Bases de Données avancés - TD1.
Des statistiques au k-anonymat

Jean-François COUCHOT
couchot[arobase]femto-st[point]fr

2 mars 2021

1 Bases de données statistiques

Exercice 1.1 (Arrondi systématique). *Deux équipes constituées d'une femme et d'un homme chacune ont participé à un test dont les résultats individuels sont des entiers qui doivent rester secrets. L'organisation a publié des arrondis au multiple de 5 le plus proche des résultats individuels et des sommes dans le tableau ci-dessous.*

	H	F	Σ
E_1	10	10	25
E_2	10	10	25
Σ	25	25	50

Montrer qu'on peut retrouver les résultats de chacun des concurrents et les trouver.

Exercice 1.2 (Trackers). *Dans cette partie, on reprend les données de la table 1, où les deux colonnes rouges sont des données sensibles. Dans cette base de données statistiques, ne sont autorisées que les requêtes de type* `count` *et* `sum` *sur des ensembles dont la cardinalité est entre 3 et 10. requêtes*

1. *En utilisant un tracker individuel, montrer qu'on peut récupérer la note de Good.*
2. *En utilisant un tracker général, montrer qu'on peut récupérer la note de Moore.*

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

TABLE 1 – Données brutes avec test et note pour chaque étudiant

2 Pseudonymisation

Exercice 2.1 (Pseudonymization). *Cet exercice est inspiré de* [1]*. L'ensemble de données d'un réseau social donné dans le tableau 2 a été pseudonymisé de la manière la plus forte possible, i.e. où le nom a été supprimé. Cependant,*

1. http://www.infsec.cs.uni-saarland.de/teaching/16WS/Cybersecurity/exercises/exercise-11.pdf

des informations supplémentaires sont à votre disposition dans le tableau 3. Dans le cadre de cet exercice, nous utilisons ces informations pour étudier comment la vie privée peut être mise à mal en reliant intelligemment les données. Supposons que pour tous les candidats sont présents dans les deux bases de données.

Name	Gender	Age	City of birth	Favorite TV Series	Relationship Status
*	male	19-25	Saarbrücken	Game of Thrones	single
*	female	16-18	Trier	Game of Thrones	in relationship
*	male	12-15	München	Friends!	in relationship
*	female	19-25	Berlin	Big Bang Theory	in relationship
*	female	19-25	Hamburg	Big Bang Theory	single
*	female	19-25	Saarbrücken	Game of Thrones	single
*	male	16-18	Trier	Game of Thrones	single
*	female	12-15	München	Game of Thrones	in relationship
*	male	19-25	Berlin	Big Bang Theory	single

TABLE 2 – Jeu de données d'un réseau social, nettoyées par pseudonymisation

Name	Email	TV Show	Rating (1=bad, 5=great)
Alice	alice1995@email.com	Friends!	1
Bob	bobbybob@email.com	Friends!	4
Charlie	s9charchar@email.com	Friends!	2
Eve	evelyn@myhighscool.com	Friends!	1
Bob	bobbybob@email.com	Game of Thrones	1
Alice	alice1995@email.com	Game of Thrones	5
Charlie	s9charchar@email.com	Game of Thrones	5
Bob	bobbybob@email.com	Big Bang Theory	3
Charlie	s9charchar@email.com	Big Bang Theory	5
Alice	alice1995@email.com	Big Bang Theory	2
Eve	evelyn@myhighscool.com	Big Bang Theory	5

TABLE 3 – Informations additionnelles

1. *Où Alice est-elle probablement née et quel est son état matrimonial le plus probable ?*
2. *Pouvez-vous également obtenir des informations personnelles sur Charlie ?*
3. *Pouvez-vous aussi apprendre des informations personnelles sur Bob ?*

3 k-anonymat

Exercice 3.1 (k-anonymat sur un exemple simple). *Cet exercice est encore inspiré de* [1].

1. *Le jeu de donné 1 de la Figure 1 satisfait-il le k-anonymat ? Si oui, quelle est la valeur maximale de k ?*
2. *Même question pour les jeux 2 et 3.*

	ID	QID			sensible
#	Nom	stat. conjugal	Age	CP	Crime
1	Joe	Séparé	29	32042	Meurtre
2	Jill	Célibataire	20	32021	Vol
3	Sue	Veuve	24	32024	Trafic
4	Abe	Séparé	28	32046	Agression
5	Bob	Veuf	25	32045	Piratage
6	Amy	Célibataire	23	32027	Indécence

TABLE 4 – Données de criminalité

ID	Age	Gender	Fav.Show
1	12-15	female	Friends!
2	19-25	male	Friends!
3	19-25	male	Friends!
4	12-15	female	Friends!
5	19-25	male	G.o.T.
6	19-25	male	G.o.T.
7	19-25	male	G.o.T.

ID	Age	Gender	Fav.Show
1	19-25	female	Grey's A.
2	19-25	female	Simpsons
3	19-25	female	Futurama
4	19-25	female	Friends!
5	19-25	male	G.o.T.
6	19-25	male	C.Minds
7	19-25	male	Br.Ba.

ID	Age	Gender	Fav.Show
1	19	male	Friends!
2	19	male	Friends!
3	19	male	Friends!
4	19	female	Friends!
5	20	male	G.o.T.
6	20	male	G.o.T.
7	20	male	G.o.T.

