Using Generalization Patterns for Fingerprinting Sets of Partially Anonymized Microdata in the Course of Disasters

Sebastian Schrittwieser*, Peter Kieseberg*, Isao Echizen†, Sven Wohlgemuth† and Noboru Sonehara†
*SBA-Research - Vienna, Austria
Email: sschrittwieser,pkieseberg@sba-research.org
†National Institute of Informatics - Tokyo, Japan
Email: iechizen,wohlgemuth,sonehara@nii.ac.jp

***Abstract*—In the event of large natural and artificial disasters, it is of vital importance to provide all sorts of data to the relief organizations (fire department, red cross,...) to enhance their effectivity. Still, some of this data (e.g. regarding personal information on health status) may be considered private. k-anonymity can be utilized to mitigate the risks resulting from disclosure of such data, however, sometimes it is not possible to achieve a suitable size for k in order to completely anonymize the data without interfering with rescue operations. Still, this data will be sensitive after the disaster recovery is finished. Thus we aim at protecting the data by devising an intrinsic fingerprinting-scheme that allows to detect the source of eventually disclosed information afterwards. Our approach uses the properties directly derived from the anonymization process to generate unique fingerprints for every data set.**

***Keywords*-k-anonymity, fingerprinting, generalization**

I. Introduction

Disastrous events such as the 2011 Tōhoku earthquake and tsunami in Japan require large-scale relief operations with thousands of members of the various participating rescue organizations. To enhance their effectivity, teams are provided with all sorts of data on people affected by the disaster, such as data on special medical treatments or medicines that need to be provided or physically disabled people who need wheelchairs. Still, some of this data (e.g., regarding personal information on health status) may be considered private. A tradeoff between privacy and data value has to be made (in the context of this paper, *data quality* is defined as the informative value of a data set, i.e. the amount of information that can be extracted).

The risks of misuse of sensitive data is especially high, when outsiders without any prior education on data security and privacy are elevated to become insiders. For example in the case of disastrous events, no time can be spared for such an education, thus data leakage prevention is often not possible. In this case, it can still be interesting to later uncover the source of data disclosure after the disaster recovery is done. With this in mind, untrained people may even be more sensible towards the risks the temptation to make money out of disclosing private data to unauthorized "customers" may result in. Due to limited preparation time and the overall chaos of a disaster situation, data breaches are difficult to trace. As a result, disclosure of sensitive microdata is generally prevented by the use of anonymization techniques. The primary technique is k-anonymity [6,9], which enables fine-grained adjustment of the degree of anonymization, i.e., the "anonymization level". While high levels of anonymity can be achieved by using a k-anonymity technique, in many cases, less than complete anonymity is desired so that the usefulness of the microdata in the course of disaster interference is maintained. Although partially anonymized data is usually subject to special usage restrictions, misuse (e.g., disclosure to unauthorized parties) is still possible, especially in the general chaos that results from a major catastrophe. One way to identify a party misusing released microdata is to use fingerprinting. In this paper, we show how k-anonymity techniques can be used to not only partially anonymize microdata but also to fingerprint it for traceability in one single step.

The main contribution of this paper lies in the adaption of intrinsic properties of k-anonymized data for the use in fingerprinting in order to find information leaks after information dissemination in cases of disastrous events.

The remainder of the paper proceeds as follows. First, in Section II, we describe the concept of anonymization, the use of k-anonymity techniques and the intrinsic properties we use in our approach. Since we use data precision for classifying the quality of data sets (and thereby generate the classes of data sets of equivalent quality), we discuss several known metrics for data precision as well. Our approach to fingerprinting sets of partially anonymized microdata is introduced in Section III. In Section III-B, we evaluate the robustness of our approach by giving an example. We conclude in Section IV with a summary of the key points and discuss future work.

II. Anonymization of Microdata

A. General Background

Typically, the first step for the anonymization of sensitive microdata, such as medical health records, is to remove

all uniquely identifying attributes such as name and identification number. However, removing these attributes does not ensure anonymity. For example, Sweeney showed that 87% of all Americans can be uniquely identified using only their ZIP code, birthdate, and gender [8]. Attributes that can be linked to uniquely identify a person are called quasi-identifiers (QIs). To prevent such linking, Sweeney proposed a technique called k-anonymity [9]. The k-anonymity criterion is satisfied if each record is indistinguishable from at least $k-1$ other records in respect to the QIs. Hence, the QI attributes must have the same values within an equivalence class to prevent de-anonymization within such a group of records. One frequently used technique for achieving k-anonymity is *generalization*, in which the granularity of a given QI is changed. For example, instead of including the actual date of birth in the microdata set, only the month and year of birth are included. *Suppression* refers to the highest possible generalization step, where the actual data is not even released.

Table I shows an example of microdata containing sensitive medical health information. The name, of course, is a uniquely identifying attribute that has to be removed for anonymization. Gender and birthdate are quasi-identifiers that could be used for de-anonymization, so they should be generalized or suppressed. The example of data anonymization shown in Table II satisfies a "2-anonymity criterion" because the values for each of the QI attributes are the same for at least two records.

Table I
ORIGINAL MICRODATA.

