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

Abstract

Bla BLA BLA WIR SIND TOLL!

1 Introduction

Nowadays data is a key factor in nearly every domain. It is comparable to the gold rush of the 19^{th}. century [9]. Furthermore, storage space and network ability increasingly become affordable [11]. This is leading to the situation that the created and stored data is often not only useful to the original data holder, but to other researchers. Also, some data is only useful if its get shared with other data and get together analyzed. But those data may contain some personal or sensitive information. Such that the data should only get releases if the privacy is protected [7].

Table 1. Basic example

SSN	Age	Postcode	Problem
680-90-2665	25	4568	procrastination
008-07-4179	34	4567	stress
391-05-7998	48	4569	stomach cancer
078-36-3853	39	4568	obesity
411-71-9290	42	4561	stomach ulcers
527-59-1948	27	4568	stress

Data like in table 1 have to get anonymized before it gets released. A very common technique archive this goal is the so-called k-anonymity, which will prevent the possibility that information about the individual gets leaked. This paper will show the barriers to implementing k-anonymity. In Section 1 explains the mandatory basic to understand k-anonymity and its purpose. Section 2 will discuss the underlying barriers of k-anonymity. In Section 3 we will explain, possible attacks which also have to be considered as barriers for k-anonymity. Section 4 will show multiple algorithms to implement k-anonymity. A summary of the whole paper will be in the last section

2 Basics

In the following subsections basics will be explained.

Microdata: First of all, 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].

Identifier: They are attributes which can identify the record owner explicitly without any other attribute. For example the full name (first name and last name), telephone number, social security number, and more [4].

Quasi-identifier: Even though explicit identifier got removed from published data (to anonymize the data). Attributes which non-explicitly identify the record owner are left. But if they get combined with other non-explicit attributes or other tables, they can reidentify the record owner. In such a case those combination of attributes are called quasi-identifier. For example Gender, Age, Postcode, weight and height [3]. Such process is shown in figure (the quasi-identifier would be the ZIP, birth date and sex) 1.

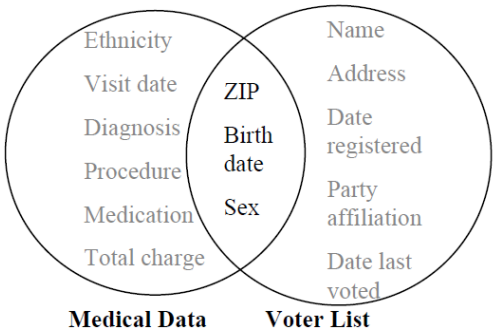

Fig. 1. Quasi-identifiers

Sensitive data: Data which is useful for example researchers but are too private and should not be known publicly nor be accessible for outsiders. This is the data which the record owner do not want to get linked to.[8].

Background-knowledge: Because its unknown what the attackers knows, we have to assume additionally to that he have access to table, the attackers knows that the table is generalized (to guarantee k-anonymity). Furthermore, the attacks is aware of the domain of the attributes.
Instance-level background knowledge: The adversary knows about specific details about his target. For example Alice (the adversary) knows that Bob do not suffer from a disease, because he does not show the symptoms. In this case Alice may can conclude what Bob is really suffers from.
Demographic background knowledge: The Adversary knows more general fact, for example P(t[condition] = cancer— t[Age] $\geq$ 40). With this information the attacker may use it to interference about records [8]

K-Anonymity: The goal of making a k-anonymized table, is to have at least (k-1) tuples of each identical tuple taking the corresponding quasi-identifiers into account [11, 7]. For example the 2-anonyminized version of the table 1 in the introduction section would be the following table:

Table 2. Basic example 2-anonymized

SSN	Age	Postcode	Problem
*	2*	456*	stress
*	3*	456*	stress
*	4*	456*	stomach cancer
*	3*	456*	obesity
*	4*	456*	stomach ulcers
*	2*	456*	stress

Disclosure: There are two kinds, **identity disclosure** if this is happening an individual gets linked to a particular record. Because of that **attribute disclosure** may happens, this is if new information about an individual gets reveled. For example, Bob gets linked to his record in 2, because of some attack (see Section 3.3). The adversary learns that he is suffering from stress [11].

Equivalence class: Is a set of all tuples with the identical quasi-identifiers of a table [7].

Global recoding/domain generalisation: This generalization technique is very common, if a attribute value get generalized then all occourences of that value gets replaced by the generalized one [11, 10, 7, 6].

Local recoding: This coding strategies works differently from the above described one. Local recording generalizes attribute values in cells. Because of that

this strategies doesn't over generalize the table and the data distortion is significantly lower [7].

3 Underlying Barriers

In the following section, we will show the basic and most challenging barriers to the implementation of k-Anonymity. First, we will show the barrier which appears if you k-anonymize the data, the so-called **distortion** of data, in some papers it also mentioned as data loss.

