Anonymity: A Formalization of Privacy - ℓ-Diversity

Michael Kern
Betreuer: Ralph Holz
Seminar Future Internet SS2013
Lehrstuhl Netzarchitekturen und Netzdienste
Fakultät für Informatik, Technische Universität München
Email: kernm@in.tum.de

ABSTRACT

Anonymization of published microdata has become a very important topic nowadays. The major difficulty is to publish data of individuals in a manner that the released table both provides enough information to the public and prevents disclosure of sensitive information. Therefore, several authors proposed definitions of privacy to get anonymous microdata. One definition is called k-Anonymity and states that every individual in one generalized block is indistinguishable from at least k - 1 other individuals. ℓ-Diversity uses a stronger privacy definition and claims that every generalized block has to contain at least ℓ different sensitive values. Another definition is called t-Closeness. It demands that the distribution of one sensitive value of a generalized block is close to its distribution in the entire table.

This paper mainly deals with the principle and notion of ℓ-Diversity. Therefore, two methods called Homogeneity and Background-Knowledge Attack are discussed to break the privacy constraints of k-Anonymity. Then a model to reason about privacy in microdata, namely Bayes-Optimal Privacy, is introduced. Based on k-Anonymity and Bayes-Optimal Privacy the principle and several instantiations of ℓ-Diversity are discussed. At the end ℓ-Diversity is applied to a real database gathered from several Android devices.

Keywords

anonymization, privacy, l-diversity, bayes-optimal privacy

1. INTRODUCTION

Many companies collect a lot of personal data of their costumers, clients or patients in huge tables. These tables often contain sensitive information about individuals like medications and diseases, income or customer data. In many cases it is useful to provide this in form of microdata (non-aggregated information per individual) to certain industries and organizations for research or analysis reasons. For that purpose companies often use suppression of identifiers like name and surname to provide a kind of anonymization of these tables. As Sweeney [10] shows in her paper, adversaries can disclose sensitive information of people with the aid of combining so called **quasi-identifiers** [10]. These are attributes like zip code, age or gender that are auxiliary for an adversary in combination with his background knowledge to reveal sensitive information of an individual. If a certain quasi-identifier (like zip code) exists both in the published microdata (containing sensitive information of one individual) and in an external database (containing the individual's name), the two datasets can be combined to get the name and the sensitive information of one individual to the corresponding zip code. This method is called **Linking Attack** [10]. For example, connecting medical data with the records of voter registration of Massachusetts led to a disclosure of medical information about the governor of Massachusetts [8]. Hence, it is inevitable to hide sensitive information from adversaries, so that certain individuals cannot be uniquely identified in published tables.

	Non-Sensitive		Sensitive
	Age	Zip Code	Medication
1	32	75235	Tamoxifen
2	49	75392	Tamoxifen
3	67	75278	Captopril
4	70	75310	Synthroid
5	54	75298	Pepcid
6	72	75243	Synthroid
7	56	75387	Tamoxifen
8	76	75355	Pepcid
9	40	75221	Erythropoietin
10	61	75391	Pepcid
11	63	75215	Synthroid
12	34	75308	Tamoxifen

Table 1: non-anonymized table

Sweeney [10] introduced **k-Anonymity**, a special definition of privacy, to enhance the anonymization of microdata. Here a published table is called k-anonymous, if every data tuple is indistinguishable from at least k - 1 other data tuples in relation to every set of quasi-identifiers. This constraint guarantees that individuals cannot be uniquely identified by using the earlier mentioned Linking Attacks.

Example 1: *Table 1 is a non-anonymized table from an imaginary hospital, collecting sensitive medication data of its patients. Identifier attributes like name and surname are removed and only age and zip code are considered to be non-sensitive and published. If an adversary knows the exact age and zip code of the individual, it is highly probable that this individual can be uniquely identified and medication is revealed.*

The next chapter deals with the disadvantages of k-Anonymity and shows two attacks to easily reveal sensitive information of such a k-anonymous table. The third chapter introduces an ideal definition of privacy, called **Bayes-Optimal Privacy** on which ℓ-Diversity is based on. Then

doi: 10.2313/NET-2013-08-1_07

	Quasi-Identifier		Sensitive
	Age	Zip Code	Medication
1	<60	752**	Tamoxifen
9	<60	752**	Erythropoietin
5	<60	752**	Pepcid
3	>=60	752**	Captopril
6	>=60	752**	Synthroid
11	>=60	752**	Synthroid
2	<60	753**	Tamoxifen
12	<60	753**	Tamoxifen
7	<60	753**	Tamoxifen
4	>=60	753**	Synthroid
8	>=60	753**	Pepcid
10	>=60	753**	Pepcid

Table 2: 3-anonymous table

ℓ-Diversity is introduced and its advantages and several instantiations are explained. At the end anonymization with ℓ-Diversity is applied to a huge dataset containing pieces of information (like device and model name, GPS location data) of several Android devices.

2. ATTACKS ON K-ANONYMITY

As mentioned in the previous section, k-Anonymity is one possible method to protect against Linking Attacks. But the definition of privacy in k-Anonymity is vulnerable. It can be easily shown that the condition of k indistinguishable records per quasi-identifier group is not sufficient to hide sensitive information from adversaries. In the following two simple attacks on k-anonymous datasets are discussed which easily reveal sensitive information.

