Barriers to the implementation of k-anonymity and related microdata anonymization techniques in a realworld application

Barriers to the implementation of k-anonymity and related microdata anonymization techniques in a realworld application

Andreas Wiegnand, 1878334
Ludwig Schallner, 1850413

1 Introduction

Nowadays data is a key factor in nearly every domain. It is comparable with the gold rush of the 19. century [9]. Furthermore storage space and network connectivity become affordable [11]. But to use the data for commercial or scientific purposes the privacy of the data holder does not have to be compromised or in other words the data holder need to know how to produce anonymous data otherwise the database can not survive, because if the information of the table get released there will be no need anymore for this data [11].

The paper The so called k-anonymity method, which produce anonymous data, theoretical. But there are practical barriers that will occur in the real world. Those prevent such implementation. The goal of k-anonymity is, to prevent the possibility to get information about the real individual, or at least with k other possible individuals. So if a individual is discriped by a tuple of $\langle f_1, ..., f_n \rangle$ features and each feature can have $\langle a_1, ..., a_n \rangle$ attributes. There are at least k other individual with the same attribute for each feature so that there is no possibility to reduce the real individual and there will be at least k individuals with the same tuple[11].

The attributes that are used to link the external data is called quasi identifiers. Typical values for them are gender, date of birth and zip code [7]. We will present techniques that override k-anonymity and get the real individual. Another problem we will introduce is, that the producing of k-anonymity of a computational view is a NP-hard problem, like meyersond and williams shown. ...

2 Basics

Microdata:
First of all it should be clear what microdata is, those data is containing records of information about individuals. The upside versus the more known summary or aggregate data is, that microdata is naturally flexible. Everyone who has this data can perform own statistics from that data [1].

Quasi-identifier:
Even tho though explicit identifier got removed of published data. Such non explicit attribute, which would not uniquely identify the record owner. But if combined (with other non explicit attributes), they become explicit identifier. Which resulting that those can link towards the owner. In such a case those attributes are called quasi-identifier [3]. Such process is shown in figure 1.

3 Barriers of implementation of k-anonymity

Like Dalenius already mentioned it is absolutely necessary that an attacker, under no circumstances, can learn about whatsoever target if he is studying the published database. Not even if the attacker has background knowledge from any other source [2]. Unfortunately like Dwork showed 2006 that such safty is impossible because of background knowledge. For example if the attacker knows that Bob get paid twice as the average german man and the attacker got access to a database which publish the average income by german men. The anonymity of Bob is compromised even if Bobs data is not in the database [4].

Mehr Text? differential privacy erklren/ bzw formalized as a relaxed version of semantic security, cannot be achieved, while semantic security for cryptosystems can be achieved

3.1 Linking data

A barrier to do the implementation of k-anonymity, the attacker can take another dataset and link both together to get rid off the k-anonymity and infer the real individual. This process is called linking data and was first described by Sweeney[11]. She showed that with a example of health care data from 37 states in the USA. The institute from which she bought the data, insures the anonymity of the individuals. Sweeney purchased the voter registration list for Cambridge Massachusettts and received information of the voters including ZIP code, birth date and gender (non explicit identifier) of each voter. She linked that information with the medical data. It was possible to deanonymize the data and get ethnicity, visit date, diagnosis, procedure, medication and total charge of some patients [11].

You got two datasets A and B. Each dataset got $\langle f_1, ..., f_n \rangle$ features and $\langle r_1, ..., r_n \rangle$ rows. Each row is then a tuple r_i with n features $\langle f_1, ..., f_n \rangle$ describing the individual. Even tho the data is k-anonimized you can get rid oft he

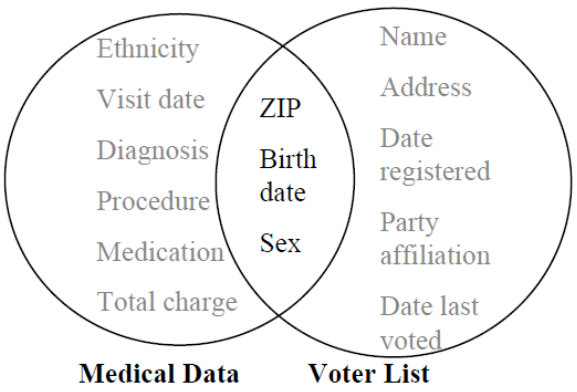

Fig. 1. linking data

anonymity of the individual by linking the A to B. So if $A \cap B \neq \emptyset$ it is possible to infer the anonymized individual [11]. As a result any attacker who know such data (ZIP Code, Birth date and sex) could easily identify with such an attack his victim. For example Peter see his ex-wife at the doctor, most likely he knows her ZIP-Code, Birth date and sex. Therefore he finds out what she is suffer from.

3.2 Unsorted matching attack against k-anonymity

There is a possibility of a leak of information, if the release k-anonymity data is in some kind of a sort release. This mean the numerical attributes are descending or ascending sorted and attributes, which be of characters are alphabetical ordered, can give the attacker Information about the sensitive data. To prevent this attack, just get the data into a random order with a pseudo randomized sorting algorithm [11]. As an example take a look at the table 3: matching attack will give an example on that. If you compare the different release generalized tables you can figure out all quasi identifier of those [11].
...

3.3 Complementary release attack against k-anonymity

The problem of complementary release attack against k-anonymitym lies by the release of other k-anoymitizet tables from the same dataset. To stop the attack the data holder should consider for each release of data if it's possible to release information with older released data. This is hard to avoid especially when the data can come from different individuals [11].
...

