" 200 453 "-" "-"
188.40.ccc.ccc - [28/Sep/2011:20:21:55 +0200] "GET / HTTP/1.0" 200 453 "-" "-"
188.40.ccc.ccc - [28/Sep/2011:20:21:56 +0200] "HEAD / HTTP/1.1" 200 276 "-"
"Visited by http://tools.geek-tools.org"
188.165.ddd.ddd - [28/Sep/2011:22:49:36 +0200] "GET /w00tw00t.at.ISC.SANS.test0:)
HTTP/1.1"400 513 "-" "-"
213.91.eee.eee - [29/Sep/2011:00:37:47 +0200] "GET /w00tw00t.at.ISC.SANS.DFind:)
HTTP/1.1" 400 513 "-" "-"
85.214. fff . fff - [29/Sep/2011:10:03:39 +0200] "GET /w00tw00t.at.ISC.SANS.DFind:)
HTTP/1.1" 400 513 "-" "-"
209.160.ggg.ggg - [29/Sep/2011:15:30:17 +0200] "GET / HTTP/1.0" 206 498 "-"
"Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"
```

Obligations de journalisation

- 2001, Loi sur la Sécurité Quotidienne (LSQ) : les opérateurs télécom doivent conserver les données de connexion pendant un an (mesure temporaire, prolongée *ad vitam*) ;
- 2004, Loi sur la Confiance dans l'Économie Numérique (LCEN) : conservation des informations identifiant les personnes déposant des contenus en ligne (étendu à tous les fournisseurs d'accès) ;
- 2011, décret d'application de la LCEN : conservation des identifiants, peudonymes, mots de passe, données de paiement, coordonnées (étendu aux hébergeurs et éditeurs de sites web).

En cas de journalisation insuffisante ?

Jusqu'à 375 k€ d'amende pour une société, 75 k€ et un an d'emprisonnement pour son dirigeant.

Qui peut accéder aux journaux ?

- La justice (commission rogatoire, décision en référé ou en instance) ;
- La police, sur réquisition simple (sans autorisation judiciaire), depuis la loi du 23 janvier 2006 sur la lutte contre le terrorisme ;
- L'administrateur système/réseau, qui « est tenu d'une **obligation de confidentialité** » (même vis-à-vis de l'employeur, en tout cas en ce qui concerne les e-mails) et peut accéder aux données « dans le cadre de sa mission de sécurité du réseau informatique » (Cour de Cassation, 17 juin 2009).

Risque opérationnel aggravé

Pour des prétextes de sécurité (lutte contre le terrorisme), on augmente le risque de dommages en cas d'intrusion et on fournit une incitation aux attaquants éventuels.

Compromis entre vie privée et d'autres notions

- Liberté d'expression ;
- Droit à l'information ;
- Diffusion et disponibilité des données (ex. des réseaux sociaux distribués) ;
- Utilisabilité (procédures, politiques de sécurité).

Complémentarités entre sécurité et vie privée

Les acteurs sont souvent les mêmes. . .

. . . mais pas toujours.

Concepteurs de systèmes, administrateurs, RSSI, fournisseurs d'accès ou de services. . . Quels sont les impératifs de chacun ?

Les problématiques sont proches

Les problèmes « durs » en sécurité et en protection des données personnelles sont quasiment les mêmes : contrôle de flux et contrôle d'usage efficaces et distants.

Complémentarités entre sécurité et vie privée

L'auditabilité au service de la vie privée

- Article 39 de la loi Informatique et Libertés ;
- Obligations de notification des brèches de vie privée ;
- Importance de démontrer sa capacité à respecter données personnelles et vie privée.

Les outils de la sécurité peuvent servir la vie privée

- Outils de chiffrement classiques ;
- Cryptologie plus exotique (IBE et consorts, garanties de répudiabilité ou de dénégation plausible, signature de groupe, chiffrement homomorphe. . .) ;
- Recherche convergente en contrôle de flux et d'usage.

Les critères communs de l'ISO

Common Criteria for Information Technology Security Evaluation

Norme ISO/IEC 15408, successeur de l'*Orange Book* du DoD.
Section 7 : protection de la vie privée.

