Sécurité Appliquée
Protection de la vie privée-PVP

Jean-François Couchot

Université de Franche-Comté, UFR-ST

Plan

Big data et vie privée

Publication non sure de données

Un premier modèle de PVP : le k-anonymat

Plan

Plan

PVP : Exacerbé par le Big Data

Big Data & Data mining

- L'exploration de données (Data Mining) : inférence de connaissances intéressantes à partir de grandes quantités de données (Big Data)
- Tendances générale : exploration des données en croissance // Big Data
- Analyse/extraction de connaissance : techniquement réalisable aujourd'hui

Big Data : application et volume

- Domaines : veille économique, découverte scientifique, santé, profilage
- Marché du Big Data en santé : ≈ 67,82 G$ en 2025 (Globe News Wire)
- 90% de toutes les données : créées au cours des deux dernières années (IBM)
- 97,2% des organisations : investissent en Big Data et IA. (New Vantage)
- Offres d'emploi dans le domaine : ≈ 2,7M en 2020. (Forbes)

Plan

Vie Privée ? [3]

Historique

- Expression du “droit d’être laissé tranquille” (the right to be let alone) [1]
- “Nul ne sera l’objet d’immixtions arbitraires dans sa vie privée, sa famille, son domicile ou sa correspondance, [. . .]. Toute personne a droit à la protection de la loi contre de telles immixtions ou de telles atteintes.” [2]
- A l’heure d’Internet et des données (personnelles) transmises par : smartphones, messageries, GPS, appareils de fitness, moteur de recherche. . .

Des exemples d’inférences problématiques

- Réfrigérateur commandant des produits consommés : $\rightsquigarrow$ nbre. de présents/absents au domicile, risque sanitaire ? assurance ?
- Application de suivi de la santé : positions, fréquences card. partagées avec Apple/Google seulement ?

1. Warren, S. D., & Brandeis, L. D. (1890). The right to privacy. Harvard law review, 193-220.
2. Organisation des Nations Unies, (1949), Déclaration universelle des droits de l’homme.
3. https ://openclassrooms.com/fr/courses/5280946-protegez-les-donnees-personnelles

Scandales de non protection de la vie privée

National Security Agency (NSA) PRISM (2007-2013)

- En 2007, création par la NSA du programme PRISM de surveillance des communications échangées sur les services en ligne des GAFAM notamment
- Transmission des données brutes à des pays tiers
- Espionnage de leaders politiques internationaux (dont A.Merkel)

Cambridge Analytica (CA) (2014-2018)

- “This is your digital life” : application pour Facebook qui permet l’aspiration de données personnelles présentes du réseau
- 87M comptes utilisateurs Facebook aspirés dès 2014
- Ciblage politique pour convaincre de voter pour D. Trump en 2016
- ”Sans CA, il n’y aurait pas eu de Brexit” selon C. Wylie

Plan

Rôles des personnes accédant à la donnée - 1

Fournisseur de contenu

- Exemple d'une photo : les personnes sur la photo + photographe ev.

Autres entités connues du fournisseur avec lesquelles il souhaite la partager

- Exemple d'une photo : amis pour partager, fournisseur de service
- Moyens utilisés : application, mécanisme de groupes et de permissions
- Règlement Général sur la Protection des Données (RGPD) :
  - A. 6 : "Données minimales collectées et traitées de manière loyale et licite"
  - A. 12 : "Transparence des informations et des communications et modalités de l'exercice des droits de la personne concernée"
  - A. 16 : "Droit à la rectification"
  - A. 17 : "Droit à l'effacement, à l'oubli"
  - A. 20 : "Droit à la portabilité" : données récupérables (format lisible)

Rôles des personnes accédant à la donnée - 2

Gestionnaire de cette donnée

- Exemple d'une photo : employés de réseau social, d'entreprises de stockage dans le nuage ;
- Catégorie critique :
  - Modèle économique du service : extraction d'information
  - Administrateurs de BD : souvent honnête mais à risque
  - Chiffrement des données : souvent une solution

Reste du monde

- Exemple d'une photo : celles/ceux avec lequel on ne veut pas la partager

Plan

Plan

BD statistique [4]