Example 2. *Table 2 shows a 3-anonymous table. Here three records of each group are put into a new block containing the same quasi-identifiers 'Age' and 'Zip Code'. These quasi-identifiers are generalized to new values, where age is divided into two intervals '>=60' and '<60', and the first same three digits of the zip code are published, whereas the last two significant digits are hidden by '*'. It can be seen that every record is indistinguishable from the other two data tuples in its group. Therefore, this table satisfies the definition of 3-Anomymity and prevents Linking Attacks on this dataset.*

2.1 Homogeneity Attack

One attack is called **Homogeneity Attack**. Let us assume that Eve is the adversary. Her neighbor and good friend Alice was taken to hospital two weeks ago and she's anxious about what kind of disease she's suffering from. She discovers the generalized table 2 on the internet and looks at the quasi-identifiers of the table. As Alice lives nearby her, she knows the exact zip code '75392' of her town. Furthermore she remembers that Alice is younger than 60 years. Thus, she comes to the conclusion that her medication lies in records 2,7 and 12. Since there's only one sensitive value 'Tamoxifen' within this group, it is quite likely that Alice has breast cancer, because this medication is often used to treat this kind of disease. The example shows that the lack of diversity of sensitive attributes in a generalized group can lead to an unintentional disclosure. So, the aim of privacy is not fulfilled in this case.

2.2 Background-Knowledge Attack

Often times adversaries have certain background knowledge that can be to used to successfully eliminate possible values for sensitive attributes of a particular individual with very high probability. Assume Eve has a friend called Mario who was also taken to the same hospital as Alice. Hence, Mario's medication must be listed in the same table 2. As she knows that Mario is older than 60 years and comes from the neighbor town with zip code '75355', his sensitive value must be contained in record 4, 8 and 10. So, Mario has to take either 'Synthroid' or 'Pepcid'. Because Eve knows that Mario eats fish almost every day, it is very unlikely that he suffers from a certain thyroid disease, and has to take the drug 'Synthroid'. Thus, Mario takes the medication 'Pepcid', which implies that he must have a certain stomach disease. Regarding this example, k-Anonymity does not take into account the background-knowledge of adversaries. Eve needs just one additional information to eliminate one sensitive value and to reveal the medication of his friend Mario. Therefore, another formalization of privacy is needed to avoid both attacks and reach "optimal privacy".

3. BAYES-OPTIMAL PRIVACY

Before describing the principle of ℓ-Diversity, the idea of ideal privacy has to be discussed first on which it is based on. This idea is called **Bayes-Optimal Privacy** (introduced in [6]) that uses conditional probabilities to model the background knowledge of an adversary and to reason about privacy in a table.

3.1 Definitions

Several notations are mentioned in the following. Let the set $T = \{t_1, t_2, ..., t_n\}$ be a simple non-anonymized table like table 1. T is assumed to be a partial quantity of some larger population Ω, where t_i denotes the i^{th} row of T, and its columns are termed attributes A_i as a subset of all possible attributes $A = \{A_1, A_2, ..., A_m\}$. Every attribute A_i itself has several varying values $\{v_1, v_2, ..., v_n\}$. Then $t_i[A_j] = v_i$ is the value v_i of attribute A_j of the i^{th} individual.

Example 3: *Table 1 with $T = \{t_1, t_2, ..., t_{12}\}$ is a fictional subset of the population of the United States Ω. The set of attributes A is $\{'Age', 'ZipCode', 'Medication'\}$. For example, the value of $t_1['Age']$ is '32' and $t_2['Medication'] =$ 'Tamoxifen'.*

Furthermore the attributes are subdivided into non-sensitive and sensitive attributes. Every attribute, whose values have to be hidden from any adversary, is called **sensitive attribute**. Then S denotes the set of all possible sensitive attributes in a table T. Every attribute that is not called sensitive is termed **non-sensitive attribute**. In table 2 for example the attributes $\{'Age', 'ZipCode'\}$ are assumed to be non-sensitive. Attribute 'Medication' has to be protected from revealing by some adversaries and thus considered a sensitive attribute. The set of non-sensitive attributes are further refined in a subset Q labelled as a set of 'quasi-identifier', defined in section 2.1 in [6]:

Definition (Quasi-identifier) *A set Q of non-sensitive attributes $\{Q_1, ..., Q_w\}$ of a table is called a quasi-identifier if these attributes can be combined with external data to*

doi: 10.2313/NET-2013-08-1_07

uniquely identify at least one individual in the general population Ω.

As mentioned in the introduction, publishing table 1 induces the danger of disclosure of sensitive information of one individual by certain adversaries. Therefore, table T has to be anonymized. One possible method is called generalization, where every value of a certain quasi-identifier is replaced by a more general value (i.e. the value of attribute 'Age' of all persons that are younger than 50 years can be generalized to '< 50'). Hence, $T \rightarrow^* T^*$ denotes the generalization of table T to T^*, and $t \rightarrow^* t^*$ means data tuple t is generalized into the data tuple t^*. An anonymized table is denoted $T^* = \{t_1^*, t_2^*, ..., t_n^*\}$ consisting of attribute values q^*, generalized from the set of quasi-identifiers Q.
With all these defnitions the probability of belief of one adversary is modeled in the next chapter.

