Lecture "Privacy Enhancing Technologies for Data Science" (PETs4DS)
Winter Semester 2016/17
Informatik 5, Lehrstuhl Prof. Decker

# Assignment 2

Date for exercise session : 6.12.2016

## Information:

1. Participants are encouraged to present their solutions for the exercises.
2. In addition, solutions to the exercises will be presented at the exercise session and published in L2P.
3. Working on the exercises in groups is welcome.
4. The assignments of this lecture series are not graded.
5. Instead, participation in the exams depends on passing the presence exercise in January.

## Exercise 1: K-anonymity theory

(a) How many equivalence classes has a table with x entries and Parameter k.

At most x/k equivalence classes

(b) What is the probability that a record can be identified, when an adversary has all the quasi identifiers of an individual?

Probability of identification = 1/k

(c) What does k-anonymity protect against and why?

Protects against re-identification by linking. By introducing indistinguishable records, it ensures that the probability of re-identification is 1/k

(d) What is the difference between a homogeneity and a background knowledge attack?

Homogeneity attack occurs when attacker can make inference due to the lack of diversity of the sensitive attributes. Background knowledge attack occurs when the attacker makes use of some known facts to make inferences

## Exercise 2: L-Diversity theory

As shown in the lecture, there are situations in which k-anonymity does not provide sufficient protection from de-anonymisation. Additional protection is provided by the l-diversity anonymity measure. L-diversity has at least two different definitions:

**distinct l-diversity:** Each equivalence class has at least l sensitive values.

**entropy l-diversity:** This is a version of l-diversity defined via measuring of entropy.

Your task is to come up with a definition of entropy based l-diversity through answering the following questions:

(a) What is entropy and how is it defined?

(b) Why does it make sense to use entropy as a measurement to calculate l-diversity?

Entropy has the ability to measure the information content of a random variable. We can use this fact to model l-diversity by setting a threshold, above which we believe that there is enough information content to provide any discriminating evidence to the attacker

(c) To calculate entropy a probability is needed. How would you define this probability in order to calculate l-diversity? Create a probability mass function for each unique sensitive attribute based on their occurrence in the equivalence class

(d) Are the results of distinct l-diversity and entropy l-diversity on the same data set correlated? Yes, since l and log l are correlated

## Exercise 3: T-closeness theory

Background knowledge attack

(a) L-diversity still does not solve the problem of privacy leaks. Name two attacks on data which has been anonymised using l-diversity and explain them.

Then there would be no point of publishing the data in the first place

(b) These attacks are possible, because sensitive attributes are released. Why are they released and not anonymized like the quasi identifiers?

(c) What is the difference between t-closeness and l-diversity and does t-closeness solve the problems of l-diversity?

t-closeness says that the distribution of sensitive attr in the equivalence classes should follow the distribution of the actual dataset.

l-diversity says that the members of the equivalence class should be “diverse enough”

Yes, skewness attack is taken care of

## Exercise 4: Apply anonymity metrics to a table

A Table representing the income based on location and age collected by the revenue office has to be processed for public release. The table should allow determining the average income per region. However the data shown in Table 1 contains personal identifiable information.

(a) Your task is to anonymize the table first with k-anonymity assuming that the name is an identifiable attribute, age and income are quasi-identifiers and location is a sensitive attribute. Furthermore ignore the change of order in the Location vector while doing k-anonymity and just keep the order for the next exercise parts. Hide sensitive attributes, partition on age 20-35, 36-70, partition on salary <=4000 and >4000. Assuming k=3

(b) Are there any entries you need to change to satisfy distinct l-diversity with l $= 3$? Yes!

(c) Find a $c$ such that recursive (c,3)-diversity is satisfied. (c,l)-diversity is described on slide 24 of Chapter 3.

(d) Calculate the entropy of the biggest equivalence class with entropy l-diversity. Is the outcome a good value?

| Name | Age | Income | Location |
|---|---|---|---|
| Alice | 23 | 450 | A |
| Bob | 25 | 697 | D |
| Charlie | 38 | 3000 | A |
| Dave | 69 | 1500 | K |
| Eve | 61 | 7000 | A |
| Felix | 52 | 5000 | D |
| Gertrud | 31 | 3000 | D |
| Hilde | 28 | 450 | A |
| Ina | 28 | 3800 | K |
| Julian | 40 | 5000 | M |
| Klaus | 58 | 4000 | A |

Table 1: A = Aachen, D = Duesseldorf, K = Koeln, M = Moenchengladbach

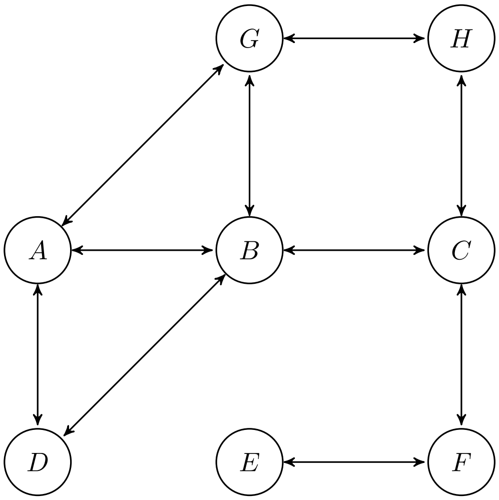
## Exercise 5: Graph anonymization with k-degree anonymity

The lecture showed the intuition behind an algorithm for k-degree anonymity. The full algorithm is described in Liu, Kun, and Evimaria Terzi. "Towards identity anonymization on graphs" Proceedings of the 2008 ACM SIGMOD international conference on Management of data. ACM, 2008.

(a) Apply the description of the intuition behind the k-degree anonymity algorithm, as presented in the lecture, to the following graph. Use k=3. Find all possible solutions.

(b) Which of these is the best solution and why?

(c) Which of these solutions does also fulfil the condition of k-neighbourhood anonymity for k=3.



## Exercise 6 (optional): De-anonymization of a graph

In the tasks above we looked at the problem of graph anonymization. A graph is defined as following: $G = \{V, E, X\}$, where X is the set of attributes a node has. In the case of a social Network X could consists of name, age, political affiliation and relationship. The function $\mu(x)$ returns the value of the attribute of an attribute $x \in X$.

(a) In the table based anonymization we found out that the approaches are vulnerable against background knowledge. What could be viable background knowledge to de-anonymise the given graph $G$?

(b) In the lecture we discussed a so called active attack, where an adversary places a small identifiable network into the larger social network. This provides the starting point for de-anonymising any anonymised release of the social network. Think of some properties this small network needs to have such that it can be found and re-identified by the adversary in the anonymised social network data. You answer only needs to take into account the anonymisation approaches which were covered in the lecture.

(c) Describe briefly how you could use this background knowledge to achieve identity disclosure or content disclosure.

(d) Assume now $|X| = 1$. Think of a case where you dont need background knowledge to disclose the hidden attribute X of a node.