



Sécurité Appliquée Protection de la vie privée-PVP

Jean-François COUCHOT

Université de Franche-Comté, UFR-ST



Extensions du k -anonymat

Ajouter du bruit aux résultats ?



Extensions du k -anonymat

Casser l'homogénéité par l -diversité

Préserver les distributions sensibles par t -proximité

Ajouter du bruit aux résultats ?

Inspiré entre autre de¹

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



Extensions du k -anonymat

Casser l'homogénéité par l -diversité

Préserver les distributions sensibles par t -proximité

Ajouter du bruit aux résultats ?

Rappels de l'exemple



Les données

H	Non-sensibles				Sensibles
	CP	Age	Genre	Nationality	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

Généralisation selon une hiérarchie

- CP : laisser, **regroupements par 2**, **suppr.**
- Age : laisser, par intervalles d'amplitudes 10, 20, **suppr.**
- Genre : laisser, **suppr.**
- Nationalité : laisser, par continent, **suppr.**

Homogénéité problématique



4-anonymat minimisant *Loss*

CP	Age	Genre	Nationalité	Pathologie	
{13053 13058}	[20; 30[*	*	trouble cardiaque	} 4 individus
	[20; 30[*	*	trouble cardiaque	
	[20; 30[*	*	infection virale	
	[20; 30[*	*	infection virale	
{14850 14853}	[40; 50[*	*	cancer	} 4 individus
	[40; 50[*	*	trouble cardiaque	
	[40; 50[*	*	infection virale	
	[40; 50[*	*	infection virale	
{13053 13058}	[30; 40[*	*	cancer	} 4 individus
	[30; 40[*	*	cancer	
	[30; 40[*	*	cancer	
	[30; 40[*	*	cancer	

/-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

Quasi-Identifiants				Sensibles
CP	Age	Genre	Nationalité	Pathologie
{13053, 13058}	[21; 41[*	*	trouble cardiaque
	[21; 41[*	*	trouble cardiaque
	[21; 41[*	*	infection virale
	[21; 41[*	*	infection virale
	[21; 41[*	*	cancer
	[21; 41[*	*	cancer
	[21; 41[*	*	cancer
{14850 14853}	[41; 50[*	*	cancer
	[41; 50[*	*	trouble cardiaque
	[41; 50[*	*	infection virale
	[41; 50[*	*	infection virale

} 3 val. sensibles \neq

} 3 val. sensibles \neq



Exemple de sur-représentation d'une valeur²

- 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.

2. 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.

/-diversité : utilité statistique ?



Exemple précédent avec une classe d'équivalence en 50/50

- Satisfait la 2-diversité distinctes
- 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)



Extensions du k -anonymat

Casser l'homogénéité par l -diversité

Préserver les distributions sensibles par t -proximité

Ajouter du bruit aux résultats ?



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



Extensions du k -anonymat

Ajouter du bruit aux résultats ?

Notations et motivation

Quelle quantité α de bruit ajouter ?



Extensions du k -anonymat

Ajouter du bruit aux résultats ?

Notations et motivation

Quelle quantité α de bruit ajouter ?

Des données, des histogrammes

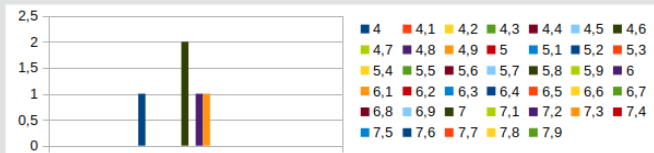


Représentation

- \mathcal{X} : l'univers des valeurs possibles
- Base D : multi-ensemble d'éléments de \mathcal{X}
- Représentée par un histogramme : $D \in \mathbb{N}^{|\mathcal{X}|}$

Exemple sur une base de tailles en pieds

$D = [5'2, 6'1, 5'8, 5'8, 6'0]$ et $\mathcal{X} = \{4'0, 4'1, \dots, 7'9\}$ par ex.



Requête de comptage sur $D = \{\dots, v_i : e_i, \dots\}$

Définition d'une requête de comptage Q_S

- $S \subseteq \mathcal{X}$: valeurs de \mathcal{X} à compter
- $Q_S(D) = \sum_{v_i \in S} e_i$

$$D_1 = \{5'2 : 1, 5'8 : 2, 6'0 : 1, 6'1 : 1\} \quad D_2 = \{5'2 : 2, 5'3 : 1, 5'8 : 1, 6'0 : 1, 6'1 : 1\}$$

- $S = \{5'2, 5'3, 5'8\}$
- $Q_S(D_1) = 1 + 0 + 2 = 3$ et $Q_S(D_2) = 2 + 1 + 1 = 4$