Exigences techniques pour assurer la vie privée

- **Anonymat** (*anonymity*) : incapacité d'un observateur à déterminer l'identité d'un utilisateur ;
- **Pseudonymat** (*pseudonymity*) : idem, mais l'utilisateur continue à répondre de ses actions ;
- **Non-traçabilité** (*unlinkability*) : incapacité d'un observateur à déterminer si deux actions ont été réalisées par le même utilisateur ;
- **Non-observabilité** (*unobservability*) : incapacité d'un observateur à déterminer si une action est en cours.

Anonymat

Intérêt et nécessité de l'anonymat

- Protection physique des personnes (témoins de crimes, personnes menacées) ;
- Protection morale des personnes (accouchement sous X, personnes citées dans la presse) ;
- Protection pénale des personnes (*whistleblowing*, sources des journalistes) ;
- ...

L'anonymat total permet l'impunité par définition, ce qui peut être une bonne ou une mauvaise chose suivant le contexte.

Tous ces exemples relèvent-ils vraiment de l'anonymat ?
Qu'en est-il de l'anonymat sur Internet ?
Y a-t-il une réelle « impunité » en ligne ?

Pseudonymat

Parfois appelé « anonymat révocable »
Souvent confondu avec l'anonymat, notamment sur Internet.

Intérêt et nécessité du pseudonymat

- En première approximation, mêmes avantages que l'anonymat ;
- Possibilité de lever le pseudonymat (non-répudiation, auditabilité).

Quelle « quantification » pour le degré de pseudonymat ?
Qui peut lever le pseudonymat ?

Voir les argumentaires de Maître Eolas (*Journal d'un Avocat*, `http://www.maitre-eolas.fr/` ou de Zythom (*Blog d'un informaticien expert judiciaire*, `http://zythom.blogspot.fr/`) sur la légitimité du pseudonymat.

Bases de données anonymes ou anonymisées

- Absence de données permettant d'identifier une personne de manière unique :
  - Retrait des nom et prénom ;
  - Remplacement par un numéro aléatoire ;
  - Remplacement par des pseudonymes arbitraires. . .
- Sondages anonymes, officiels ou non ;
- Sondages et questionnaires dont la partie identifiante est ensuite désolidarisée du reste.

Cadre juridique des bases de données anonymes

Si l'on considère qu'il n'y a pas de « données à caractère personnel » parce qu'il n'y a pas d'éléments identifiants, alors la loi Informatique et Libertés **ne s'applique pas** !

Conséquence :

- **Aucun droit** pour les personnes concernées ;
- **Aucune obligation** pour les responsables de traitements ;
- **Aucune restriction** à la conservation, la publication, l'exploitation, le rapprochement avec d'autres bases de données.

Mais. . . aucun problème puisque tout est anonyme ?

L'anonymisation parfaite est impossible

L'anonymisation d'une base de données
NE SUFFIT PAS À EMPÊCHER
L'IDENTIFICATION DES INDIVIDUS
dans la majorité des cas.

On considère généralement que l'anonymisation est une opération impossible dans le cas général et que le terme est donc impropre. On préfère souvent parler de « pseudo-anonymisation » ou de « désidentification » (ou de « pseudonymisation », mais cela peut désigner autre chose).

L'affaire des requêtes AOL

Juillet 2006

AOL rend public un historique de 20 millions de requêtes (3 mois pour 658000 utilisateurs).

De bonnes intentions

- Mise à la disposition de la communauté scientifique, à fins de recherche ;
- Requêtes « anonymisées » (mais identifiant unique par utilisateur).

L'affaire des requêtes AOL

Diffusion des données

- Données republiées par de nombreux sites miroirs ;
- Développement de nombreux outils d'analyse spécialisés.

Désanonymisation

De nombreux utilisateurs sont identifiés.

L'utilisateur n° 4417749 a interrogé AOL sur :

- Les célibataires de 60 ans ;
- Les tremblements de la main ;
- Les effets de la nicotine ;
- Les chiens qui urinent partout ;
- ...

L'affaire des requêtes AOL

L'utilisateur n° 4417749

Thelma Arnold, 62 ans, Lilburn, Georgie.

Identifiée par le *New York Times*, identité publiée avec l'autorisation de la personne concernée.

Picture by Erik. S. Lesser, NYT

L'affaire des requêtes AOL

Conséquences

Nombreuses compilations de requêtes, exotiques ou artistiques, œuvres dérivées.