Exemple de base d'étudiants

Nom	Sexe	Dpt.	Année	Test	Note
Allen	F	Info	00	600	3,4
Baker	F	Ing	00	520	2,5
Cook	H	Ing	98	630	3,5
Davis	F	Info	98	800	4,0
Evans	H	Bio	99	500	2,2
Frank	H	Ing	01	580	3,0
Good	H	Info	98	700	3,8
Hall	F	Psy	99	580	2,8
Ilies	H	Info	01	600	3,2
Jones	F	Bio	99	750	3,8
Kline	F	Psy	01	500	2,5
Lane	H	Ing	98	600	3,0
Moore	H	Info	99	650	3,5

- Identifiant : Nom
- Champs sensibles : Note, Test

Statistiques sensibles

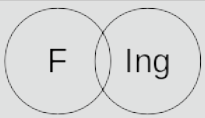
- Informations uniquement agrégées par `sum`, `count`,... : $|H.\text{Info}| = 3$
- Pas d'informations identifiantes : $F.\text{Info} = \{\text{Allen}, \text{Davis}\}$
- Stat. sensible $\equiv$ information sensible issue d'un singleton : $|F.\text{Ing}| = 1$, $\texttt{sum}(F.\text{Ing}, \text{Note}) = 2,5 \rightsquigarrow$ à interdire !

4. Denning, D. E. R. (1982). Cryptography and data security (Vol. 112). Reading : Addison-Wesley.

Requêtes statistiques dans $[n, N-n]$: pas sures

Tracker individuel, ordre 2

Nom	Sexe	Dpt.	Année	Test	Note
Baker	F	Ing	00	520	2,5
Cook	H	Ing	98	630	3,5
Frank	H	Ing	01	580	3,0
Lane	H	Ing	98	600	3,0



- $|F.\text{Ing}| = 1$: sensible !
- $|\text{Ing}| = 4$, $[\overline{F}.\text{Ing}| = 3$: OK
- $\texttt{sum}(F.\text{Ing}, \text{Note}) = \texttt{sum}(\text{Ing}, \text{Note}) - \texttt{sum}(\overline{F}.\text{Ing}, \text{Note}) = 12,0 - 9,5 = 2,5$

Tracker individuel, ordre 3

Nom	Sexe	Dpt.	Année	Test	Note
Evans	H	Bio	99	500	2,2
Hall	F	Psy	99	580	2,8
Jones	F	Bio	99	750	3,8
Moore	H	Info	99	650	3,5

99 \ H.Bio	00	01	11	10
0				
1			X	

- $q(H.\text{Bio}.99) + q((\overline{H} + \overline{\text{Bio}}).99) = q(99)$
- $q(H.\text{Bio}.99) = q(99) - q((\overline{H} + \overline{\text{Bio}}).99)$
- $\texttt{sum}(H.\text{Bio}.99, \text{Note}) = \texttt{sum}(99, \text{Note}) - \texttt{sum}((\overline{H} + \overline{\text{Bio}}).99, \text{Note}) = 12,3 - 10,1 = 2,2$

Inconvénient : nécessite de caractériser l'individu

Requêtes statistiques dans $[n, N-n]$: pas sures

Tracker général T : partager l'espace en deux

	T	$\overline{T}$
C	w	x
$\overline{C}$	y	z

- $q(\text{Tous}) = q(T) + q(\overline{T}) = w + x + y + z$
- Essayer $q(C) = q(C+T) + q(C+\overline{T}) - q(\text{Tous})$
- Sinon $q(C) = 2 \times q(\text{Tous}) - q(\overline{C}+T) - q(\overline{C}+\overline{T})$

Jones possède F.Bio, sans test d'unicité

- Essai avec le tracker $T = H$.
- $|\text{Tous}| = |H| + |\overline{H}| = 7 + 6 = 13$
- $|F.\text{Bio}| = |F.\text{Bio} + H| + |F.\text{Bio} + \overline{H}| - |\text{Tous}| = 1\,!$
- $sum(\text{Tous}, \text{Note}) = sum(H, \text{Note}) + sum(\overline{H}, \text{Note}) = 41,2$
- $sum(F.\text{Bio}, \text{Note}) =$
  $sum(F.\text{Bio} + H, \text{Note}) + sum(F.\text{Bio} + \overline{H}, \text{Note}) - sum(\text{Tous}, \text{Note}) = 3,8$