FIGURE 1 – 3 small generalized datasets

Exercice 3.2 (Différentes méthodes de k-anonymat sur un même micro exemple). *On considère le jeu de données D de la table 4, inspiré de* [2].

1. *Proposer une version 2-anonyme D^g_{k2} en considérant les généralisations suivantes :*
   - *Statut marital : séparé(e), célibataire, veuf/veuve, célibataire $\rightsquigarrow$ **
   - *Age : 20,23,24$\rightsquigarrow$[20,24], 25,28,29 $\rightsquigarrow$ [25,29], $\rightsquigarrow$ **
   - *CP :32021,32024,32027$\rightsquigarrow$ 3202*, 32042,32045,32046$\rightsquigarrow$ 3204*, $\rightsquigarrow$ **
2. *Proposer une version 2-anonyme D^m_{k2} en considérant l'algorithme de Mondrian.*
3. *Calculer C_{AVG} et de Loss sur D^m_{k2} et D^g_{k2} et émettre une conjecture.*
4. *La figure 2 compare l'utilité de différentes méthodes de k-anonymisation sur deux jeux de données différents (Adult et Irish). Que peut-on conclure de ces expérimentations ?*

2. , V., McDonagh, P., Cerqueus, T., & Murphy, L. (2014). A systematic comparison and evaluation of k-anonymization algorithms for practitioners. Transactions on data privacy, 7(3), 337-370.

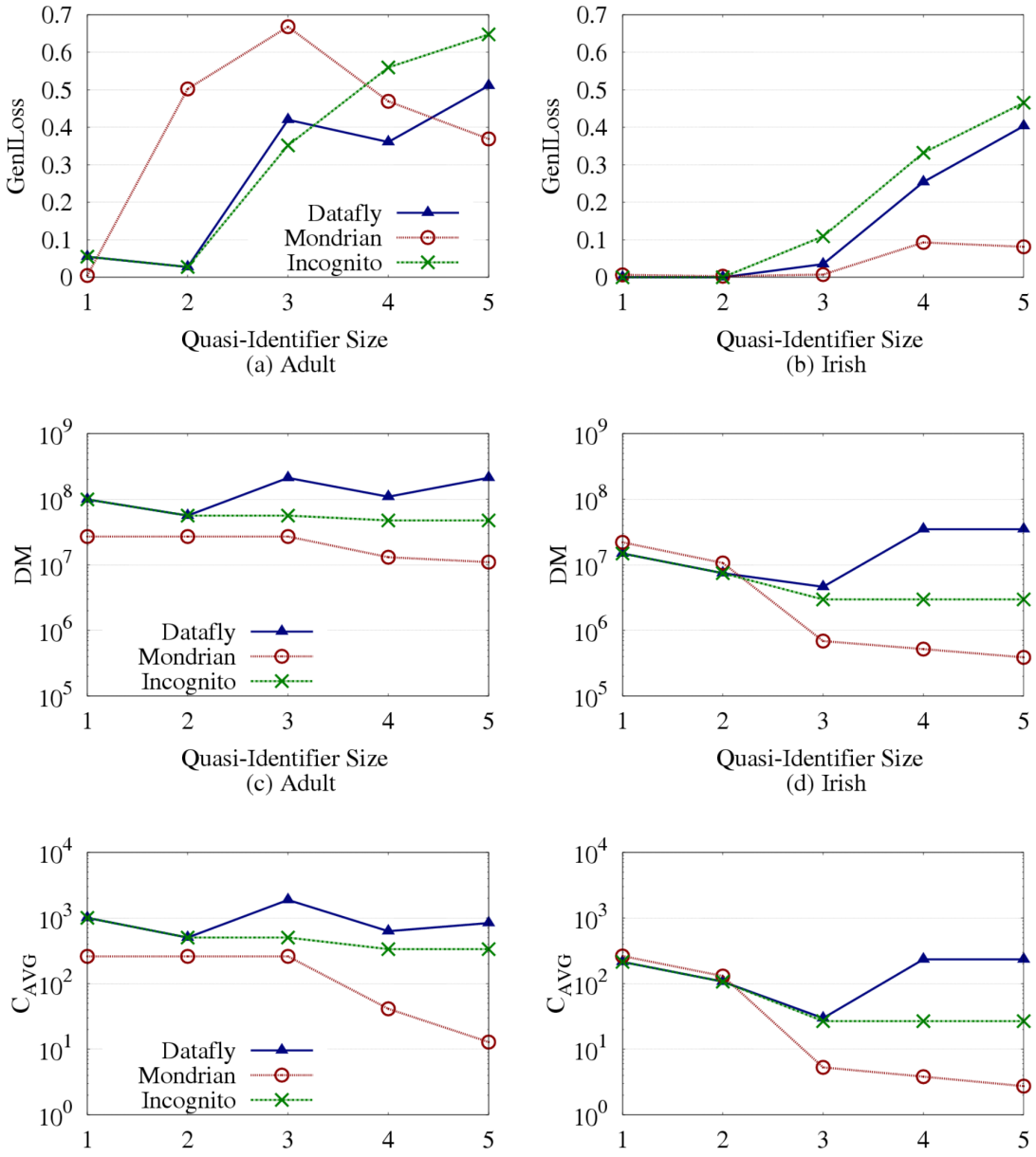

FIGURE 2 – Mesures d'utilité d'algorithmes de k-anonymat[2]