Impact commercial désastreux pour AOL, deux licenciements, démission du *Chief Technology Officer*, une *class action* en Californie.

Impact pédagogique : accent mis sur le risque lié aux moteurs de recherche, pas sur la désanonymisation de données (peu technique ici).

L'affaire Netflix

Une démarche inoffensive

2010 : Netflix publie des données d'évaluation anonymes, dans le cadre d'un concours (*Netflix prize*, 1M$) visant à améliorer son algorithme de recommandation.

Un chercheur recoupe les données anonymes avec celles du site IMDb et « désanonymise » la base. Les goûts cinématographiques des utilisateurs deviennent des données identifiantes !
La connaissance de deux notes suffit à identifier 68 % des utilisateurs.

Plainte fédérale, Netflix se rétracte et met fin au concours pour cause de risque pour la vie privée.

Les aventures de Latanya Sweeney, épisode 1

L'affaire du GIC

Milieu des années 90 : le *Group Insurance Commission* du Massachusetts décide de rendre publiques des données « anonymisées » concernant les hospitalisations des employés de l'état.

L. Sweeney, étudiante à Carnegie Mellon, recoupe ces données avec les listes électorales et envoie le détail de son dossier médical au gouverneur.

Le gouverneur fait faire marche arrière au GIC...

Les aventures de Latanya Sweeney, épisode 2

Naissance du k-anonymat

2000 : L. Sweeney montre que 87 % des citoyens U.S. peuvent être identifiés de manière unique par leur sexe, leur date de naissance et leur code postal (recoupements faciles avec les registres publics).

Publication en 2002 : introduction du concept de *k-Anonymity*, qui mesure de manière mathématique le degré d'anonymat d'une base de données « anonyme ».

Exemple de base de données « anonyme »

Sondage anonyme (fictif) sur les étudiants d'un campus

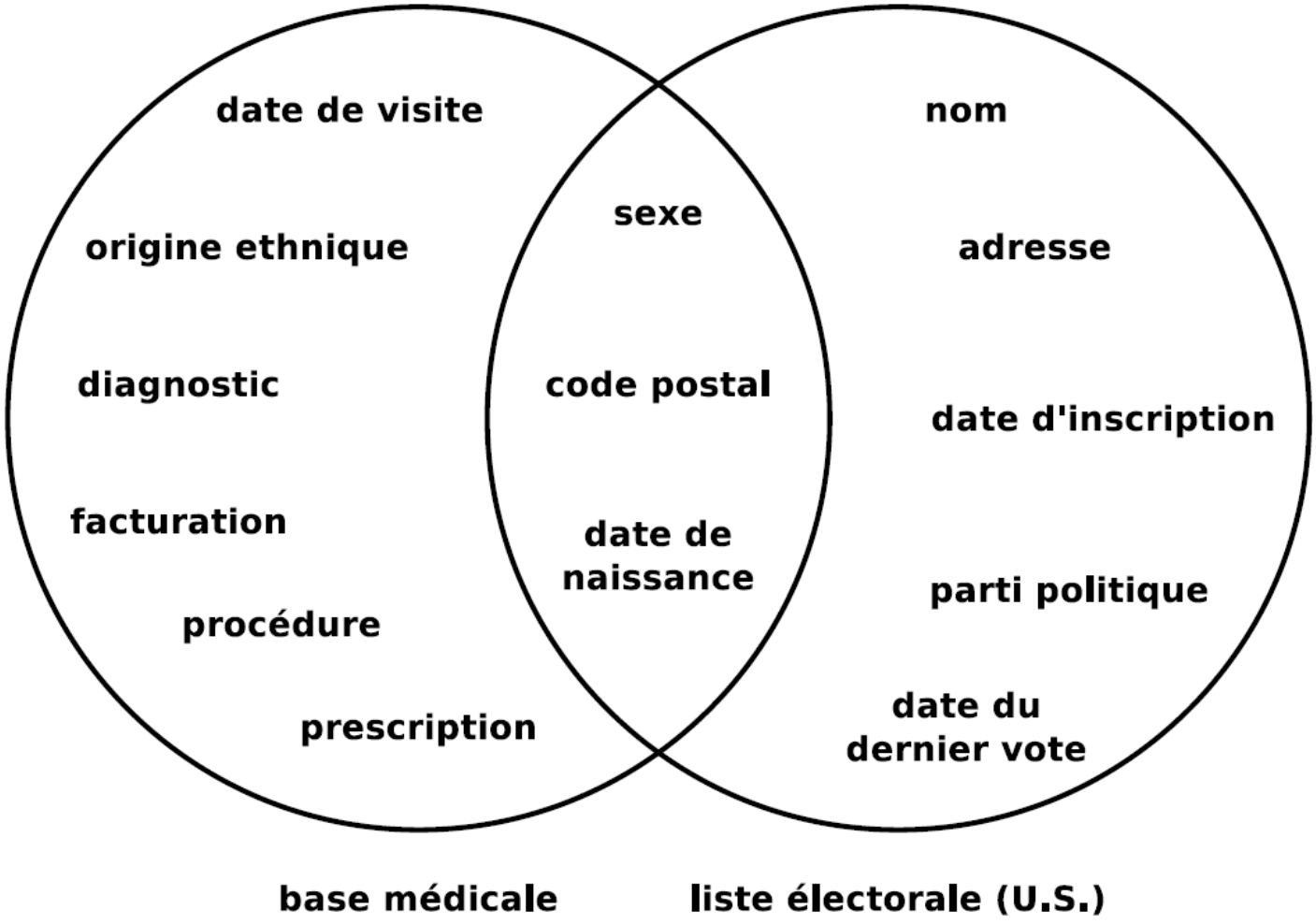
Sexe	Taille	Orientation sexuelle
...	...	...
M	170-180	hétérosexuel
M	170-180	hétérosexuel
M	170-180	hétérosexuel
M	170-180	bisexuel
M	170-180	hétérosexuel
M	170-180	hétérosexuel
M	170-180	hétérosexuel
M	170-180	hétérosexuel
M	180-190	hétérosexuel
M	180-190	homosexuel