Name	Gender (a)	Birthdate (b)	Complaint
Bob	m	19.03.1970	diabetic
Dave	m	20.03.1970	disability
Alice	f	18.04.1970	blind
Eve	f	21.04.1970	disability

Table II
ANONYMIZED VERSION WITH $k = 2$.

Birthdate	Gender	Complaint
1970	F	diabetic
1970	M	disability
1970	F	blind
1970	M	disability

Table III
DIFFERENTLY ANONYMIZED VERSION WITH $k = 2$.

Birthdate	Gender	Complaint
03.1970	P	diabetic
03.1970	P	disability
04.1970	P	blind
04.1970	P	disability

Since its introduction, several improvements to the idea of k-anonymity have been proposed. The ℓ-diversity model, for example, enhances anonymity in cases where there is little diversity among sensitive attributes by requiring that each equivalence class has at least ℓ well-represented values for each sensitive attribute [4]. Another example is t-closeness, which requires that the distribution of a sensitive attribute in any equivalence class be close to the distribution of the attribute in the original table [3]. Both ℓ-diversity and t-closeness tighten the criteria for k-anonymity; they do not replace it. The examples we present in this paper are based on the basic k-anonymity criterion for simplicity, and the introduction of further criteria would not require any modification to our approach for fingerprinting.

B. Related Work

Data precision metrics measure the loss of information caused by generalization. A lattice diagram can be used to visualize all possible combinations of generalization operations (see Figure 1). Data precision is our criteria for defining equivalency classes, which we use for fingerprinting, since we want to hand out the same level of data quality to each receiver. Several metrics for measuring the loss of data precision caused by generalization have been used in previous studies. We describe four of the most widely used ones. The important factors are that the metric is able to provide a reasonable classification of the data loss (with respect to the needs derived from the actual data) and that the equivalency classes generated are big enough to provide all receiving organizations with their own, unique set of data.

The *Samarati metric* [5] is very simple. It uses the height of a node in the lattice diagram used to visualize all possible combinations of generalization operations. Although it is intuitive and easy to calculate, this metric takes into account the actual level of generalization only, but not the total number of generalization levels. For example, generalizing {`female, male`} to {`person`} results in much greater information loss than generalizing a date (e.g., birthdate) from day granularity to month granularity. Although both operations are one-level generalizations, the perceived information loss is drastically different. Additionally, this means that an identifier possessing more levels of granularity will have a greater impact on the data precision metric.

The *Precision metric* [7] was introduced by Sweeney in 2002. Its main difference from the Samarati metric is that the generalizations are not weighted equally. Instead, they are weighted with respect to the number of generalization levels possible for the each QI. If a QI has more than one level of generalization (i.e., two levels of granularity), the weighting is given by

$$\frac{\text{Number of levels generalized}}{\text{Number of possible generalization levels}}.$$

For example, if one QI can be granulated into five different levels (the lowest being the exact data), generalization of one level yields a weight of $\frac{1}{4}$, whereas the generalization from {female, male} to {person} would have a weight of 1. The drawback of the precision metric is that it does not factor in the size of the levels, only the number of generalization levels.

The *Discernability metric* [1] takes into account the number of elements distributed to each equivalency class; i.e., the equality of the distribution of the elements to the respective classes is taken into consideration. This stands in direct contrast to the two metrics above, which consider only the classification itself, apart from the actual data and can thus be calculated independently beforehand and for every data set using the same QIs and generalization strategies.

Let N be the number of classes, n_i the number of elements of the i-th class and $n = \sum_N n_i$ the number of all elements. Then

$$DM = \sum_{n_i \geq k} n_i^2 + \sum_{n_i < k} (n \cdot n_i).$$

Despite its name, the Discernability metric is not an actual metric since the criteria of monotonicity is not fulfilled [2].

The *Modified Discernability metric* (DM*), a slightly modified version of the Discernability metric, was proposed by El Emam et al. [2]. Contrary to the original discernability metric it is monotonic.

Again, N denotes the number of classes and n_i the number of elements of the i-th class.

$$DM^* = \sum_N n_i^2$$

The following example shows the data precision value with these different metrics. For simplicity, we use sample medical microdata containing only two quasi-identifiers: gender and birthdate. See Table I for the detailed data. We construct generalizations as depicted in Figure 1. Note that the highest granularity (level 0) is the level of the original data.

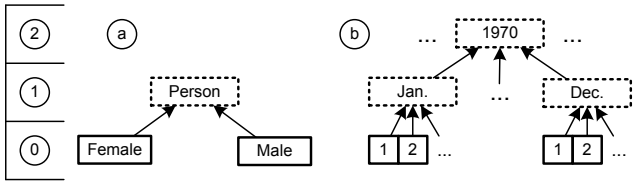

Figure 1. Different levels of generalization for (a) gender and (b) birthdate.

Using these generalization levels, we can build a lattice diagram with the granularity levels used for the gender and birthdate identifiers. The edges show the direct generalizations that can be derived by generalizing only one identifier by one level.

