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.

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

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

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

Exercise 1: Solution

(a) The size of an equivalence classes is $\geq k$. Therefore the maximum number of equivalence classes is $\lfloor \frac{x}{k} \rfloor$. However an equivalence class can also consist of x entries, so the number n of equivalence classes is $1 \leq n \leq \lfloor \frac{x}{k}. \rfloor$

(b) The probability of re-identifying an individual is $\frac{1}{k}$.

(c) k-anonymity protects against re-identification but not against attribute disclosure.

(d) A homogeneity attack is possible when an equivalence class only contains one value of a quasi-identifier, or if the values of quasi-identifiers in an equivalence class is very similar. A background knowledge attack uses knowledge from other sources to aide in re-identification the subject.

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?

(c) To calculate entropy a probability is needed. How would you define this probability in order to calculate l-diversity?

(d) Are the results of distinct l-diversity and entropy l-diversity on the same data set correlated?

Exercise 2: l-diversity solutions

(a) Entropy in this context is a measure for uncertainty of an item. In general, entropy is defined as $H(X) = -\sum_{i=1}^{n} p(x_i)log_2(p(x_i))$, where $p(x_i)$ is the probability the outcome x_i for the discrete random variable X.

(b) It makes sense to use it, because you can calculate the uncertainty of an item, which in this case is the value of an attribute in an Equivalence class. We are interested in a high Entropy value, because then the equivalence class can be said to be save.

(c) This probability is simply the fraction how often a value appears in an Equivalence class. This can be denoted as $p(E, s)$, where the sum is taken over the $s \in S$ which are all values of a sensitive attribute. If we use this, we get the entropy of an equivalence class E as follows: $H(E) = -\sum_{s \in S} p(E, s)log_2(p(E, s))$

(d) Yes, both measure the same property of the anonymisation, mainly how many values of a sensitive attribute are in each equivalence class. However, while distinct l-diversity counts the distinct values of the sensitive attribute, entropy l-diversity measures the entropy of the sensitive values. Therefore the results of l-diversity and entropy l-diversity are correlated.

Exercise 3: T-closeness theory

(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.

(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?

Exercise 3: T-closeness theory solutions

(a) l-diversity does not protect against the skewness attack or against the similarity attack. The skewness attack is possible when the distribution of a sensitive value in an equivalence class is very different from the distribution of the value in the table as a whole. This difference can be used to infer properties of the members of the equivalence class. The similarity attack is possible when the different values of a sensitive attribute in an equivalence class are very similar, in the sense that they are different, but with domain knowledge e.g. about medicine, they are similar enough to give an attacker a lot of knowledge about all the individuals in an equivalence class.

(b) Releasing the sensitive value is the reason why the data set is released at all. Changing / removing them would make the data set useless.

(c) l-diversity measures the number of different values for a sensitive attribute, whereas t-closeness measures the similarity between values of a sensitive attribute. T-closeness addresses the skewness and similarity attacks, however it still cant mitigate all forms of background knowledge attacks.

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.

Name	Age	Income	Location
Alice	23	450	A
Bob	25	697	D
Charlie	38	3200	A
Dave	69	1500	K
Eve	61	6999	A
Felix	52	4900	D
Gertrud	31	2600	D
Hilde	28	450	A
Ina	28	3800	K
Julian	41	4440	M
Klaus	58	3100	A

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

(a) Your task is to anonymize the table first with k-anonymity for $k = 3$ 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.

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

(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?

Exercise 4: Solution

(a) (i) First remove identifiable attributes and select a quasi identifier, to which all other quasi identifiers will be grouped to and order them by size. We select Age for this. Furthermore group quasi identifiers to intervals.

(ii) The second step is to check if the Age column satisfies k-anonymity with k=3. And if not resize the equivalence classes.

(iii) Final step is to resize the income attribute equivalence classes.

Name	Age	Income	Location
*	20-30	0-1	A
	20-30	0-1	D
	20-30	0-1	A
	20-30	3-4	K
	30-40	3-4	A
	30-40	2-3	D
	40-50	4-5	D
	50-60	4-5	A
	50-60	3-4	K
	60-70	1-2	M
	60-70	6-7	A

Name	Age	Income	Location
*	20-30	0-1	A
	20-30	0-1	D
	20-30	0-1	A
	20-30	3-4	K
*	30-50	2-3	A
	30-50	2-3	D
	30-50	4-5	D
*	50-70	4-5	A
	50-70	3-4	K
	50-70	1-2	M
	50-70	6-7	A

Name	Age	Income	Location
*	20-30	0-4	A
	20-30	0-4	D
	20-30	0-4	A
	20-30	0-4	K
*	30-50	2-5	A
	30-50	2-5	D
	30-50	2-5	D
*	50-70	1-7	A
	50-70	1-7	K
	50-70	1-7	M
	50-70	1-7	A

(b) Yes the second equivalence class needs to have 3 distinct location values. One D could be changed into a K, but this would modify the sensitive values.

(c) $r1 = A$, $r2 = D$, $r3 = K$, $r4 = M$,
with $|r1| = 5$, $|r2| = 3$, $|r3| = 2$, $|r4| = 1$.
Therefore $5 < c(2 + 1)$ and $c > 5/3$.