Présentation agrégée : pas toujours sure

Macrostatistiques

Effectifs par année/sexe

Sexe	98	99	00	01	Total
F	1	2	2	1	6
H	3	2	0	2	7
Total	4	4	2	3	13

Tests par année/sexe

Sexe	98	99	00	01	Total
F	800	1330	1120	500	3750
H	30	1150	0	1120	4260
Total	2730	2480	0	1180	8010

Suppressions

Tests par année/sexe

Sexe	98	99	00	01	Total
F	*	1330	1120	*	3750
H	1930	1150	0	1120	4260
Total	2730	2480	0	1180	8010

Tests par année/sexe

Sexe	98	99	00	01	Total
F	*	1330	1120	*	3750
H	*	1150	0	*	4260
Total	2730	2480	0	1180	8010

Pas de garantie de sûreté.

Arrondi systématique de valeurs : pas sûr

Arrondi systématique de q dans une base b

- Soit $b' = \left\lfloor \frac{b+1}{2} \right\rfloor$ et $d \equiv q \mod (b)$
- $$r(q) = \begin{cases} q & \text{si } d = 0 \\ q - d & \text{si } d < b' \\ q + (b - d) & \text{si } d \geq b' \end{cases}$$
- $r(q) \in [r(q) - b' + 1, r(q) + b' - 1]$

	$r(q_i)$	l_i	u_i
q_1	15	13	17
q_2	10	8	12
q_3	15	13	17
q_4	20	18	22
$\sum$		52	**68**
Q	70	$L = \mathbf{68}$	$U = 72$

Cas d'arrondis réversibles pour un attribut A

- Ensembles disjoints $C_1, \ldots, C_m$ et $q_i = sum(C_i, A)$, $1 \leq i \leq m$,
- $Q = q_1 + \cdots + q_m$ le total de toutes les sommes
- $[l_i, u_i]$ l'intervalle estimant $r(q_i)$, $[L, U]$ l'intervalle estimant $r(Q)$
- Si $L = \sum_i u_i \rightarrow \forall i, 1 \leq i \leq m, q_i = u_i$
- Si $U = \sum_i l_i \rightarrow \forall i, 1 \leq i \leq m, q_i = l_i$

Plan

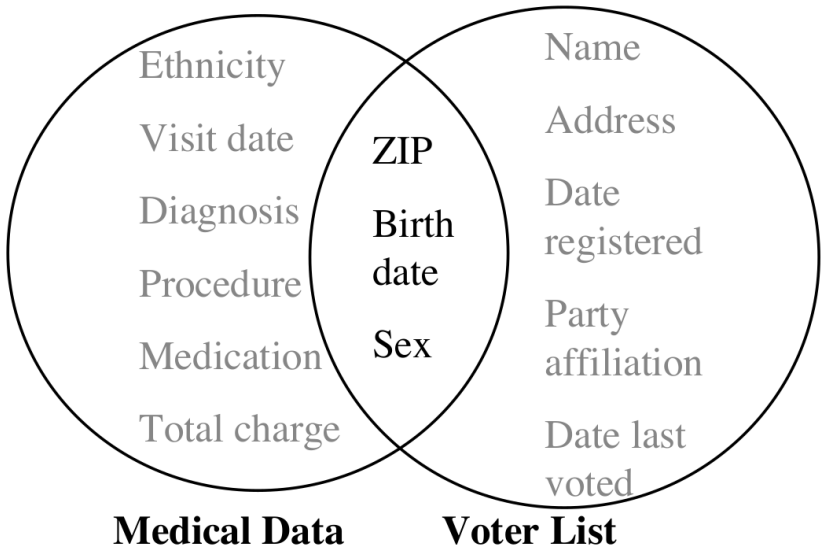
Pseudonymisation

- Champs identifiants : supprimés et remplacés par un id (H(NSS)).
- Avantage : calculs identiques à ceux sur la base de données initiale (Age moyen/cancer=37,8)

	Non-sensibles				Sensibles
H	CP	Age	Genre	Nationalité	Pathologie
1	13053	28	H	russe	trouble cardiaque
2	13068	29	H	américaine	trouble cardiaque
3	13068	21	F	japonaise	infection virale
4	13053	23	H	américaine	infection virale
5	14853	49	H	indienne	cancer
6	14853	48	F	russe	trouble cardiaque
7	14850	47	H	américaine	infection virale
8	14850	49	F	américaine	infection virale
9	13053	31	H	américaine	cancer
10	13053	37	H	indienne	cancer
11	13068	36	F	japonaise	cancer
12	13068	35	F	américaine	cancer