3.2 Probability of Belief

Every adversary has a different level of background knowledge that can be used to reveal sensitive information. Because one company is not able to possess all different levels of knowledge, it is necessary to describe such an adversary's knowledge mathematically.
It is assumed that every adversary has the maximum possible knowledge. Considering the Example 1, where Eve wants to find the sensitive value corresponding to Alice, she knows the complete joint frequency distribution f of sensitive attribute S, conditioned on the non-sensitive quasi-identifiers Q (for example, she knows the frequency of heart diseases of people being older than 60 years in the United States). Furthermore she knows all quasi-identifier values q of Alice. So, she knows that $t_{Alice}[Q] = q$, and wants to discover her sensitive value $t_{Alice}[S] = s$. Her belief of Alice's sensitive value being s, given that q is her non-sensitive value, is classified in [6] into **prior-belief** and **posterior belief**. The prior-belief is just Eve's background knowledge

$$\alpha_{(q,s)} = P_f(t[S] = S|t[Q] = q)$$

which denotes the probability that Alice's sensitive value must be s on condition that her non-sensitive value is q.
Now Eve encounters the anonymized table T* published from the hospital. After analyzing the records of the table, her belief changes to a posterior-belief:

$$\beta_{(q,s,T*)} = P_f(t[S] = s|t[Q] = q \wedge \exists t^* \in T^*, t \rightarrow^* t^*)$$

This posterior-belief can be put into a mathematical formula, whose derivation and proof can be found in theorem 3.1 in [6] (technical report):

Theorem 3.2 *Let q be a value of the non-sensitive attribute Q in the base table T; let* q^* *be the generalized value of q in the published table* T^**; let s be a possible value of the sensitive attribute; let* $n_{(q^*,s')}$ *be the number of data tuples* $t^* \in T^*$ *where* $t^*[Q] = q^*$ *and* $t^*[S] = s'$ *and let* $f(s'|q^*)$ *be the conditional probability of the sensitive value conditioned on the fact that the non-sensitive attribute Q can be generalized to* q^**. Then the following relationship holds:*

$$\beta_{(q,s,T*)} = \frac{n_{(q^*,s)} \frac{f(s|q)}{f(s|q^*)}}{\sum_{s' \in S} n_{(q^*,s')} \frac{f(s'|q)}{f(s'|q^*)}}$$

Theorem 3.2 takes into account both the counts of one sensitive value proportional to all sensitive values in a q^*-block and the frequency distribution f of one sensitive value compared to all possible sensitive values in a certain population. It is also useful to measure the quality of the privacy.

3.3 Privacy Principle

When talking about privacy there are two different possibilities of revealing sensitive information.
Positive disclosure denotes that an adversary can correctly identify the sensitive value of one individual with very high probability. Consider the homogeneity attack in section 2.1 where Eve could be sure that Alice has breast cancer. Hence, after observing the published table, her posterior-belief has become very high (here $\beta_{(q,s,T*)} \rightarrow 1$).
In contrast to that, the process of correctly eliminating possible sensitive values for one individual with very high probability is called **negative disclosure**. This takes place, if the posterior belief becomes very small (or $\beta_{(q,s,T*)} \rightarrow 0$). Regarding section 2.2 again Eve could successfully eliminate the possible sensitive value 'Thyroid' using her very good background-knowledge.
The ideal principle of privacy is that prior and posterior-belief of one adversary should not differ very much from each other after observing the published table. For example, Eve's prior belief that Alice has the sensitive value s, if her non-sensitive value is q, is assumed to be about 50 percent. After considering the generalized table T^* it raises to nearly 100 percent, because there's only one possible candidate. Then positive disclosure takes place and the sensitive information could be correctly revealed.

One possible measurement of privacy in a certain table is the difference between prior and posterior belief. This can be modeled by using and defining boundaries for each of the two beliefs. Here, the privacy of one table is violated, if the prior belief is below its upper boundary and the posterior belief exceeds its lower boundary, which implies that the difference of the two beliefs is too high and the adversary can infer positive or negative disclosure (explained in section 3.2 in [6]).

Although Bayes-Optimal Privacy is a good definition to gain optimal privacy, it has some disadvantages to overcome. First of all it is very likely that the company which publishes a table does not know the complete distribution of all sensitive and non-sensitive attributes over the general population Ω. Then it is even more unlikely that the publisher knows the adversary's level of knowledge. Third there are instances of knowledge that even cannot be described mathematically (regarding section 2.2), when Mario told Eve that he eats fish almost every day. And there are always more then one adversary. Each of them has a different level of background-knowledge, which a publisher cannot handle as well.

4. L-DIVERSITY

In order to eliminate the above-mentionend disadvantages of Bayes-Optimal Privacy, the principle and basic notion of ℓ-Diversity is described. Then, two definitions of this principle for realization in practice are introduced and at last the advantages and disadvantages of ℓ-Diversity are discussed.

	Non-Sensitive		Sensitive
	Age	Zip Code	Medication
1	<60	75***	Tamoxifen
7	<60	75***	Tamoxifen
5	<60	75***	Pepcid
2	<60	75***	Tamoxifen
12	<60	75***	Tamoxifen
9	<60	75***	Erythropoietin
3	>=60	75***	Captopril
6	>=60	75***	Synthroid
11	>=60	75***	Synthroid
4	>=60	75***	Synthroid
8	>=60	75***	Pepcid
10	>=60	75***	Pepcid

Table 3: 3-diverse table

4.1 Principle of l-Diversity

The principle of ℓ-Diversity is based on the theorem 3.2 for posterior-belief. For that reason, observe table 2 marked by T^*. This table is subdivided into several q^*-blocks, whose non-sensitive attributes q are generalized to q^*. Regarding the dataset of T^*, data tuples 1, 2 and 6 are one q^*-block, where attributes 'Age' and 'Zip Code' are generalized to '<60' and '75***'.
In order to infer positive disclosure the number of occurrences of the sensitive value s must be much higher than the counts of all other sensitive values. Or the frequency distribution of all other sensitive values not including s in a certain population is very small. Thus, the theorem 3.2 can be rearranged as follows:

$$\exists s, \forall s' \neq s : n_{(q^*,s')} \frac{f(s'|q)}{f(s'|q^*)} \ll n_{(q^*,s)} \frac{f(s|q)}{f(s|q^*)}$$