Comme un produit de vecteurs

- $D_1 = (\dots, 1, 0, \dots, 2, 0, 1, 1, \dots)$, $D_2 = (\dots, 2, 1, \dots, 1, 0, 1, 1, \dots)$
- $S = (\dots, 1, 1, \dots, 1, 0, 0, 0, \dots)$
- $Q_S(D_1) = S \cdot D_1^T = 3$, $Q_S(D_2) = S \cdot D_2^T = 4$

Distance entre bases de données



Norme 1 pour un vecteur $x = (x_1, \dots, x_k)$ de \mathbb{R}^k

La norme 1 de x est :

$$\|x\|_1 = |x_1| + \dots + |x_k|$$

Norme 1 entre deux bases D_1 et D_2

- Exprimer D_1 et D_2 comme deux vecteurs
- Norme-1 du vecteur de différences entre les deux $\|D_1 - D_2\|_1$

$$D_1 = \{5'2 : 1, 5'8 : 2, 6'0 : 1, 6'1 : 1\} \quad D_2 = \{5'2 : 2, 5'3 : 1, 5'8 : 1, 6'0 : 1, 6'1 : 1\}$$

- $D_1 = (\dots, 1, 0, \dots, 2, 0, 1, 1, \dots) \rightsquigarrow \|D_1\|_1 = |1| + |2| + |1| + |1| = 5$
- $D_2 = (\dots, 2, 1, \dots, 1, 0, 1, 1, \dots) \rightsquigarrow \|D_2\|_1 = |2| + |1| + |1| + |1| + |1| = 6$
- $\|D_1 - D_2\|_1 = \|(\dots, 1, 1, \dots, -1, 0, 0, 0, \dots)\|_1 = |1| + |1| + |-1| = 3$

Borner le bruit du comptage



Motivation

- Peut-on trouver un algorithme respectueux qui retourne un comptage proche de l'original ?

Définition du bruit borné par α^3

Un algorithme Q' ajoute un bruit borné par α si pour toute base D et toute requête S on a

$$|Q_S(D) - Q'_S(D)| \leq \alpha$$

Exemple de comptage bruité

$Q'_S(D) = Q_S(D) + b$ avec un bruit b t.q.

- $b \sim \mathcal{U}[-\alpha, \alpha]$
- $b \sim \mathcal{N}(0, 3\alpha)$ et $\text{abs}(b) < \alpha$
- ...

3. Dinur, I., & Nissim, K. (2003, June). Revealing information while preserving privacy. In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 202-210).



Extensions du k -anonymat

Ajouter du bruit aux résultats ?

Notations et motivation

Quelle quantité α de bruit ajouter ?

Borner par α : D est \approx reconstructible !



Théorème : attaque d'un algorithme Q' ajoutant un bruit borné par α

- si $Q'_S(D)$ connu pour tout S : on peut construire D' tq $\|D - D'\|_1 \leq 4\alpha$
- En pratique, D' est encore plus proche de D

Preuve

- Construction d'une solution D' :
 - par brute force sur l'ensemble des possibilités sur \mathcal{X}
 - et qui vérifie $|Q_S(D') - Q'_S(D)| < \alpha$ pour chaque requête (finie) S
- $S_0 = D' \setminus D$ et $S_1 = D \setminus D'$
- Si $\|D - D'\|_1 = \|S_0\|_1 + \|S_1\|_1 > 4\alpha$
 - $\max(\|S_0\|_1, \|S_1\|_1) > 2\alpha \rightsquigarrow$ on suppose $\|S_0\|_1 > 2\alpha$
 - or $Q_{S_0}(D) = 0 \rightsquigarrow Q'_{S_0}(D) \leq \alpha$ (ajout de bruit par Q' borné par α)
 - or $Q_{S_0}(D') > 2\alpha \rightsquigarrow |Q_{S_0}(D') - Q'_{S_0}(D)| > |2\alpha - \alpha| = \alpha \frac{1}{2}$

Mauvaise nouvelle ? A nuancer



Non démontré

- $\alpha < \frac{n}{40}$: bruit ajouté faible mais 90% des données sont restructuribles
- $\alpha = \frac{n}{2}$: données protégées mais utilité ?

construction de D' pour $\alpha = \mathcal{O}(n)$?

1. Générer une distribution sur \mathcal{X}
2. Générer toutes les requêtes S et tester, reprendre en 1 ev.
3. Conclusion : en $\mathcal{O}(\exp(n))$

Et pour un bruit plus petit ?

- Si bruit $\alpha = \mathcal{O}(\sqrt{n})$: attaquable en $\mathcal{O}(n \log(n))$