Pseudonymisation : attaque par QID, Sweeney [5]

- Base de donnée médicale pseudonymisée et publique



- Notion de Quasi IDentifiants : CP, date de naissance, genre
- Liste d'électeurs publique, recensement USA, 1990 : "87% of the population in the US had characteristics that likely made them unique based only on 5-digit Zip, gender, date of birth"
- Identification de données médicales du gouverneur Weld

5. Sweeney, L. (2002). k-anonymity : A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

Pseudonymisation : attaque par intersection [6]

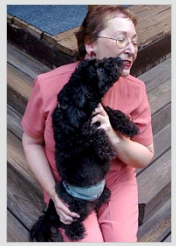
2006, diffusion par AOL de 20M requêtes, 658K utilisateur sans nom

AnonID	Query	QueryTime
1326	"holiday mansion houseboat"	2006-03-29
1326	"back to the future"	2006-04-01
591476	"english spanish translator"	2006-03-20
591476	"panama vacations"	2006-03-20
591476	"breast reduction"	2006-03-23
591476	"volunteer work at hospitals in brooklyn"	2006-05-24
591476	...	...
591476	"how to secretly poison your ex"	2006-03-12

Thelma Arnold, 62 ans, veuve vivant à Lilburn, Ga., réidentifiée en 3 j.

AnonID	Query
4417749	"people with last name 'Arnold"'
4417749	"landscapers in Lilburn, Ga"
4417749	"60 single men"
4417749	"dog that urinates on everything"
4417749	dog-related queries

$\rightsquigarrow$ Suppression hâtive des données sur le site d'AOL.



6. BARBARO, Michael, ZELLER, Tom, et HANSELL, Saul. A face is exposed for AOL searcher no. 4417749. New York Times, 2006, vol. 9, no 2008, p. 8.

Plan

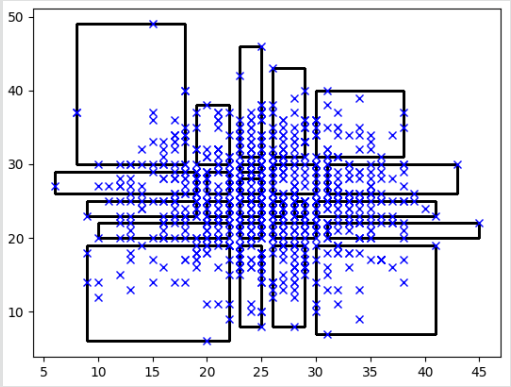
Big data et vie privée

Publication non sure de données

Un premier modèle de PVP : le k-anonymat
- Deux algorithmes pour l'atteindre
- Mésures d'utilité
- Attaques du k-anonymat

Le k-anonymat[5]

Intuition graphique : regrouper



Principe : classes d'équivalence de taille au moins k

- QID : attributs dont les combinaisons mènent à une identification
- Niveau de détail des valeurs des QID : à réduire pour qu'il y ait au moins k individus différents dont les QIDs sont égaux
- Individus avec mêmes QIDs : font partie de la même classe d'équivalence
- A résoudre :
  - quelle valeur pour k ?
  - quelle est la perte d'information ?

k-anonymat par généralisation

Réduction des niveaux de détail

- CP : suppr. chiffre D→G : XXXX*, XXX**,XX***, X****, suppr.
- Age : par intervalles d'ampl. croissante : 10, 20, 40, suppr.
- Genre : suppr.
- Nationalité : par continent, supprimer.
- $\rightsquigarrow 6 \times 5 \times 2 \times 3 = 180$ combinaisons de généralisation !