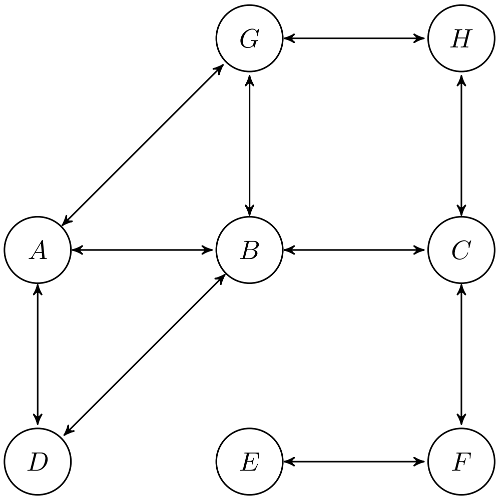
(d) Entropy of one equivalence class E: $H(E) = -\sum_{s \in S} p(E, s)log_2(p(E, s))$, where $p(E, s)$ is the fraction of the number of attributes taken value s in the equivalence class E. If we consider the 3rd equivalence class, we get $-\sum_{s \in S} p(E, s)log_2(p(E, s)) = -((1/2 * log_2(1/2)) + (1/4 * log_2(1/4)) + (1/4 * log_2(1/4))) = 1.5$. As we used the logarithm with a base of 2, the outcome of the formula is 2 Shannon. 1 Shannon is the equivalent to the uncertainty of a coin toss (1 Shannon is approx. 0.693). Since we only got 3 distinct values this outcome can be considered good.

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.

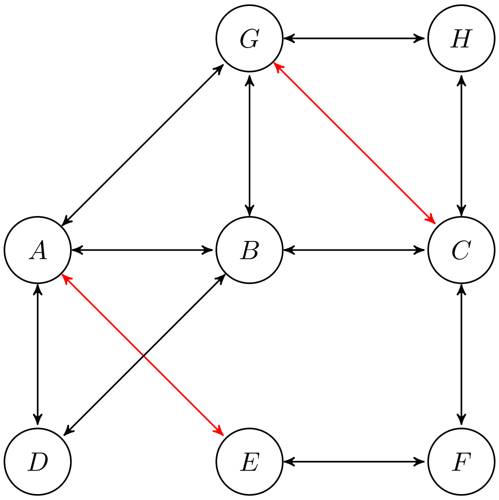
(a) Apply the description of the intuition behind the k-degree anonymity algorithm, as presented in the lecture, to the following graph. Use k=3. Find all possible solutions.

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



Solution 5

(a) The degree vector is : $(4, 3, 3, 3, 2, 2, 2, 1)$ to statisfie k=3 there are 3 possible solutions. $(4, 4, 4, 4, 4, 4, 4, 4)$, $(4, 4, 4, 4, 2, 2, 2, 2)$, $(4, 4, 4, 4, 4, 2, 2, 2)$. $(4, 4, 4, 3, 3, 3, 3, 3)$ is not a solution, because the sum of the degree vector results in an odd number. This is a problem because to an odd degree vector there does not exist a graph, even if it does satisfies k-degree-anonymity.



(b) The best one is the 3rd vector $(4, 4, 4, 4, 2, 2, 2, 2)$, because it results in the least added edges and therefore introduces the minimal amount of information loss, since a lot of information is stored in the structure of the network.

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.

Solution 6

(a) Background knowledge to de-anonymize a graph could be another graph. For example, a social network publicly available could be used to de-anonymize another secured social network. This technique was used to de-anonymize the secure Twitter graph with the help of the not anonymized and publicly available Flicker graph.
Furthermore graph metrics such as clustering coefficient, power law distribution, node degree information, edge count, closeness and centrality measures of the network before the anonymization could be useful background information.
Furthermore a lot more graph measurements are suited for graph de-anonymization. On top of this the background knowledge suited for table based attacks are also suited, leaving the question open as to how secure anonymised networks truly are.

(b) This network is also called a seed network. This network needs to have some unique properties such that after the anonymization these unique properties are preserved. We don't know which algorithm was used for anonymization. However it can be assumed that the anonymization technique used in the attack, has the property of keeping the general structure of the network preserved. For the attack discussed in the lecture, it turned out to be useful to use a k-clique as a seed network, because cliques of certain sizes are very rare and can be re-identified using many graph metrics.

(c) One possible attack could be, if you take a seed network G_{seed} and another publicly available not anonymized auxiliary network G_{aux} as background knowledge, then you could construct a matching attack. First you locate the seed network in the anonymized graph G_{anony} and G_{aux}. After this you could compute similarities between the nodes from G_{aux} and G_{anony}. The similarity measure can be based on a subset of graph measurements. Starting from the seed network, the network can be de-anonymized for instance with a suited machine learning algorithm.

(d) No background knowledge is needed if the graph itself reveals enough information about their nodes. This is for example the case in a social network. Imagine private and public profiles. Poorly constructed networks could show private profiles in the friend list of a public profile. To predict the hidden value, one could for instance compute a distribution over all public values connected to a private profile. The distribution attributes for the public nodes might allow inferring the attribute of the private profile.