Nom: Prénom:

L3 informatique. Bases de données avancées.
Partie J.-F. COUCHOT.

Seule une feuille A4 de notes manuscrites personnelle est autorisée. Tous les moyens de communication sont interdits. Toutes les réponses doivent être justifiées. Sans justification, une réponse est considérée comme fausse.

1 Attaque d'une publication de données agrégées

Une entreprise publie tous les jours des statistiques de fréquentation d'une zone géographique donnée sous la forme :

- le premier jour j_1 : nombre de personnes uniques présentes sur cette zone ;
- le second jour j_2 :
  - nombre de personnes uniques présentes sur cette zone le jour j_2 ;
  - nombre de personnes uniques présentes sur cette zone les jours j_2 ou j_1 ;
- le troisième jour j_3 :
  - nombre de personnes uniques présentes sur cette zone le jour j_3 ;
  - nombre de personnes uniques présentes sur cette zone les jours j_3 ou j_2 ;
  - nombre de personnes uniques présentes sur cette zone les jours j_3 ou j_2 ou j_1 ;
- ...

On va montrer que dans certains cas, cette publication permet d'avoir des informations beaucoup plus précises et donc potentiellement problématiques.

L'étude commence un lundi et les données suivantes pour lundi et mardi sont publiées :

- lundi soir, publication de : "770 personnes sont venues lundi" ;
- mardi soir, publication de : "780 personnes sont venues mardi et 1530 personnes sont venues mardi ou lundi".

1. Montrer qu'on peutdéduire (et le faire) toutes les informations suivantes. Attention, il n'y a pas d'ordre dans les questions suivantes.
   - le nombre a de personnes présentes lundi mais pas mardi ;
   - le nombre b de personnes présentes à la fois lundi et mardi ;
   - le nombre c de personnes présentes mardi mais pas lundi.

2. Vous connaissez un groupe de musiciennes et musiciens de 40 personnes se déplaçant toujours ensemble. Vous savez que le groupe était présent sur la zone. Quelle information pouvez vous déduire de la question précédente ?

2 Pseudonymisation : mise en place et attaque

La relation T1 donnée à la figure 1 est une extraction brute de notes d'étudiants inscrits à certains modules.

1. Les étudiants vous demandent de diffuser leurs notes. Peut-on diffuser ce tableau ? Pourquoi ?

Nom: Prénom:

NoEtudiant	Nom	Prenom	CodeModule	Resultat
23794	Artemia	Adrien	BD_L1	15
23794	Artemia	Adrien	APOO_L1	12
32911	Grandet	Simon	BD_L1	8
32911	Grandet	Simon	APOO_L1	10
33818	Colin	Cynthia	APOO_L1	13
34812	Mollenski	Marie	BD_L1	5
34812	Mollenski	Marie	BD_L1	18

FIGURE 1 – Relation T1 extraite d'un systèmes de de gestion de notes

SHA2(str, hash_length)
Calculates the SHA-2 family of hash functions (SHA-224, SHA-256, SHA-384, and SHA-512). The first argument is the plaintext string to be hashed. The second argument indicates the desired bit length of the result, which must have a value of 224, 256, 384, 512, or 0 (which is equivalent to 256).
The return value is a string in the connection character set.

```
SELECT SHA2('abc', 224);
'23097d223405d8228642a477bda255b32aadbce4bda0b3f7e36c9da7'
```

...

FIGURE 2 – Documentation MySQL de la fonction `SHA2`

2. Il est décidé de publier ceci, mais après une mise en place de pseudonymisation. C'est la fonction de hachage `SHA2` qui a été retenue. Quel(s) attribut(s) va-t-on hacher ? Quel(s) attribut(s) va-t-on conserver pour publication ? Justifier à chaque fois.