Regroupement par classes d'équivalence de card. $\geq$ à k

CP	Age	Genre	Nationalité	Pathologie
	[21; 31[	*	*	trouble cardiaque
{13053	[21; 31[	*	*	trouble cardiaque
13058}	[21; 31[	*	*	infection virale
	[21; 31[	*	*	infection virale
	[41; 50[	*	*	cancer
{14850	[41; 50[	*	*	trouble cardiaque
14853}	[41; 50[	*	*	infection virale
	[41; 50[	*	*	infection virale
	[31; 41[	*	*	cancer
{13053	[31; 41[	*	*	cancer
13058}	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer

4 individus

4 individus

4 individus

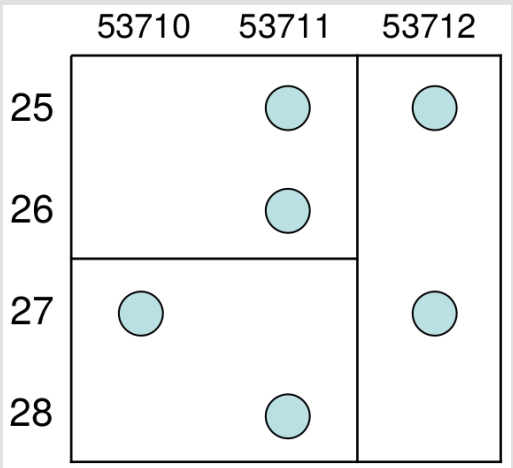
k-anonymat par Mondrian [7]

Algorithme général, en factorielle

1. Pour chaque séq. $[q_1, \ldots, q_f]$ de QID, tq'il reste plus de k éléments par groupe
   - partitionner intelligemment selon la médianne des données de q_i
2. généraliser les valeurs des attributs : un groupe $\equiv$ une généralisation
3. evaluer la perte d'information

2-anonymat, partitionnement selon [CP, Age, Genre]

Age	Genre	CP	Pathologie
25	H	53711	Grippe
25	F	53712	Hépatite
26	H	53711	Bronchite
27	H	53710	Bras cassé
27	F	53712	SIDA
28	H	53711	Ongle perdu



Age	Genre	CP	Pathologie
[25-26]	H	53711	Grippe
[25-27]	F	53712	Hépatite
[25-26]	H	53711	Bronchite
[27-28]	H	[53710-53711]	Bras cassé
[25-27]	F	53712	SIDA
[27-28]	H	[53710-53711]	Ongle perdu

7. LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006, April). Mondrian multidimensional k-anonymity. In 22nd International conference on data engineering (ICDE'06) (pp. 25-25). IEEE.

C_{AVG} : nbre. moy. d'élements par classe normalisé

Définition pour $|T|$ enregistrements et propriétés

$$C_{AVG} = \frac{|T|}{|EQs|} \times \frac{1}{k}$$

- et $|EQs|$: nbre. de classes d'équiv.
- $\frac{|T|}{|EQs|}$ nbre. moy. d'éléments par classe $(\geq k) \rightsquigarrow C_{AVG} \geq 1$
- Utilité décroissante à mesure que C_{AVG} croît.

Exemple avec 4-anonymat

CP	Age	Genre	Nationalité	Pathologie
	[21; 31[	*	*	trouble cardiaque
{13053	[21; 31[	*	*	trouble cardiaque
13058}	[21; 31[	*	*	infection virale
	[21; 31[	*	*	infection virale
	[41; 50[	*	*	cancer
{14850	[41; 50[	*	*	trouble cardiaque
14853}	[41; 50[	*	*	infection virale
	[41; 50[	*	*	infection virale
	[31; 41[	*	*	cancer
{13053	[31; 41[	*	*	cancer
13058}	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer

- $C_{AVG} = \frac{12}{3} \times \frac{1}{4} = 1$
- Optimal pour cette métrique

Loss : Perte dûe à la généralisation

Définition pour $|T|$ enregistrements, n QID et propriétés

$$Loss = \frac{1}{n|T|} \sum_{j=1}^{|T|} \sum_{i=1}^{n} \frac{R_{i_j}}{R_i}$$

- R_{i_j}/R_i : rapport acquis/acquérables
  - discret : $(|\text{généralisation } i_j| - 1)/(|\text{Qid } i| - 1)$
  - continu : $(U_{i_j} - L_{i_j})/(U_i - L_i)$
- Moyenne des acquis/acquérables, utilité décroissante // *Loss* croît