M	180-190	hétérosexuel
M	180-190	autre
M	180-190	hétérosexuel
M	190-200	hétérosexuel
M	200-210	homosexuel

Orientation sexuelle : **sensible** au sens de l'article 8 de la loi « Informatique et Libertés ».

MAIS : sondage complètement anonyme... donc hors du champ de la loi !

- Réalité de cet « anonymat » ?
- Les étudiants sont-ils tous égaux devant cet « anonymat » ?
- À quelles questions répondez-vous lors de sondages « anonymes » ?

Problème central : l'interconnexion de bases de données



Principes du k-anonymat

Quasi-identifiant

Ensemble d'attributs d'une base de données pouvant permettre, dans au moins un cas, d'identifier un tuple à l'aide d'informations externes.

N'importe quel attribut peut appartenir à un quasi-identifiant !

Nouvel éclairage sur la notion de « donnée à caractère personnel » (art. 2 de la loi « Informatique et Libertés »).

k-Anonymat

Une base de données est dite **k-anonyme** si tout tuple est indistingable d'au minimum $k - 1$ autres tuples de la base projetée sur tout quasi-identifiant.

Retour sur la base d'exemple

Sondage anonyme (fictif) sur les étudiants d'un campus

Sexe	Taille	Orientation sexuelle
...	...	...
M	170-180	hétérosexuel
M	170-180	hétérosexuel
M	170-180	hétérosexuel
M	170-180	bisexuel
M	**170-180**	**hétérosexuel**
M	170-180	hétérosexuel
M	170-180	hétérosexuel
M	170-180	hétérosexuel
M	180-190	hétérosexuel
M	180-190	homosexuel
M	180-190	hétérosexuel
M	180-190	autre
M	180-190	hétérosexuel
M	190-200	hétérosexuel
M	200-210	homosexuel

La base est **1-anonyme** : c'est le pire cas !

Au moins une personne (deux ici) est ré-identifiable à l'aide d'une base de données externe facile à concevoir.

On peut dire que l'individu en gras est 8-anonyme dans la base.

k-anonymisation

À partir d'une base 1-anonyme, obtention d'une base « publiable » k-anonyme :

- Falsification ou obfuscation de certains attributs ;
- Agrégation ;
- Projection relationnelle ;
- Introduction de tuples artificiels...

Autres techniques (pas forcément liées au k-anonymat) :

- Injection de bruit dans certains attributs ;
- Sélection relationnelle...

Attention !

