



Bases de Données Avancées *k*-anonymat et extensions

Jean-François COUCHOT

Université de Franche-Comté, UFR-ST



Un premier modèle de PVP : le k -anonymat

Extensions du k -anonymat



Un premier modèle de PVP : le k -anonymat

Introduction au k -anonymat

Algorithme pour l'atteindre (parmi ...)

Mesures d'utilité

Attaques du k -anonymat

Extensions du k -anonymat



Un premier modèle de PVP : le k -anonymat

Introduction au k -anonymat

Algorithme pour l'atteindre (parmi ...)

Mesures d'utilité

Attaques du k -anonymat

Extensions du k -anonymat

Le k -anonymat³ et les quasi-identifiants -1

Quasi-identifiants : QID

- QID, intuition¹ : “des éléments qui ne sont pas en eux-mêmes des identificateurs uniques, mais qui sont suffisamment bien corrélés avec une entité pour pouvoir être combinés avec d'autres quasi-identifiants afin de créer un identificateur unique”
- QID, définition² : Les attributs de $Q \subseteq \{A_1, \dots, A_M\}$ sont quasi-identifiants de la relation T si la requête suivante retourne au moins un résultat

```
SELECT Q FROM T
GROUP BY Q
HAVING COUNT(*)=1
```

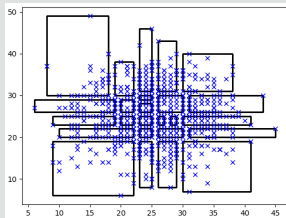
- (CP, genre, date de naissance) : triplets uniques dans 87% des cas \rightsquigarrow quasi-identifiants

1. <https://en.wikipedia.org/wiki/Quasi-identifiant>

2. Nguyen, B., & Castelluccia, C. (2020). Techniques d'anonymisation tabulaire : concepts et mise en oeuvre. arXiv preprint arXiv :2001.02650.

Intuition : regrouper les QID pour casser l'unicité

- 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



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



Un premier modèle de PVP : le k -anonymat

Introduction au k -anonymat

Algorithme pour l'atteindre (parmi ...)

Mesures d'utilité

Attaques du k -anonymat

Extensions du k -anonymat

Rappel de l'exemple pseudonymisé



H	Non-sensibles				Sensibles
	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

k-anonymat par généralisation



Réduction des niveaux de détail

Sur l'exemple :

- CP : laisser, **regroupements par 2**, **suppr.**
- Age : laisser, par intervalles d'amplitudes 10, 20, **suppr.**
- Genre : laisser, **suppr.**
- Nationalité : laisser, par continent, **suppr.**
- $\rightsquigarrow 3 \times 4 \times 2 \times 3 = 72$ combinaisons de généralisation !

Regroupement par classes d'équivalence de card. \geq à 4

CP	Age	Genre	Nationalité	Pathologie	
{13053 13068}	[21; 31[*	*	trouble cardiaque	} 4 individus
	[21; 31[*	*	trouble cardiaque	
	[21; 31[*	*	infection virale	
	[21; 31[*	*	infection virale	
{14850 14853}	[41; 50[*	*	cancer	} 4 individus
	[41; 50[*	*	trouble cardiaque	
	[41; 50[*	*	infection virale	
	[41; 50[*	*	infection virale	
{13053 13068}	[31; 41[*	*	cancer	} 4 individus
	[31; 41[*	*	cancer	
	[31; 41[*	*	cancer	
	[31; 41[*	*	cancer	



Un premier modèle de PVP : le k -anonymat

Introduction au k -anonymat

Algorithme pour l'atteindre (parmi ...)

Mesures d'utilité

Attaques du k -anonymat

Extensions du k -anonymat

C_{AVG} : nbre. moy. d'éléments 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
{13053 13068}	[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 13068}	[31; 41[*	*	cancer
	[31; 41[*	*	cancer
	[31; 41[*	*	cancer
	[31; 41[*	*	cancer

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

Loss : Perte due à la généralisation



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

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

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

Exemple avec 4-anonymat

CP	Age	Genre	Nationalité	Pathologie
{13053 13068}	[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 13068}	[31; 41[*	*	cancer
	[31; 41[*	*	cancer
	[31; 41[*	*	cancer
	[31; 41[*	*	cancer

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



Un premier modèle de PVP : le k -anonymat

Introduction au k -anonymat

Algorithme pour l'atteindre (parmi ...)

Mesures d'utilité

Attaques du k -anonymat

Extensions du k -anonymat



Exemple

CP	Age	Genre	Nationalité	Pathologie	
{13053 13068}	[21; 31[*	*	trouble cardiaque	} 4 individus
	[21; 31[*	*	trouble cardiaque	
	[21; 31[*	*	infection virale	
	[21; 31[*	*	infection virale	
{14850 14853}	[41; 50[*	*	cancer	} 4 individus
	[41; 50[*	*	trouble cardiaque	
	[41; 50[*	*	infection virale	
	[41; 50[*	*	infection virale	
{13053 13068}	[31; 41[*	*	cancer	} 4 individus
	[31; 41[*	*	cancer	
	[31; 41[*	*	cancer	
	[31; 41[*	*	cancer	

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, $P(\text{trouble cardiaque}|\text{japonais})=\text{faible} \rightsquigarrow$ infection virale.