Table 1. matching attack

Age	ZIP
2	91058
4	91058
50	27785
52	27785
20	32105
21	32105
31	67676
32	67676

Age	ZIP
*	91058
*	91058
5*	27785
5*	27785
2*	32105
2*	32105
3*	67676
3*	67676

Age	ZIP
2	91*
4	91*
50	27*
52	27*
20	32*
21	32*
31	67*
32	67*

3.4 Temporal attack against k-anonymity

The data can be change over time. New tuples might be added or persistent ones can be changed. If the GT0 was release at time T=0 and on a later time GT1 will be released at time T = 1 with new tuples of information. Both tables at their time stamp T are k-anonymity, but it will not be checked if they are k-anonymity between them. So their is a possibility of information leaking and a failure oft he k-anonymity in both tables [11].
...

3.5 Homogeneity Attack

3.6 Background Knowledge Attack

Taking background knowledge attack from a person and take it into account to derive the sensitive data. In our example of table 1 this could be that we know one person my name, age, and nationality. Additionally we know, because that she is asian, in would be unusual that she got diabetes, because diabetes is a uncommon sickness in japan [7].
...

3.7 Complexity of producing k-anonymity

Till now we only looked at problems of information leaking and privacy problems for individuals. Data is personal-specific information which is structured as a table in rows and columns. Rows a tuple. The columns are attributes with are a set of values which describe the certain attribute. A tuple specify a person. K-anonymity is about protecting the identity of a person not relationships of companies or governments. So the goal of k-anonymity is, not getting more information by linking the data to external data. The bridge between the data and external data is called "quasi-identifier". Examples for that would be ZIP, gender, birth date etc..

Generalization mean, replacing a value with a less specific but semantic identical value. For example we got a list of forenames of buys, (Achmed, Achilles, Achim). To generalize this names you can just (Ach*,Ach*, Ach*) delete the last chars of the name. So there is a less specific domain and now more generalize through this mapping. Suppression on the other hand means not releasing the value at all.
...

4 Barriers against L-Diversity

K-anonymity gives protection on the identity disclosure. L-Diversity more with the same sensitive data.

Skewness Attack When the overall distribution is skewed, satisfying l-diversity does not prevent attribute disclosure. Now consider an equivalence class that has 49 positive Records and only one negative record. It would be distinct 2-diverse and has higher entropy than the overall table (and thus satisfies any entropy l-diversity that one can impose), even though anyone in the equivalence class would be considered 98% positive, rather than 1% percent. In fact, this equivalence class has exactly the same diversity as a class that has one positive and 49 negative records, even though the two classes present very different levels of privacy risks [7].
...

Similarity Attack When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information [8]. Positive and negative disclosure: the homogeneity attack where A determined that B has cancer is an example of positive disclosure, whereas when an adversary eliminates some possibilities of sensitive tuples is known as negative disclosure. This negative disclosure uses background knowledge attack. L-diversity also fails in the case of multiple sensitive attributes. In short, distributions that have the same level of diversity may provide very different levels of privacy, because there are semantic relationships among the attribute values. Furthermore different values have very different levels of sensitivity, and privacy is also affected by the relationship with the overall distribution. l-diversity does not consider semantic meanings of sensitive values. l-diversity cannot provide privacy for the multiple sensitive attributes [7].
...

5 Barriers against t-Closeness

The publication of Ninghui, Tiancheng and Suresh is about the t-closeness, which takes into account the unavoidable gain of knowledge of an attacker (the weakness of ℓ-diversity) [6]. To archive t-closeness a special distance measurement

method the Earth Mover's Distanz (EMD) is used [10]. Furthermore t-closeness "requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table" [6].

The weakness of t-closeness by using EMD is that it does not takes the into account if a change of pairs is more significant than another one, which an attacker can use [6]. A combination of EMD and Kullback-Leibler Distance [5] will help to prevent this attack [6].

The distribution of the sensitive data in the table is the same like the real data[6]. The t-closeness principle: An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness[6].
...

$$Entropy(E) = -\sum_{s \in S} p(E,s) \log p(E,s)$$

Fig. 2. Definition of a entropy of an equivalence class E

References

1. Ipumsl-confidentiality, https://web.archive.org/web/20070823010133/http://international.ipums.org/international/
2. Dalenius, T.: Towards a methodology for statistical disclosure control. Statistik Tidskrift 15, 429–444 (1977)
3. Dalenius, T.: Finding a needle in a haystack or identifying anonymous census records. Journal of official statistics 2(3), 329 (1986)
4. Dwork, C.: Differential privacy. In: Encyclopedia of Cryptography and Security, pp. 338–340. Springer (2011)
5. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Statist. 22(1), 79–86 (03 1951), https://doi.org/10.1214/aoms/1177729694
6. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. pp. 106–115. IEEE (2007)
7. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. In: Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on. pp. 24–24. IEEE (2006)
8. Maheshwarkar, N., Pathak, K., Chourey, V.: Privacy issues for k-anonymity model. International Journal of Engineering Research and Application 1(4), 1857–1861 (2011)
9. Rossi, B.: Data revolution: the gold rush of the 21st century, http://www.information-age.com/data-revolution-gold-rush-21st-century-2-123460039/
10. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International journal of computer vision 40(2), 99–121 (2000)
11. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05), 557–570 (2002)