3.1 Distortion of data as Barrier

A basic underlying barrier of k-anonymity is, how to measure if a implementation has been successful or leads to a satisfying result. This can be measured by a simple calculation. The **modification rate** is representing the fraction of cells which got modified within the attribute set of the quasi-identifier [7].

Table 3. a: original table,b: example for local recording, c: example for domain generalization

Gender	Birthday	Problem
male	13.08.1962	stress
male	28.10.1967	obesity
male	20.01.1977	stress
female	15.09.1973	obesity
female	15.03.1985	stress
female	28.05.1986	obesity

Gender	Birthday	Problem
male	13.08.1962	stress
male	28.10.1967	obesity
*	197*	stress
*	197*	obesity
female	15.03.1985	stress
female	28.05.1986	obesity

Gender	Birthday	Problem
*	196*	stress
*	196*	obesity
*	197*	stress
*	197*	obesity
*	198*	stress
*	198*	obesity

Example: for table2 a, the modification rate is $33,\overline{33}\%$ (4 out of 12 quasi-identifier got changed) for table 2c: its is 100% (12 out of 12 quasi-identifier got changed). Like this simple example shows the modification rate calculation is a unsatisfying procedure. Because of that the **weighted hierarchical distance** got introduced by Li, Wong, Fu and Pei. To calculate the **weighted hierarchical distance** of a cell, which got generalized from level p to level q, following formula is used [7].

$$WHD(p,q) = \frac{\sum_{j=q+1}^{p} \omega_{j,j-1}}{\sum_{j=2}^{h} \omega_{j,j-1}} \text{[7]}$$

Let the hierarchy of birth date be {D/M/Y, M/Y, Y, 10Y, C/T/G/P, *}. Where D/M/Y would be day.month.year, 10Y a 10 years interval and C/T/G/R for Child/Teen/Grownup/Pensioner.

Example with uniformed weight $w_{j,j-1} = 1$ **where** $2 \leq j \leq h$ [7]: For the above example Birthday gets generalized from D/M/Y to 10Y, which corresponds into $WHD_{Birthday}(6,3) = \frac{3}{5} = 0,6$. For the Gender generalization it

would be $WHD_{gender}(2,1) = \frac{1}{1} = 1$. Which means for generalize 5 cells of age from D/M/Y to 10Y one will have the same data distortion as if 3 cells of gender gets generalized from Male/Female to *. This calculation shows a much better way to address the distortion of data than the **modification rate** but it does not take how near a generalization is to the root (which would be *).

Example with height weight: $w_{j,j-1} = 1/(j-1)^{\beta}$ **where** $\mathbf{2 \leq j \leq h}$ **and** $\beta = \mathbb{R} \geq \mathbf{1}$ [7]: β would be chosen by the user. For example $\beta = 1$. For $WHD_{Birthday}(6,3) = \frac{0,\overline{33}+0,25+0,20}{1+0.5+0,\overline{33}+0,25+0,20} \sim 0,3431$. For $WHD_{gender}(2,1) = \frac{1}{1} = 1$. The distortion of nearly 3 changed cell of birthday from D/M/Y to 10Y have the same amount as if one cell of gender, from Female/Male to *, gets generalized.

Conclusion Because research need the information out of the tables, like of the examples. Its very important that as less as necessary information gets lost during the anonymization process. To show the importance of this an additionally example, consider a table with surviver of a **idiom disaster beyond all expectations**. Researchers trying to find out the long-time effects of this disaster. Thats why the want to find out if victims get more likely to life a long and happy life if the live far away or close to the disasters location. If the data gets to much generalized by location its maybe useless for researchers to work with.

3.2 NP Hard

3.3 Attacks as Barrier

Furthermore, also attacks have to be considered as barriers for the implementation, because if the implementation ignores the weaknesses which the attacks use, k-anonymity will be useless. 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 sources [2]. Unfortunately like Dwork showed 2006 that such safety is impossible because of impossibility of predict what the attacker may know. For example, if the adversary knows that Bob get paid twice as the average German man and the attacker got access to a database which publishes the average income by German men. The anonymity of Bob is compromised even if Bob's data is not in the database [5]. Therefore its important and necessary that the implementation takes possible attacks into account and implement countermeasures, but because attacks are not the main part of this paper it will be only a short introduction.

Homogeneity attack As an example, let Alice be the adversary and let be Bob her target. They are neighbors, some day Bob get transported with an ambulance to an hospital. Assume the hospital published the table **??**, where all current patients with them Nationality, Age, ZIP, and Problem are listed,

but this table got 4-anonymized before release. Alice knows that Bob is a 31 old, American who lives in ZIP Code 02239. She can conclude that either he is entry 3, 5,6, or 11. Furthermore, all of these entry have the same Problem, Cancer. Alice can conclude Bob is suffering from Cancer even if the table the table got 4-anonymized [11, 8]. To counter such attacker **diversity** is needed [8]. Such method is the so-called l-diversity which will not addressed further in this paper.