This means that the probability of every other sensitive value s' is much lower than the likelihood of Alice's probable sensitive value candidate s. Thus, it is very unlikely that Alice's sensitive value must be s'. So, Eve can successfully determine the correct sensitive value s of Alice with high probability. This event takes place in two cases: lack of diversity and very good background knowledge.
Lack of diversity occurs when there's nearly one sensitive value s in this block. This means the number $n_{(q^*,s)}$ of data tuples for s in the q^*-block is much higher than the counts $n_{(q^*,s')}$ of all the block's other sensitive values s'.
With **Strong Background Knowledge** an adversary can often eliminate possible sensitive value of one individual with very high probability by knowing the frequency distribution $f(s'|q)$ of sensitive values s' in a certain population Ω. For example, the frequency distribution f of 'breast cancer' for men is low in general, as it is very improbable that men have this disease.
In order to avoid these two privacy-destroying cases, every q^*-block should have at least ℓ different sensitive attributes, so that an adversary must have at least $\ell-1$ different amount of information to eliminate the other possible values with high probability. Thus, the following principle is used to define ℓ-Diversity in [6]:

ℓ-Diversity Principle *A q^*-block is ℓ-diverse if contains at least ℓ "well-represented" values for the sensitive attribute S. A table is ℓ-diverse if every q^*-block is ℓ-diverse.*

Example 4: *Consider the table 3. Here, the records are grouped into two q^*-blocks whose non-sensitive attributes are generalized. Every block contains 6 indistinguishable individuals and three different values for the sensitive attribute 'Medication' (e.g. the first q^*-block contains 'Tamoxifen', 'Pepcid' and 'Erythropoietin'). Such a table is called 3-diverse, as every adversary who wants to reveal sensitive information of one individual needs to have at least $\ell-1=2$ pieces of information to eliminate the "wrong" sensitive values and to identify the "correct" one. Regarding the example of the background-knowledge attack (section 2.2), Eve can successfully determine that Marco cannot take the medication 'Synthroid', but she still has to find out if Marco takes 'Captopril' or 'Pepcid'. Hence, she needs one additional information to gain a positive disclosure.*

This example shows that ℓ-Diversity takes into account every level of background-knowledge of any adversary. Therefore, the publisher can control the amount of protection that is given by an ℓ-diverse table only by modifying the parameter ℓ to the desired level without knowing the background-knowledge level of all adversaries.
Several instantiations are introduced in the next section that can be used to define the "well-representation" of sensitive attributes in a practical manner.

4.2 Realization of l-Diversity

One realization of ℓ-Diversity is called Entropy ℓ-Diversity. It uses the definition of entropy in information theory to quantify the uncertainty of possible sensitive values [7]. The following condition states that every q^*-block has not less than ℓ different and nearly "well-represented" sensitive values:

Entropy ℓ-Diversity: [6]
A table is Entropy ℓ-diverse if for every q^*-block

$$H_\ell = -\sum_{s \in S} p_{(q^*,s) log(p_{(q^*,s)})} \geq log(\ell)$$

where $p_{(q^*,s)} = \frac{n_{(q^*,s)}}{\sum_{s' \in S} n_{(q^*,s')}}$ is the fraction of tuples in the q^*-block with sensitive attribute value equal to s.

Using the notion of entropy, the higher the value of H_ℓ is, the more pieces of information are needed to infer positive disclosure. For example, consider the case that there's only one possible sensitive value s in a certain q^*-block. Then $p_{(q^*,s)} = 1$ and $p_{(q^*,s')} = 0, \forall s' \in S, s' \neq s$. This yields $H_\ell = 0$, which means there's no information needed to determine the possible sensitive value, as there is only one given in the q^*-block. Because the maximal value of entropy $H_\ell = log(\ell)$ is only achieved if $p_{(q^*,s')}$ is equal for at least ℓ existing sensitive values s' in the block, the entropy of the whole table must be greater or equal $log(\ell)$.

Example 5: *Applying this definition to table 3, the two entropies of the two blocks are calculated. The first block yields an entropy $H_1 = -(\frac{4}{6} \cdot \log(\frac{4}{6}) + 2 \cdot \frac{1}{6} \cdot \log(\frac{1}{6})) \approx 0.378$, and the second block results in $H_2 = -(\frac{3}{6} \cdot \log(\frac{3}{6}) + \frac{2}{6} \cdot \log(\frac{2}{6}) + \frac{1}{6} \cdot \log(\frac{1}{6})) \approx 0.439$. In order to fulfill the condition every entropy of each q^*-block has to be at least $log(\ell)$. So, the minimum entropy H_1 of the table has to be chosen to quantify ℓ. In this case $H_1 \geq \log(\ell) \Leftrightarrow 10^{H_1} \approx 2.387 = \ell$. Thus, table 3*

doi: 10.2313/NET-2013-08-1_07

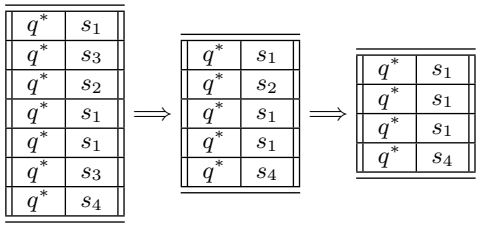

Table 4: q^*-block of a fictional anonymization table

is at least 2.3-diverse, which states that every block contains at least two different "well-represented" sensitive values.