Exemple avec 4-anonymat

CP	Age	Genre	Nationalité	Pathologie
{13053 13058}	[21; 31[	*	*	trouble cardiaque
	[21; 31[	*	*	trouble cardiaque
	[21; 31[	*	*	infection virale
	[21; 31[	*	*	infection virale
{14850 14853}	[41; 50[	*	*	cancer
	[41; 50[	*	*	trouble cardiaque
	[41; 50[	*	*	infection virale
	[41; 50[	*	*	infection virale
{13053 13058}	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer

- Nat. $= \{\text{russe, am., jap., ind.}\} \rightsquigarrow$ $R_{\text{Nat}} = 3$, $R_{\text{CP}} = 3$, $R_{\text{Genre}} = 1$,
- $R_{\text{Age}} = 50 - 21 = 29$
- $Loss = \frac{1}{4\times 12}(8 \times (\frac{1}{3} + \frac{10}{29} + \frac{1}{1} + \frac{3}{3}) + 4 \times (\frac{1}{3} + \frac{9}{29} + \frac{1}{1} + \frac{3}{3})) = \frac{2}{3}$

Discernabilité

Définition pour $|T|$ enregistrements, et propriétés

$$Disc = \sum_{EQ, |EQ| \geq k} |EQ|^2 + \sum_{EQ, |EQ| < k} |T| \times |EQ|$$

- EQ, une classe d'équivalence
- Utilité décroissante à mesure que $Disc$ croît.

Exemple avec 4-anonymat

CP	Age	Genre	Nationalité	Pathologie
	[21; 31[	*	*	trouble cardiaque
{13053	[21; 31[	*	*	trouble cardiaque
13058}	[21; 31[	*	*	infection virale
	[21; 31[	*	*	infection virale
	[41; 50[	*	*	cancer
{14850	[41; 50[	*	*	trouble cardiaque
14853}	[41; 50[	*	*	infection virale
	[41; 50[	*	*	infection virale
	[31; 41[	*	*	cancer
{13053	[31; 41[	*	*	cancer
13058}	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer

- $Disc = 4^2 + 4^2 + 4^2 = 48$

Divergence de Kullback-Leibler

Mesure de dissimilarité entre deux distributions

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log_2 \left(\frac{P(x)}{Q(x)} \right).$$

- P (resp. Q) distrib. bassée sur les données réelles (resp. modifiées)
- Positive, utilité décroissante avec l'accroisement de D_{KL}

Exemple avec 4-anonymat

CP	Age	Genre	Nationalité	Pathologie
{13053 13058}	[21; 31[	*	*	trouble cardiaque
	[21; 31[	*	*	trouble cardiaque
	[21; 31[	*	*	infection virale
	[21; 31[	*	*	infection virale
{14850 14853}	[41; 50[	*	*	cancer
	[41; 50[	*	*	trouble cardiaque
	[41; 50[	*	*	infection virale
	[41; 50[	*	*	infection virale
{13053 13058}	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer

- $P(13053, 28, H, russe) = \frac{1}{12}$ et $Q(13053, 28, H, russe) = \frac{1}{3} \times \left(1 - (1 - \frac{1}{2 \times 10 \times 2 \times 4})^4\right)$
- $D_{\mathrm{KL}}(P \parallel Q) = 8\frac{1}{12} \log_2(\frac{\frac{1}{12}}{Q(13053,28,H,russe)}) + 4\frac{1}{12} \log_2(\frac{\frac{1}{12}}{Q(14850,47,H,am)}) \approx 3,28$

k-anonymat : attaques

Exemple

CP	Age	Genre	Nationalité	Pathologie
	[21; 31[	*	*	trouble cardiaque
{13053	[21; 31[	*	*	trouble cardiaque
13058}	[21; 31[	*	*	infection virale
	[21; 31[	*	*	infection virale
	[41; 50[	*	*	cancer
{14850	[41; 50[	*	*	trouble cardiaque
14853}	[41; 50[	*	*	infection virale
	[41; 50[	*	*	infection virale
	[31; 41[	*	*	cancer
{13053	[31; 41[	*	*	cancer
13058}	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer

4 individus

4 individus

4 individus

Attaques

- Homogénéité :
  - $\oplus$ Patient de 35 ans connu $\rightsquigarrow$ cancer.
  - $\ominus$ Patient de 29 ans connu $\rightsquigarrow$ ~~cancer~~.
- Connaissance supplémentaire : un japonais de 21 ans, P(trouble cardiaque|japonais)=faible $\rightsquigarrow$ infection virale.