Un premier modèle de PVP : le k -anonymat

Extensions du k -anonymat

Casser l'homogénéité par l -diversité

Préserver les distributions sensibles par t -proximité



Un premier modèle de PVP : le k -anonymat

Extensions du k -anonymat

Casser l'homogénéité par l -diversité

Préserver les distributions sensibles par t -proximité

/-diversité pour contrer l'homogénéité



Définition et remarque

- Classe d'équivalence : l -diverse si contient au moins l valeurs "représentatives" par **donnée sensible**
- Base de données : l -diverse si toutes ses classes d'équivalence le sont
- Représentatives : valeurs distinctes \subset répartition sensée...

4-anonymat et 3-diversité distinctes

H	Quasi-Identifiants				Sensibles
	CP	Age	Genre	Nationalité	Pathologie
1	130**	[21; 41[*	*	trouble cardiaque
2	130**	[21; 41[*	*	trouble cardiaque
3	130**	[21; 41[*	*	infection virale
4	130**	[21; 41[*	*	infection virale
9	130**	[21; 41[*	*	cancer
10	130**	[21; 41[*	*	cancer
11	130**	[21; 41[*	*	cancer
12	130**	[31; 41[*	*	cancer
5	148**	[41; 50[*	*	cancer
6	148**	[41; 50[*	*	trouble cardiaque
7	148**	[41; 50[*	*	infection virale
8	148**	[41; 50[*	*	infection virale

} 3 val. sensibles \neq

} 3 val. sensibles \neq

(c, l) -diversité récursive



Définition pour chaque classe d'équivalence EQ

$$\forall \text{EQ} \frac{r_1}{r_l + r_{l+1} + \dots + r_m} < c$$

- r_1, r_2, \dots, r_m : fréq. décroissantes des valeurs des attributs sensibles
- c : facteur (réel, >1) de tolérance de variations des extrêmes

4-anonymat et $(2.01, 3)$ -diversité récursive

H	Quasi-Identifiants				Sensibles
	CP	Age	Genre	Nationalité	Pathologie
1	130**	[21; 41[*	*	trouble cardiaque
2	130**	[21; 41[*	*	trouble cardiaque
3	130**	[21; 41[*	*	infection virale
4	130**	[21; 41[*	*	infection virale
9	130**	[21; 41[*	*	cancer
10	130**	[21; 41[*	*	cancer
11	130**	[21; 41[*	*	cancer
12	130**	[31; 41[*	*	cancer
5	148**	[41; 50[*	*	cancer
6	148**	[41; 50[*	*	trouble cardiaque
7	148**	[41; 50[*	*	infection virale
8	148**	[41; 50[*	*	infection virale

- Dans $\text{EQ}_{130^{**}}$ et $\text{EQ}_{148^{**}}$: $r_1 = 0.5$,
 $r_2 = r_3 = 0.25$
- $\frac{r_1}{r_3} = 2 < 2.01$



Exemple de sur-représentation d'une valeur⁴

- Données originales 10000 personnes :
 - Un seul attribut sensible : résultat de test viral.
 - Deux valeurs : positive (1%) et négative (99%).
- Valeurs avec des degrés de sensibilité très \neq :
 - Peu d'opposition que l'on sache que le test est négatif (comme 99 % de la population)
 - Forte réticence à être connu positif
- 2-diversité :
 - Au maximum $10000 \times 1\% = 100$ classes d'équivalence
 - \rightsquigarrow perte d'informations importante.

4. Li, N., Li, T., & Venkatasubramanian, S. (2009). Closeness : A new privacy measure for data publishing. IEEE Transactions on Knowledge and Data Engineering, 22(7), 943-956.

Exemple précédent avec une classe d'équivalence équilibrée

- Satisfait la 2-diversité :
 - distincte
 - $(c, 2)$ récursive, pour $c > 1$.
- Toute personne de cette classe : considérée 1 fois/2 positive

Exemple précédent avec une classe d'équivalence en 49/1

- 49 positifs / 1 négatif : satisfait la 2-diversité :
 - distincte
- Mais une personne de cette classe : considérée positive à 98% (vs. 1%).
- Mais cette classe a exactement la même diversité qu'une classe avec 1 positif / 49 négatifs

Avec des éléments \neq mais sémantiqu^t proches ?

Exemple agrégé avec salaire et maladies sensibles

CP	Age	Salaire	Pathologie
476**	2*	3K	ulcère gastrique
476**	2*	4K	gastrite
476**	2*	5K	cancer de l'estomac
4790*	> 40	6K	gastrite
4790*	> 40	11K	grippe
4790*	> 40	8K	bronchite
476**	3*	7K	bronchite
476**	3*	9K	pneumonie
476**	3*	10K	cancer de l'estomac

- Satisfait
 - la 3-diversité distincte.

Fuite due à la non prise en compte de la proximité de valeurs.