It can be easily seen that Entropy ℓ-Diversity is very restrictive and hard to achieve. Consider the first q^*-block of a fictional table with the same sensitive attribute as table 3. Is it assumed that the medication value 'none' is listed as well, which indicates that the patient is good in health again. Furthermore let the number of records 'none' be much more higher than the counts of all other sensitive values. Then Entropy ℓ-Diversity cannot be satisfied by such a table, as the probability of value 'none' is too high compared to the likelihood of all other sensitive values.
Because most of the patients are already healthy again, the hospital does not need to bother about positive disclosure of the medication value 'none', as this information cannot be misused by an adversary.
Therefore, another definition definition is used to resolve this problem, which is called Recursive (c,ℓ)-Diversity. Here one q^*-block of an anonymized table contains $\{s_1, s_2, ..., s_m\} \in S$ possible sensitive values. Their frequencies $n_{(q^*,s_i)}$ (number of data tuples within the block) are put into the set of the overall frequencies $\{n_1, n_2, ..., n_m\}$ sorted in descending order. So n_1 means the frequency of the most frequent sensitive value in the q^*-block, n_2 the second most frequent value and so on. It is assumed that the adversary needs to eliminate $\ell - 1$ different sensitive values to gain positive disclosure (with some sensitive values being allowed to reveal or $\ell \leq m - 1$). So, in order to prevent positive disclosure, the most frequent sensitive value should not exist too often in a table. This is satisfied, if the following definition (introduced in [6]) holds:

Recursive $(\mathbf{c}, \ell)$-Diversity: *In a given q^*-block, let n_i denote the number of times the i^{th} most frequent sensitive value appears in that q^*-block. Given a constant c, the q^*-block satisfies Recursive (c, ℓ)-Diversity if $n_1 < c(n_\ell + n_{\ell+1} + ... + n_m)$. A table T^* satisfies Recursive (c, ℓ)-diversity if every q^*-block satisfies Recursive (c, ℓ)-Diversity. 1-Diversity is assumed to be always fulfilled.*

The constant c is defined manually by the user and can be used to determine, how often the most frequent sensitive value may occur in relation to the total amount of the other sensitive attribute values. **Recursive** in this definition states that, if any sensitive value s' within a (c, ℓ)-diverse q^*-block is eliminated by an adversary, the remaining block (not regarding the tuples containing s') has to be at least $(c, \ell - 1)$-diverse.

Example 6: *Table 4 shows one q^*-block of a fictional table. Here the set of all possible sensitive values S is $\{s_1, s_2, s_3, s_4\}$, where every non-sensitive attribute is generalized to q^*. Then the set of all the sensitive value frequencies is $\{n_1 = 3, n_2 = 2, n_3 = 1, n_4 = 1\}$. It is assumed that the adversary has to eliminate at least $\ell = 3 - 1 = 2$ different sensitive values to infer positive disclosure. Applying the previous definition, let the constant c be 2. Then this block is (2,3)-diverse, if $n_1 < c(n_3 + n_4)$. It can be seen that this equation holds for $c = 2$, as $3 < 2 \cdot 2 = 4$. Now the second most frequent sensitive value s_3 is eliminated by the adversary. Then the resulting block has to be (2,2)-diverse, or more respectively the equation $n_1 < 2(n_2 + n_3)$ has to be satisfied. Regarding the table in the middle of table 4, this is also fulfilled, as $n_1 = 3 < 2 \cdot (1 + 1) = 4$. It can be easily recalculated that (2,2)-diversity holds, if any other sensitive value is eliminated first instead of s_3. After removing a second sensitive value (compare the right of table 4), it has to be examined, if this remaining block is (2,1)-diversity. As this is always satisfied by definition, this q^*-block can be considered Recursive (2,3)-diverse.*

In some cases a company wants to release not only one but multiple sensitive attributes when publishing an anonymized table which provides a certain level of ℓ-Diversity. Like [6] shows, if multiple sensitive attributes are treated and tested separately against ℓ-Diversity, it is not guaranteed that this table satisfies ℓ-Diversity for all sensitive attributes as well. Using the other (non-generalized) sensitive attributes the privacy definition of one single sensitive attribute can be broken and Linking Attacks are possible. Therefore, for every sensitive attribute all other sensitive attributes have to be treated as quasi-identifiers, as well.

4.3 Discussion

The principle of ℓ-Diversity avoids the disadvantages that arise with Bayes-Optimal Privacy. When using the definition of ℓ-Diversity, a publisher does not require knowledge of the full distribution of sensitive and non-sensitive attributes in any population. Furthermore, the publisher does not have to know the level of any adversary's knowledge, as he can decide with the parameter ℓ how many pieces of knowledge the adversary needs to gain full positive disclosure.
But Li et al. [5] show that the principle of ℓ-Diversity is not sufficient to avoid sensitive attribute disclosure. He mentions two attacks that can break the privacy definition of this principle and reveal sensitive information of individuals. Imagine that a sensitive value in the ℓ-diverse table is extremely frequent, whereas the sensitive value is very unlikely in the whole population. Then **Skewness Attack** is possible and implies that it is very likely for a certain person which is associated to this table to have this sensitive value, because most of the individuals have this same and seldom sensitive value, as well. Another attack is called **Similarity Attack**. Consider a q^*-block that contains ℓ diverse possible sensitive values that all depict a special kind of heart disease. Then the adversary can infer positive disclosure if he can assign an individual to this q^*-block. In order to avoid such attacks Li [5] introduces the principle of **t-Closeness**, where the distribution of a sensitive value in any q^*-block should be close to the distribution of the value in the entire table.
After all this theory and privacy definitions, it is interesting to know, how the principle of ℓ-Diversity can be applied to real existent databases.