Table 4. My caption

	Nationality	Age	ZIP	Problem	Nationality	Age	ZIP	Problem
1	American	42	02135	Viral Infect	*	≥40	021**	Viral Infect
2	Japanese	41	02133	Hearth disease	*	≥40	021**	Hearth disease
3	Germany	38	02238	Hearth disease	*	3*	0223*	Cancer
4	Japanese	29	02139	Fever	*	≤30	021**	Fever
5	Indina	37	02232	Viral Infection	*	3*	0223*	Cancer
6	Native-american	34	02236	Cancer	*	3*	0223*	Cancer
7	Russia	53	02138	Viral Infection	*	≥40	021**	Viral Infection
8	China	23	02139	Cancer	*	≤30	021**	Cancer
9	American	23	02141	Short of breath	*	≤30	021**	Short of breath
10	Indian	46	02139	Viral Infection	*	≥40	021**	Viral Infection
11	American	31	02239	Vomiting	*	3*	0223*	Cancer
12	American	28	02130	Viral Infection	*	≤30	021**	Viral Infection

Background Knowledge Attack This attack use the demographic background knowledge, which got explained in the basics, of an adversary. Assume Alice have a college, which get also to the same hospital. This college is 32 years old, Japanese and have the ZIP 93607. Everyone with the same quasi-identifiers (Age = 3* and ZIP = 936**) have a cancer or a hearth disease. Because she knows that Japanese have a very low risk of a hearth disease she conclude her college has cancer [8].

Table 5. Background Knowledge Attack

	ZIP Code	Age	Disease	ZIP Code	Age	Disease
1	93677	29	Liver Disease	936**	≤30	Liver Disease
2	93602	22	Liver Disease	936**	≤30	Liver Disease
3	93909	52	Cancer	9390*	≥40	Cancer
4	93906	47	Flu	9390*	≥40	Flu
5	93673	36	Hearth Disease	936**	3*	Hearth Disease
6	93607	32	Cancer	936**	3*	Cancer

Unsorted matching attack against k-anonymity This attacks is based on the very common strategy to release 2 tables separately

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].

Table 6. My caption

Age	ZIP	Age	ZIP	Age	ZIP
42	91058	*	91058	42	91050
44	91058	*	91058	44	91050
50	27785	*	27785	50	27780
52	27785	*	27785	52	27780
20	32105	*	32105	20	32100
21	32105	*	32105	21	32100
31	67676	*	67676	31	67670
32	67676	*	67676	32	67670

4 Algorithm

This section will show some algorithms which goals is to archive k-anonymity through generalization.

4.1 The KACA Algorithm

This algorithm idea is to archive k-anonymity by clustering attribute hierarchical structures. The algorithm choose a random equivalent class, which is smaller than k. The next step is to form a lager equivalent class by merging the chosen one with the closest equivalent class. Which is resulting in a larger combined equivalent class. Through repeating this process the final result is that each

equivalent class consists of at least k tuples [7].

Algorithm 1: K-Anonymization by Clustering in Attribute hierarchies (KACA) [7]

```
1 form equivalence classes from the data set
2 while there exists an equivalence class of size < k do
3 |   randomly choose an equivalence class C of size k k
4 |   evaluate the pairwise distance of C and all other equivalence classes
5 |   find the equivalence class C' with the smallest distance to C
6 |   generalise the equivalence classes C and C'
7 end
```

This algorithm has a runtime of $O(nlogn + |E|^2)$. Li, Wong, Fu, and Pei have shown that their KACA-Algorithm is resulting in a 5.57 times smaller amount of distortion as the well known Incognito algorithm. The reason is lying in the technique which Incognito is using. Its a global recoding algorithm, which is resulting in a over-generalized table [7].

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. Domingo-Ferrer, J., Torra, V.: A critique of k-anonymity and some of its enhancements. In: Availability, Reliability and Security, 2008. ARES 08. Third International Conference on. pp. 990–993. IEEE (2008)
5. Dwork, C.: Differential privacy. In: Encyclopedia of Cryptography and Security, pp. 338–340. Springer (2011)
6. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: Efficient full-domain k-anonymity. In: Proceedings of the 2005 ACM SIGMOD international conference on Management of data. pp. 49–60. ACM (2005)
7. Li, J., Wong, R.C.W., Fu, A.W.C., Pei, J.: Achieving k-anonymity by clustering in attribute hierarchical structures. In: International Conference on Data Warehousing and Knowledge Discovery. pp. 405–416. Springer (2006)
8. 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)
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. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05), 571–588 (2002)
11. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05), 557–570 (2002)