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

Aucun document n'est autorisé. Toutes les réponses doivent être justifiées. Sans justification, une réponse est considérée comme fausse.

Dans toute cette partie, on considère le jeu de données de la Table 1. Cette table contient 12 enregistrements précisant les attributs tels que

— *Nom* est l'identifiant,
— *Genre*, *Département* (d'enseignement) et *Statut* sont les quasi-identifiants ;
— *Salaire* est l'attribut sensible.

Nom	Genre	Département	Statut	Salaire ($K)
Adams	F	Info.	Etud.	10
Baker	H	Maths.	Prof.	60
Cook	F	Maths.	Prof.	100
Dodd	F	Info.	Admin	38
Engel	H	Stats.	Prof.	72
Flynn	F	Stats.	Prof.	88
Grady	H	Info.	Admin	40
Hayes	H	Maths.	Prof.	72
Irons	F	Stats.	Etud.	12
Jones	H	Stats.	Etud.	15
Knapp	F	Maths.	Prof.	100
Lord	H	Info.	Etud.	10

TABLE 1 – Jeu de données concernant une université fictive

1 Attaque d'une base de données statistique

Dans cette section on considère que cette table est une base de données statistique telle que $n = 3$.

1. Préciser quelles sont les requêtes autorisées et lesquelles ne le sont pas dans une base de données statistique telle que $n = 3$.

2. Vous savez qu'Engel est un professeur en statistiques présent dans cette base. A l'aide d'un tracker général et tout en respectant les contraintes d'une base de données statistiques :
   (a) trouver le nombre total d'enregistrements dans cette base ;

   (b) trouver la somme de tous les salaires des personnes dans cette base ;

(c) montrer que Engel est le seul professeur en statistiques présent dans cette base ;

(d) trouver son salaire.

2 k-anonymat

On souhaite publier des indicateurs par département d'enseignement, tout en protégeant la vie privée des personnes dans cette base de données.

On considère la hiérarchie de généralisation suivante :

— *Genre* : H, F $\rightsquigarrow$ *

— *Département* : Info., Maths, Stats $\rightsquigarrow$ *

— *Statut* : Etud., Prof., Admin. $\rightsquigarrow$ *

1. Donner une version 2-anonyme de cette table en dernière page préservant l'attribut Département.
2. Quels problèmes peut-on rencontrer lorsqu'on publie une base de données k-anonyme ? Montrer que le problème apparaît ici.

3. Calculer la valeur de Loss pour cette transformation.

4. On a vu en CM/TD la méthode de Mondrian pour produire une base de données 2-anonyme. Quel est l'intérêt d'un tel algorithme ? Cela est-il sensé corriger le problème soulevé à la question 2 ? **Attention**, on ne demande pas d'appliquer la méthode de Mondrian pour produire une base 2-anonyme.

Réponse à la question 2.1