3. Comment un étudiant peut-il retrouver ses notes dans le fichier pseudonymisé publié ?

4. La figure 2 donne un extrait de la documentation MySQL de la fonction `SHA2`. Donner le code SQL qui

Age	Pr. systolique	Pr. Diastolique	Glycémie	Fréq. Card.	Risque
15	76	49	6.4	77	faible
15	120	80	6.8	70	faible
19	90	70	7.8	80	faible
19	120	80	7.0	70	moyen
19	120	80	7.0	70	moyen
20	100	90	7.5	88	faible
22	120	85	7.5	88	faible
23	140	90	6.8	70	élevé
23	130	70	7.01	78	moyen
23	120	90	7.8	60	moyen
25	140	100	7.2	80	élevé
32	140	90	18.0	88	élevé
35	100	70	7.5	66	faible
38	110	70	7.9	80	moyen
55	100	65	7.5	66	faible
56	120	80	13.0	70	élevé

FIGURE 3 – Extrait de données médicales sur les risques liées aux maternités

pseudonymise ce qui doit l'être et ne publie que les attributs qui doivent l'être. Pour rendre l'approche plus robuste, ordonner correctement les résultats.

5. Que penser de la robustesse d'une telle publication ? Proposer le cas échéant une attaque permettant de retrouver les notes des autres étudiants.

3 Apprentissage respectueux de la vie privée

On considère les données de la figure 3 qui sont un très petit extrait (issues des recherches [1]) de celles recueillies dans différents hôpitaux, cliniques communautaires et maternités au Bangladesh et donnant un niveau de risque sur les maternités de patientes. On a les attributs suivants :

— âge : âge en années de la patiente enceinte ;
— pression systolique : valeur supérieure de la pression artérielle en mmHg ;
— pression diastolique : valeur inférieure de la pression artérielle en mmHg ;
— BS : glycémie exprimée en termes de concentration molaire, en mmol/L ;
— fréquence cardiaque : fréquence cardiaque normale au repos, en battements par minute ;
— niveau de risque : niveau d'intensité du risque de la grossesse. précédents.

On vous demande de mettre en place une démarche d'apprentissage machine pour prédire le niveau de risque de la grossesse à partir des valeur des autres attributs. On vous alerte sur le fait que ces données brutes ne peuvent pas vous être données car vous n'avez pas le consentement pour travailler dessus.

1. Expliquer comment on peut mettre en place une démarche d'apprentissage tout en respectant la vie privée.

2. Quel(s) quasi-identifiant(s) peut-on définir pour cette base de données ? Justifier.

1. Ahmed M., Kashem M.A., Rahman M., Khatun S. (2020) Review and Analysis of Risk Factor of Maternal Health in Remote Area Using the Internet of Things (IoT). In : Kasruddin Nasir A. et al. (eds) InECCE2019. Lecture Notes in Electrical Engineering, vol 632. Springer, Singapore.

Nom: Prénom:

Age	Pr. systolique	Pr. Diastolique	Glycémie	Fréq. Card.	Risque
{15,19}	76	49	6.4	77	faible
	120	80	6.8	70	faible
	90	70	7.8	80	faible
	120	80	7.0	70	moyen
	120	80	7.0	70	moyen
{20,22, 23,25}	100	90	7.5	88	faible
	120	85	7.5	88	faible
	140	90	6.8	70	élevé
	130	70	7.01	78	moyen
	120	90	7.8	60	moyen
	140	100	7.2	80	élevé
{32,35,38}	140	90	18.0	88	élevé
	100	70	7.5	66	faible
	110	70	7.9	80	moyen
*	100	65	7.5	66	faible
*	120	80	13.0	70	élevé

FIGURE 4 – Jeu de données généré pour la question 4. du dernier exercice

3. Proposer ci dessous une version 2-anonyme autre que celle présentée à la figure 4. Inutile de recopier toutes les valeurs des différents attributs.

Age	Pr. systolique	Pr. Diastolique	Glycémie	Fréq. Card.	Risque

4. On considère que l'attribut sensible est le risque. Pour cela on a construit la table donnée à la figure 4.
   (a) Décrire cette table en termes de k-anonymat, en terme de l-diversité.

   (b) Calculer l'indice de t-proximité associé.