Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Max Berrendorf, Julian Busch

Knowledge Discovery and Data Mining I

WS 2018/19

Exercise 3: Privacy, Frequent Itemset Mining

Exercise 3-1 Privacy

Given the following table

Key	**Quasi-Identifier**			**Sensitive**
Name	*Sex*	*Age*	*Zip*	*Disease*
Alice	F	24	10000	Heart Disease
Bob	M	22	10000	Lung Cancer
Charlotte	F	24	10000	Breast Cancer
Dave	M	22	10000	Lung Cancer
Emma	F	20	10000	Heart Disease
Francis	M	20	10000	Heart Disease
Garry	M	22	10000	Lung Cancer
Harry	M	20	10000	Heart Disease
Iris	F	21	10001	Flu
John	F	21	10001	Flu
Kendra	F	20	10000	Heart Disease
Lisa	F	20	10000	Lung Cancer

(a) k-Anonymity:

(i) Determine the largest k such that the table is k-anonym. Explain which rows contradict the $(k+1)$-anonymity.

(ii) You may now use suppression on the columns. Assume that by removing one digit from *Age* or *Zip*, or suppressing the *Sex* attribute, you lose one ”value”. What is the minimal value loss required to achieve 5-anonymity?

(b) Distinct l-Diversity

(i) What is one shortcoming of k-anonymity compared to l-diversity? Which attack exploits this weakness?

(ii) Given that a dataset is k-anonymous, but not $(k+1)$-anonymous. What implications does this have on the distinct l-diversity of the dataset? Give a lower and upper bound for l.

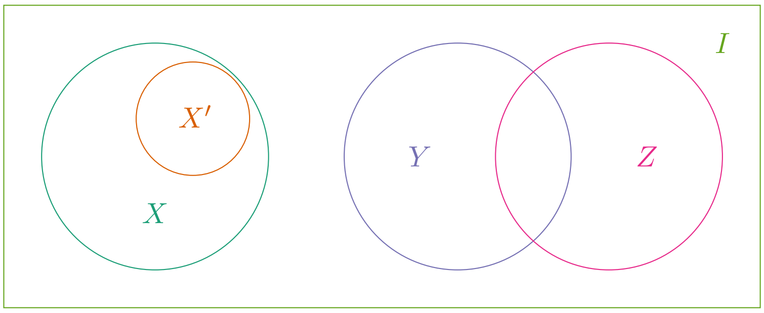
(iii) Knowing only the distribution of the sensitive attribute values; What bounds can you derive for l in distinct l-diversity?

(iv) What is the largest l such that the above mentioned dataset is distinct l-diverse?

(v) Assume suppressing the last digit of the *Zip* column and generalising *Age* to $\{(-\infty, 22], (22, +\infty)\}$. For what value of l can distinct l-diversity now be guaranteed.

Exercise 3-2 Apriori Principle

The apriori principle can be used to prune candidates for frequent itemsets and association rules. Let I be the set of all items. You can use the following Venn diagram as help to understand the subset relations in the following tasks.



Give proofs or counterexamples for the following claims:

(a) Let $X \subseteq I$ be an arbitrary itemset. Then, $supp(X') \geq supp(X)$ holds for any non-empty subset $\emptyset \subset X' \subseteq X$.

(b) Let $Y, Z \subseteq I$ be arbitrary itemsets with $|Y| > |Z|$. Then, $supp(Y) \leq supp(Z)$.

(c) Let $X \subseteq I$ be a frequent itemset. Then, every non-empty subset $\emptyset \subset X' \subseteq X$ must also be frequent.

(d) Let $X \Rightarrow Y$ be not strong. Then, for all $Z \subseteq I$ holds $X \Rightarrow (Y \cup Z)$ not strong.

(e) Let $X \Rightarrow Y$ be not strong. Then, for all $X' \subseteq X$ holds $(X \setminus X') \Rightarrow (Y \cup X')$ not strong.

Exercise 3-3 Apriori Algorithm

Given a set of items $\{a, b, c, d, e, f, g, h\}$ and a set of transactions T according to the following table

TID	Items
1	ag
2	bcg
3	eg
4	dg
5	dfg
6	dg
7	ag
8	ag
9	ae
10	ag
11	afh
12	af
13	ade
14	dfg

(a) Using the Apriori algorithm, compute all frequent itemsets for $minSup = 0.1$ (i.e. 2 transactions are needed for an itemset to be frequent).

(b) Which of the found frequent itemsets are closed/maximal? Is there a dependency between those two concepts?