device	model	version	network	latitude
leo	HTC HD2	10	Vodafone.de	49.265111
lgp970	LG-P970	8	movistar	40.475134
crespo	Nexus S	10	Swisscom	47.430045
vision	HTC Vision	8	vodafone UK	50.872790
vision	HTC Vision	8	vodafone UK	50.872688

Table 5: Excerpt from the dataset of the Android geodata

5. ANONYMIZATION IN PRACTICE

Several algorithms (like [9], [4]) have been introduced so far to realize k-Anonymity, ℓ-Diversity or t-Closeness [5] in an efficient manner. Hence, it is useful to develop toolboxes that provide these algorithms, and that can be applied to any arbitrary published dataset. Therefore, two universities developed such toolboxes which are briefly introduced in the following section.

5.1 Anonymization Toolboxes

One toolbox is called **Cornell Anonymization Toolkit (CAT)** [12] and was developed by the Department of Science at Cornell University. It is a Windows-based software containing an interactive GUI for visualization and analyzing of (anonymous) databases. For anonymization it uses the definitions of Recursive (c, ℓ)-Diversity and t-Closeness [5].
Another toolbox was created by the University of Dallas (UTD) and is called **UTD Anonymization Toolbox** [1]. It is a platform-independent software for anonymization of random datasets. Here nearly every privacy method (including k-Anonymity, ℓ-Diversity and t-Closeness) is implemented using algorithms like Datafly [9] and Incognito [5]. Both tools require databases in form of text files as input and certain hierarchy trees like value generalization trees ([10]) of non-sensitive attributes or quasi-identifiers to apply the implemented algorithms.

5.2 Android Geodata

In the context of the Bachelor Thesis of Wagner [11], several kinds of data from Android-based devices (like smart-phones and tablets) is gathered via an Android application in order to analyze the user's behavior and the general network structure. Information like name of model, name of device, Android sdk version, network provider and exact GPS locations (in latitude and longitude angles) are collected and stored in json-files each device with different datasets per timestamp. Every device gets its own unique 'device id' within the whole dataset. The entire dataset has overall 166060 records composed of 989 distinct devices with averaged 169 different datasets per device. The dataset itself is transformed into a sqlite database consisting of the attributes {'device id', 'model', 'device', 'version', 'network', 'longitude', 'latitude'}.
Table 5 shows an excerpt from the sqlite database generated from the Android datasets. Imagine that Eve is an adversary and has one friend Tom that takes part in Wagner's study. It is assumed that she knows the name of Tom's device called 'lpg970'. As this device name is unique in the whole dataset, she can correctly determine her friend's longitude and latitude angles. With this information Eve is able to look up, where her friend is located currently or was situated in the past. Hence, this table must not be released in its raw form.
In order to publish such a table to any research group or the public, it has to be anonymized to make it difficult for any adversary to disclose sensitive information.

5.3 Generalization Process

The method of generalization is used for anonymization. First it has to be figured out which set of quasi-identifier attributes Q are auxiliary for an adversary to reveal sensitive information of an individual, and which published attributes have to be considered sensitive. As [3] shows, an adversary can easily identify individuals by knowing only one attribute value q of Q = {'model', 'device', 'version', 'network'}. For example, 56 individuals can be identified by their unique 'device' value. All the worse, sensitive information of 214 individuals can be revealed by the knowledge of all four attributes in Q. For that reason, every attribute $q \in Q$ has to be taken as quasi-identifiers for generalization. Furthermore, it is interesting to know, where the users come from and which path route they took within a certain time interval. So, the values of their exact world position should be published as well and regarded sensitive, as GPS location data linked to an unique individual can be misused by an adversary.

5.3.1 *Building Up Generalization Hierarchies*

The major challenge is to disguise the quasi-identifiers Q in such a manner that the resulting anonymous table both provides enough information to an observer and satisfies the privacy definition to prevent sensitive information disclosure. For example, the concealment 'GT-I*' of value 'GT-I9100' means to an expert that this user uses a smart-phone created by the producer SAMSUNG, but does not disclose which exact type of the smart-phone's family he actually uses.
For that purpose a hierarchy of generalization has to be created for Q with different levels of disguise. This is useful, as various generalization hierarchies can be combined to divide the table into different generalized q^*-blocks, and to achieve the desired privacy definition. Such an hierarchy is implemented as a tree, where the root denotes the highest level of generalization and every parent node denotes the generalization of all its child nodes.
One possible generalization of the device names is classifying the value alphabetically. For example, all values with first letter 'G' are generalized to the alphabetical range 'A-G'. Such an hierarchy does not make sense when providing this data to researcher groups. For statistical analysis no information can be acquired and gathered from such a hierarchy, as the value 'A-G' does not specify, what kind of device the participant has used.

```
SELECT COUNT(*) as cnt, device
FROM android_data
GROUP BY device ORDER BY cnt;
```

Listing 1: Show all occurrences of the device values

So, first the SQL-query in listing 1 is executed on the database to show all distinct values of device names and their counts in the full dataset in descending order, which can be seen in listing 2. It shows that the devices with sub-strings 'GT-*', 'GT-S' and 'GT-P' are very frequent in the database. So, 'GT-I*' belongs to the first generalization level G_1 of attribute 'device', and contains values with sub-string 'GT-I'.