- En k-anonymisant, on limite l'intérêt (utilité) de la base ;
- La taille du quasi-identifiant dépend de l'ensemble des bases dans le monde extérieur.

k-anonymisation

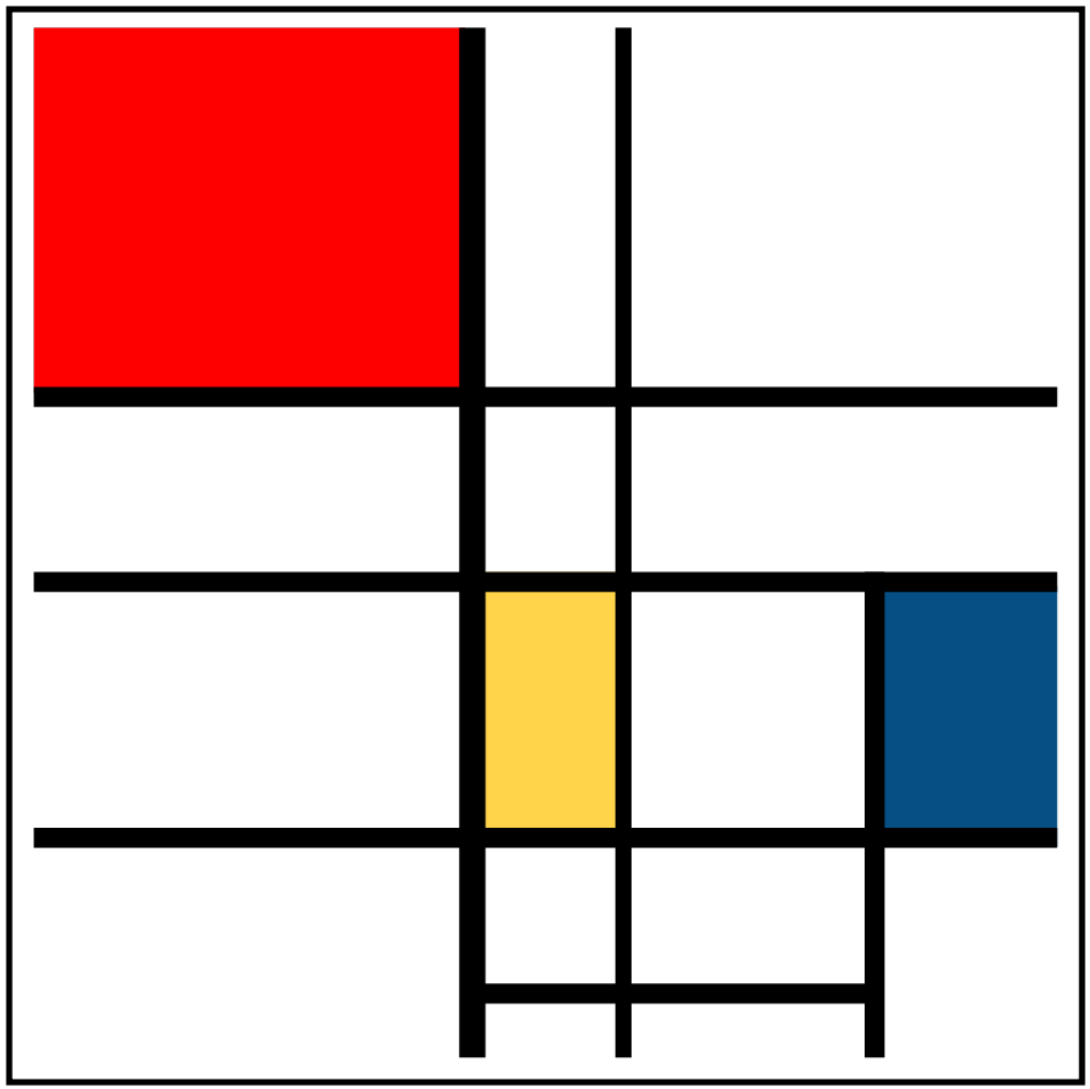
Exemple de k-anonymisation : algorithmes de type Mondrian

- **Hypothèse :** un attribut de la base est soit partie d'un QI, soit une donnée sensible ;
- **Objectif :** rendre le plus équivoque possible le lien entre une valeur de QI et les données sensibles correspondantes ;
- **Mécanisme :** partitionner l'espace des QI de manière à former des groupes d'au moins k éléments, puis remplacer dans la BD les QI par l'identifiant de la partition.

