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.

The dataset is 2-anonymous, as there is no Quasi-Identifier-tuple which occurs only once. It is not 3-anonymous, as e.g. $(F, 24, 10000)$ occurs only twice.

(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?

5-anonymity can be achieved by suppressing the last digit of Age and the last digit of Zip. Hence, the minimal value is at most 2. It is not 1 as:

- Suppressing *Sex* leads to 2-anonymity, e.g. $(*, 24, 10000)$ occurs only twice.
- Suppressing the last digit of *Age* leads to 2-anonymity, e.g. $(F, 2*, 10001)$ occurs only twice. Suppressing the first digit does not give any benefit, as all age numbers begin with "2".
- Suppressing the last digit of Zip leads to 2-anonymity, e.g. $(F, 24, 1000*)$ occurs only twice. Suppressing any other digit does not give any benefit, as all zip codes begin with "1000".

(b) Distinct l-Diversity

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

k-anonymity only regards the quasi-identifiers, but does not investigate the distribution of the sensitive attribute within one equivalence-class w.r.t. the quasi-identifier. This can be exploited by the *Background-Knowledge Attack*.

(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.

The smallest equivalence-class w.r.t. to the Quasi-Identifier has size k. Hence, in this class there can only be at most k different values for the sensitive attribute. Thus, l can be bounded from above as $l \leq k$. Trivially, $1 \leq l$ holds as lower bound. As k-anonymity does not make any statement about the distribution of the sensitive attribute, we cannot guarantee a larger lower bound, i.e. the following bounds are tight: $1 \leq l \leq k$.

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

Let L be the number of different sensitive attribute values. Then, there can also be at most L different values within each equivalence class w.r.t. to an Quasi-Identifier. Thus, $l \leq L$.
Additional information: This bound is independent of the bound from (ii), as the former one operates only on the Quasi-Identifier columns and this one solely considers the sensitive attribute.

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

The dataset is distinct 1-diverse as $QI = (F, 21, 10001) \implies Disease = Flu$.

(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.

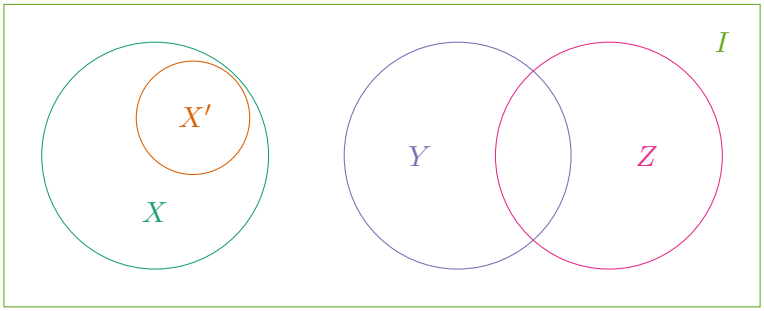
There are the following equivalence classes

Sex	Age	Zip	Diseases	l
F	$(-\infty, 22]$	$1000*$	{Flu, Heart Disease, Lung Cancer}	3
M	$(-\infty, 22]$	$1000*$	{Heart Disease, Lung Cancer}	2
F	$(22, \infty)$	$1000*$	{Breast Cancer, Heart Disease}	2

Hence, the table is now distinct 2-diverse.

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 $\emptyset \subset X' \subseteq X$.

For all $T \subseteq I$ we have $X \subseteq T \implies X' \subseteq T$ by transitivity of the subset relation. Hence, $\{T \in D \mid X' \subseteq T\} \supseteq \{T \in D \mid X \subseteq T\}$ and thus $supp(X') \geq supp(X)$.

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

Counterexample: Consider $I = \{a, b, c\}$, $Y = \{a, b\}$, $Z = \{c\}$ and $D = \{t\}$ with $t = \{a, b\}$. Then, $supp(Y) = 1 > 0 = supp(Z)$.

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

As X is frequent, $supp(X) \geq minSup$. Then, by (a) it follows: $supp(X') \overset{(a)}{\geq} supp(X) \geq minSup$. Thus, by definition, X' is frequent as well.

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

$X \Rightarrow Y$ is not strong. Hence,

$$minConf > conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} \overset{(a)}{\geq} \frac{supp(X \cup Y \cup Z)}{supp(X)} = conf(X \Rightarrow (Y \cup Z))$$

Thus, $X \Rightarrow (Y \cup Z)$ is also 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.

$X \Rightarrow Y$ is not strong. Then,

$$\begin{aligned} minConf > conf(X \Rightarrow Y) &= \frac{supp(X \cup Y)}{supp(X)} \overset{(a),\clubsuit}{\geq} \frac{supp((X \setminus X') \cup (Y \cup X'))}{supp(X \setminus X')} \\ &= conf((X \setminus X') \Rightarrow (Y \cup X')) \end{aligned}$$

where $\clubsuit$ also exploits that $(X \setminus X') \cup (Y \cup X') = X \cup Y$ for $X' \subseteq X$.

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).

k	candidate	prune	count	threshold	closed	maximal
1	a		8	a	✓	
	b		1			
	c		1			
	d		5	d	✓	
	e		3	e	✓	
	f		4	f	✓	
	g		10	g	✓	
	h		1			
2	ad		1			
	ae		2	ae	✓	✓
	af		2	af	✓	✓
	ag		4	ag	✓	✓
	de		1			
	df		2	df		
	dg		4	dg	✓	
	ef		0			
	eg		1			
	fg		2	fg		
3	aef	with ef				
	aeg	with eg				
	afg		0			
	dfg		2	dfg	✓	✓

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

Maximal implies *closed*. To this end, observe that if X is frequent and maximal, then for all $Y \supset X$ holds $supp(Y) < minSup$. As X is frequent, $supp(X) \geq minSup$. Hence, for all $Y \supset X$ holds $supp(Y) < minSup \leq supp(X)$, which implies X being closed.