doi: 10.2313/NET-2013-08-1_07

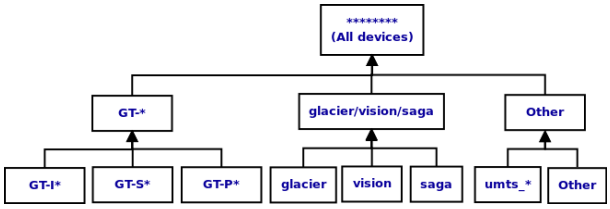

Figure 1: Generalization hierarchy tree of the non-sensitive attributes 'device'

The same is done for the values 'GT-S*' and 'GT-P*'. These generalizations can be generalized further to all devices with sub-string 'GT-' as part of the second generalization level G_2.

'Count'	'device'
128	glacier
108	GT–P1000
103	vision
102	GT–P1000L
57	saga
38	GT–I9000
36	umts_sholes
24	GT–S5570

Listing 2: List of some distinct device names with their counts in the android dataset

Furthermore the device values 'glacier', 'vision' and 'saga' are generalized to 'glacier/vision/saga' as these are all devices from the vendor HTC. Every other device which does not belong to the earlier mentioned generalizations is put into the category 'Other'. As the sub-string 'umts_*' is very common in the dataset as well, 'Other' is further divided into 'umts_*' and 'Other'.
The third and highest generalization level G_3 is 'All devices' and involves all the second generalization levels and thus all the device values in the entire dataset. This is the same as saying the attribute 'device' is completely suppressed.
Now the generalization hierarchy of 'device', which can be seen in figure 1, consists of three different levels: G_3 ={'All devices'}, G_2 ={'GT-*', 'glacier/vision/saga', 'Other'} and G_1 ={'GT-I*', 'GT-S*', 'GT-P*', 'glacier', 'vision', 'saga', 'umts_*', 'Other',}. In this context G_0 is composed of all distinct, not generalized values of 'device'.
The hierarchies for the remaining quasi-identifiers are created in the same manner. For example, the hierarchy for attribute 'model' looks very similar to the one in figure 1. Here the branch 'glacier/vision/saga' is replaced by the generalization of all 'HTC models', as these models are produced by the vendor HTC. This generalization is further divided into the set {'HTC Desire*', 'HTC Glacier', 'HTC Vision', 'other HTC models'}. As many participants use Android sdk 'version' 8 (Android version 2.2.x), 'All versions' is further refined into the set {'<=8', '>8'}, which is shown in figure 2. Last but not least, the attribute 'network' is generalized into the first generalization level G_1 = {'T-Mobile', 'Vodafone*', 'Telefonica', 'No network', 'Other network'} and the highest level G_2 = {'All networks'}.

5.3.2 *Suppression*

It is possible that a chosen generalization hierarchy in relation to an attribute is not good enough to satisfy a certain

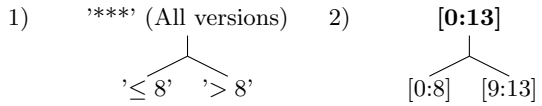

Figure 2: Generalization tree of attribute 'version' in 1) and its value generalization hierarchy in 2)

privacy definition. Therefore, suppression [9] is a method to overcome this problem. Here, all values of one attribute are not published at all. Instead, they are disguised by a simple string like '***'. So, the suppression is the same as applying the highest generalization level of the quasi-identifier (compare the top of the root in figure 1).

5.3.3 *Diversity of Sensitive Attributes*

In order to generate an ℓ-diverse table from the evaluated Android data set, diversity of the sensitive attributes longitude and latitude has to be defined. When regarding table 5, the pure degree of latitude itself is not well suited for diversity. There are sparsely populated regions, where villages or even houses are more than one kilometers apart from each other. For example, two latitude angles that differ in the fourth decimal place (the second of angle) can be assigned by an adversary to the same village or house, as the resulting positions are very close to each other
Therefore, the minute of angle is used to determine diversity. Two points that differ in one minute of latitude angle, are $\approx 1.83km$ away from each other and assumed to be diverse. As the angles have to be compared mathematically, they are transformed into the float format 'xxx.yy', where 'xxx' is the angle and 'yy' depicts the minute of angle ('21.45' means an angle of $21°45'$).
Now all requirements are fulfilled to start the anonymization of the Android database.

5.4 Anonymization with UTD Toolbox

The toolbox of the University of Dallas is used to perform anonymization of the Android database, applying the created hierarchies. This tool uses a **Value Generalization Hierarchy** (VGH, [9]) per quasi-identifier for its anonymization process. Here, every value of the quasi-identifier attribute is mapped onto a distinct integer number. Then, every next generalization level compromises a certain range of specified integer numbers (e.g. the range [x:y] covers the numbers from x to y). Consequently, the highest generalization level covers the entire number range.