Déductions possibles de la connaissance que Bob est dans la classe 1 :

- son salaire ([3K-5K]) est relativement bas
- souffre de l'estomac (toutes les pathologies y sont liées)



Un premier modèle de PVP : le k -anonymat

Extensions du k -anonymat

Casser l'homogénéité par l -diversité

Préserver les distributions sensibles par t -proximité



Définition : vérification de la t – proximité

- dans chaque classe EQ si pour chaque attribut sensible, la *distance* entre sa distribution dans EQ et sa distribution dans la table complète est $\leq t$
- dans la base complète si toutes ses classes d'équivalence la respectent
- attribut sensible \equiv pas de généralisation

EMD entre les distributions $P = (p_1, p_2, \dots, p_n)$ et $Q = (q_1, q_2, \dots, q_n)$

- Travail minimal à fournir pour modifier un tas de terre en un autre
- Pour un attribut numérique : $v_1 < v_2 < \dots < v_m$

$$D(P, Q) = |p_1 - q_1| + |p_2 - q_2 + p_1 - q_1| + \dots + |p_m - q_m + \dots + p_1 - q_1|$$

- Pour un attribut catégoriel : $\{v_1, v_2, \dots, v_m\}$ par distance au sol :

$$D(P, Q) = \frac{1}{2} \sum_{i=1}^m |p_i - q_i|$$

t-proximité sur un exemple



Données originales et généralisées

QID				Sensibles
CP	Age	Genre	Nationalité	Pathologie
13053	28	H	russe	trouble cardiaque
13068	29	H	américaine	trouble cardiaque
13068	21	F	japonaise	infection virale
13053	23	H	américaine	infection virale
14853	49	H	indienne	cancer
14853	48	F	russe	trouble cardiaque
14850	47	H	américaine	infection virale
14850	49	F	américaine	infection virale
13053	31	H	américaine	cancer
13053	37	H	indienne	cancer
13068	36	F	japonaise	cancer
13068	35	F	américaine	cancer

$$q_{tc} = \frac{1}{4} \quad q_{iv} = \frac{1}{3} \quad q_c = \frac{5}{12}$$

QID				Sensibles
CP	Age	Genre	Nationalité	Pathologie
*	*	F	*	infection virale
*	*	F	*	trouble cardiaque
*	*	F	*	infection virale
*	*	F	*	cancer
*	*	F	*	cancer
*	*	H	*	trouble cardiaque
*	*	H	*	trouble cardiaque
*	*	H	*	infection virale
*	*	H	*	cancer
*	*	H	*	infection virale
*	*	H	*	cancer
*	*	H	*	cancer

$$p_{tc}^F = \frac{1}{5}, \quad p_{iv}^F = \frac{2}{5}, \quad p_c^F = \frac{2}{5} \quad \text{et} \quad p_{tc}^H = \frac{2}{7}, \quad p_{iv}^H = \frac{2}{7}, \quad p_c^H = \frac{3}{7}$$

Distances entre les distributions P^F et Q puis P^H et Q

- $D(P^F, Q) = \frac{1}{2} \left(\left| \frac{1}{5} - \frac{1}{4} \right| + \left| \frac{2}{5} - \frac{1}{3} \right| + \left| \frac{2}{5} - \frac{5}{12} \right| \right) = \frac{1}{15} \approx 0.06666666$
- $D(P^H, Q) = \frac{1}{2} \left(\left| \frac{2}{7} - \frac{1}{4} \right| + \left| \frac{2}{7} - \frac{1}{3} \right| + \left| \frac{3}{7} - \frac{5}{12} \right| \right) = \frac{1}{21} \approx 0.0476190$
- Données généralisées $\frac{1}{15}$ -proches des originales

Mesure du gain d'information moyen $\mathcal{A}_{\text{know}}$

Définition⁵ pour chaque classe EQ

$$\mathcal{A}_{\text{know}} = \frac{1}{|T|} \sum_{eq \in EQs} |eq| d(P^{eq}, Q)$$

- Un tuple de quasi-identifiants \rightsquigarrow classe $eq \rightsquigarrow$ quantité moyenne d'informations apprises sur les attributs sensibles

$\mathcal{A}_{\text{know}}$ sur l'exemple

QID				Sensibles
CP	Age	Genre	Nationalité	Pathologie
*	*	F	*	infection virale
*	*	F	*	trouble cardiaque
*	*	F	*	infection virale
*	*	F	*	cancer
*	*	F	*	cancer
*	*	H	*	trouble cardiaque
*	*	H	*	trouble cardiaque
*	*	H	*	infection virale
*	*	H	*	cancer
*	*	H	*	infection virale
*	*	H	*	cancer
*	*	H	*	cancer

- $\mathcal{A}_{\text{know}} = \frac{1}{12} \left(5 \frac{1}{15} + 7 \frac{1}{21} \right) = \frac{1}{18} \approx 5.5\%$
- Femme ds la base connue et publication \rightsquigarrow chance d'avoir une IV augmentée

5. Brickell, J., & Shmatikov, V. (2008, August). The cost of privacy : destruction of data-mining utility in anonymized data publishing. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 70-78).