- On part de l'ensemble de tous les QI ;
- On partitionne au fur et à mesure (par dichotomie par exemple) jusqu'à obtenir un maximum de zones de k éléments au minimum.

Ces algorithmes sont paramétrables par les caractéristiques des distributions dans chaque zone, par exemple.

k-anonymisation



Attaques contre les bases k-anonymes

Homogeneity attack

Age	CP	Diagnostic
...	...	...
20-25	35000	Colite
20-25	35000	Liposarcome
20-25	35000	Rhume des foins
20-25	35000	Entorse
25-30	35000	Grippe aviaire
25-30	35000	Angine virale
25-30	35000	Coqueluche
25-30	35000	Pneumonie
25-30	35510	Syphilis
25-30	35510	Syphilis
25-30	35510	Syphilis
25-30	35510	Syphilis

La base est 4-anonyme, mais pourtant on peut apprendre avec certitude des informations sensibles sur certains individus.

La cause en est une trop grande homogénéité dans les résultats de certaines classes (pas assez de **diversité**.

Attaques contre les bases k-anonymes

l-diversité (*l-diversity*)

Une classe (dans une base k-anonyme, par exemple), est dite **l-diverse** s'il y a au moins *l* valeurs *bien représentées* pour l'attribut sensible.

« *l* valeurs *bien représentées* » peut signifier :

- qu'il y a au moins *l* valeurs distinctes ;
- que l'entropie de la classe (par référence à l'espace des valeurs pour l'attribut sensible) est supérieure à $log_2(l)$;
- que, suivant d'autres métriques ((c-l)-diversité), la valeur la plus courante n'apparaît pas *trop fréquemment* et que la valeur la moins courante n'est pas *trop rare* ;
- ...

Attaques contre les bases k-anonymes

Complementary release attack

Corrélation de deux extraits de la même base, k-anonymisés de manière différente.

Unsorted matching attack

Valide uniquement pour des bases où l'ordre des tuples a un sens.
Principe : on effectue des correspondances entre deux bases de données k-anonymisées, mais dont les tuples sont publiés dans le même ordre.

Attaques temporelles

Comparaison du résultat de requêtes faites à des moments différents, corrélées avec des informations extérieures.

Comment éviter les attaques de type *data-mining* ?

Restriction sur les requêtes multiples

Base de départ : médecin $\times$ patient $\times$ prescription

- Une requête patient/prescription est très sensible (interdite) ;
- Une requête médecin/patient est moyennement sensible ;
- Une requête médecin/prescription est peu sensible.

Problème

Corrélation possible entre les résultats des deux dernières requêtes, possibilité de désanonymisation partielle.
Faut-il interdire toutes les requêtes portant sur les prescriptions ?

Autres métriques de l'anonymat

- k-anonymity (Sweeney 2002) ;
- l-diversity (Machanavajjhala 2006) ;
- **Differential privacy** (Dwork 2006) ;
- t-closeness (Li 2007) ;
- (c, k)-Safety (Martin 2007) ;
- 3D-Privacy (Chen 2007) ;
- (d, γ)-Privacy (Rastogi 2007) ;
- ϵ-Privacy (Machanavajjhala 2009) ;
- Towards a unified theory of privacy and utility (Kifer 2010) ;
- ...

À retenir

- Une base de données anonyme ou anonymisée n'est **jamais vraiment anonyme** ;
- Le **recoupement** entre plusieurs bases de données peut permettre la **réidentification** (parfaite ou imparfaite, complète ou partielle) ;
- On peut être identifié par **n'importe quel type d'information**.

Besoin de nouvelles définitions ?

Qu'est-ce que l'anonymat ? Puisque ce n'est plus une grandeur binaire, comment doit-on le mesurer ?

Qu'est-ce qu'une donnée à caractère personnel ?

Qu'est-ce qu'une « information personnellement identifiante » ?

L'obligation de sécurité devrait-elle également être appliquée aux bases de données « anonymes » ?

Crédits iconographiques

Erik. S. Lesser, New York Times (`http://www.nytimes.com/2006/08/09/technology/09aol.html`)

Hay Kranen, *Mondrian lookalike.svg* (CC-BY 2.5, Wikimedia Commons)