Example 7: *Consider figure 2. Let the values of version be {none, 8, 9, 10, 11, 12, 13} and mapped onto the values {0, 8, 9, 10, 11, 12, 13}. Then the first generalization '≤ 8' covers the range [0:8] and '>8' contains the values in range [9:13]. Consequently 'All versions' compromises the entire range [0:13].*

Then, the anonymization process of the UTD toolbox is started to gain Entropy ℓ-Diversity using the efficient algorithm **Incognito** [4]. So, first the VGHs of all attributes in Q are generated. Then, every generalization level of one quasi-identifier is combined with every generalization level of all the other quasi-identifiers. For each of the different combinations the resulting table is checked against the de-

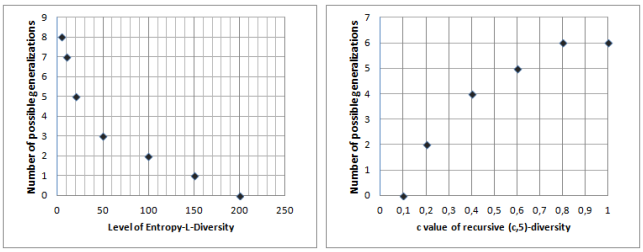

Figure 3: Number of possible generalizations, left: with different ℓ-Levels of Entropy ℓ-Diversity, right: with various constants c of Recursive $(c, 5)$-Diversity

sired privacy definition. When the anonymization process is finished all possible generalizations are listed and the best one is selected by the Incognito algorithm [4].
The toolbox is applied on the Android database to create an anonymous table containing the attributes {'model', 'device', 'version', 'network', 'latitude', 'longitude'}. It uses the privacy definitions Entropy ℓ-Diversity and Recursive (c, ℓ)-Diversity with different values for ℓ and c. Figure 3 shows the number of possible generalizations suggested by the toolbox for the set of ℓ values {5, 10, 20, 50, 100, 150, 200} on the left and different c values {0.1, 0.2, 0.4, 0.6, 0.8, 1.0} on the right. Listing 3 shows an Entropy 5-diverse table excerpt of the anonymous Android database, using the generalization levels $\{G_3, G_2, G_2, G_1\}$ (suggested by the toolbox) of Q = {'model', 'device', 'version', 'network'}. Here, the attributes 'model' and 'version' are completely suppressed to satisfy the privacy definition.

```
(model, device, version, network, latitude, longitude)
('*', 'GT-*', '*', 'No network', -15.45, -47.53)
('*', 'GT-*', '*', 'Other networks', 28.39, 77.11)
('*', 'GT-*', '*', 'T-Mobile', 52.21, 5.37)
('*', 'GT-*', '*', 'Telefonica', -15.45, -47.53)
('*', 'GT-*', '*', 'Vodafone***', 51.24, 8.35)
('*', 'Other devices', '*', 'Other networks', 51.39, -0.05)
('*', 'saga/vision/glacier', '*', 'No network', 50.52, -1.17)
('*', 'saga/vision/glacier', '*', 'Telefonica', 50.06, 14.28)
```

Listing 3: Excerpt of Entropy 5-diverse anonymized table generated by the UTD toolbox

6. RELATED WORK

Greschbach [2] utilizes the ℓ-Diversity definition to gain location privacy. When using Location Based Services (LBS) a provider may obtain GPS location data per time from a user's device, and is able to reconstruct a movement profile and by association the behavior of the user. This can be avoided if several dummies simulate the same device of the user and use the same LBS at the same time as the user's device. These dummies then fake different path routes to disguise the real path route of the user.
Zhou [13] extends the privacy definition of k-Anonymity and ℓ-Diversity from relational data to social network data in order to reach privacy in social networks.

7. CONCLUSION

This paper has presented reasons why a publisher should never publish microdata to researcher groups in its raw form, as adversaries can use their background knowledge or link quasi-identifier attributes with external databases to reveal sensitive information of certain individuals. K-Anonymity, ℓ-Diversity and t-Closeness are principles to gain a certain anonymity level, and to preserve privacy of individuals listed in the published microdata. Every principle has its own advantages and disadvantages that have to be considered when applying such principles to microdata. Chapter 5 shows that generalization in combination with the above-mentioned principles can be used to anonymize microdata. But this process is not trivial and has to be well thought out. It has to be considered which generalization makes sense, so that researcher groups or statistical analysts can work with the published tables. In this context, anonymization toolboxes like UTD Anonymization Toolbox can help to discover the most suitable anonymization for arbitrary datasets.

8. REFERENCES

[1] UTD Data Security and Privacy Lab. UT Dallas Anonymization Toolbox. http://cs.utdallas.edu/dspl/cgi-bin/toolbox/

[2] Benjamin Greschbach. *Location Privacy l-diversity durch realistische Dummies*. Studienarbeit, Albert-Ludwigs-Universität Freiburg, Freiburg, 2009.

[3] Janosch Maier. *Anonymity: Formalisation of Privacy - k-Anonymity*. Seminar paper, Technische Universität München, Munich, 2013.

[4] K LeFevre, DJ DeWitt, and R Ramakrishnan. Incognito: Efficient full-domain k-anonymity. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, June 2005.

[5] T Li. t-closeness: Privacy beyond k-anonymity and ldiversity. *International Conference on Data Engineering (ICDE)*, 2007.

[6] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L -diversity. *ACM Transactions on Knowledge Discovery from Data*, March 2007.

[7] CE Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review 5(1)*, pages 3–55, 2001.

[8] Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, pages 1–34, 2000.

[9] Latanya Sweeney. Achieving k-Anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05)*, pages 571–588, October 2002.

[10] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5)*, pages 557–570, 2002.

[11] Simon Wagner. *User-assisted analysis of cellular network structures*. Bachelor thesis, Technische Universität München, 2011.

[12] G. Wang. Cornell Anonymization Toolkit. http://anony-toolkit.sourceforge.net/, 2011.

[13] Bin Zhou and Jian Pei. The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. *Knowledge and information systems*, pages 1–38, 2011.