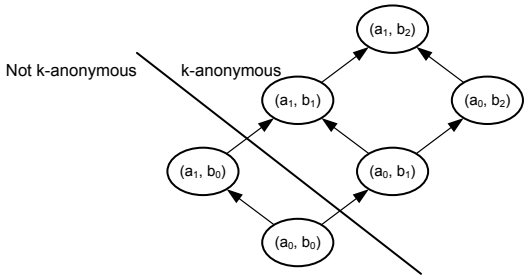

Figure 2. Lattice diagram for all possible generalization variations based on defined levels with $k = 2$.

Table IV shows the data precision values for the different metrics with respect to the generalization model and the underlying data of our example.

Table IV
DATA PRECISION.

Node	Sam	Prec	DM*	k
(a_0, b_0)	0	0	4	1
(a_1, b_0)	1	1	4	1
(a_0, b_1)	1	0.5	8	2
(a_0, b_2)	2	1	8	2
(a_1, b_1)	2	1.5	8	2
(a_1, b_2)	3	2	16	4

III. GENERALIZATION PATTERN BASED FINGERPRINTING

A. General idea

Imagine microdata containing sensitive medical health information that is to be released to three disaster responding units. To protect privacy and avoid linking persons to sensitive attributes, the records are anonymized before release. As anonymization would likely reduce data precision and thus diminish the usefulness of this data, a trade-off between the anonymity level and data precision must be made. The result will likely be partially anonymized microdata sets that still contain sensitive information. They must therefore be protected against unauthorized disclosure. Fingerprinting, a passive form of security, can be used to help identify the source of an unauthorized disclosure. In contrast to watermarking, where hidden data for identification is embedded into the host data, our fingerprinting method uses existing properties of the host data to identify it.

Our approach is to uniquely identify the sets of anonymized microdata, or even subsets thereof, solely based on the generalization levels of the quasi-identifiers. We hypothesize that different generalization operations can be used to achieve the same levels of anonymity and data precision while the resulting microdata sets have unique characteristics that enable their robust identification.

Figure 3 illustrates our idea of clustering datasets with similar anonymity and data precision values. The black dots refer to the different generalization patterns that are arranged along the axis of a coordinate system based on their anonymity and utility values. Similar values above the minimum level of anonymity can be clustered and used for the generation of comparable sets in terms of data quality (this is denoted by the two circles). In our example, the three disaster responding units would receive three different sets of partially anonymized microdata. All sets, however, would have the same or very similar levels of anonymity and data precision. If a subset of one of these microdata sets is disclosed to an unauthorized party, the source could easily be identified on the basis of the unique generalization pattern that was used for anonymization. As discussed before, the quality of a generalization (and thus its affiliation to a quality class) is largely defined by the data precision metric in use. Thus, further research on the comparability of known metrics is still needed in order to enhance the approach described in this paper.

As shown in the example datasets in Tables I–Table III, the same level of anonymity can be achieved with different generalization strategies. While Table II generalizes the birthdate to year granularity, Table III suppresses the gender and generalizes the birthdate to month granularity. The datasets in Tables II and III achieve the 2-anonymity criterion.

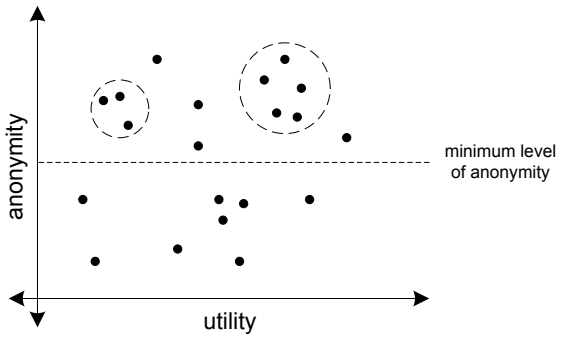

Figure 3. Performing different generalization/suppression operations to achieve the same level of anonymity.

The unique generalization patters can be used as a fingerprint. Consider a result cluster C_1 containing two anonymized data sets with the generalization patterns (a_1, b_1) and (a_0, b_2) for the QIs gender and birthdate. The first one is given to user U_1, and the second one to user U_2. If, for example, data record $(m, 1970)$ is disclosed, it is easy to identify user U_2 as the data source since user U_1 is not in possession of gender data with this granularity. Figure 4 illustrates the approach of determining the generalization levels of the QIs and comparison to stored generalization patterns of released data sets.

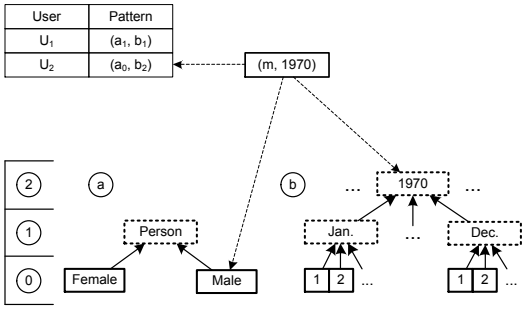

Figure 4. Identifying source of data disclosure on basis of generalization pattern.

B. Evaluation

Robustness is one of the most important aspects of the performance of a fingerprinting method. The robustness of our approach heavily depends on the number and on the data types of the quasi-identifiers. The greater the number of attributes that are quasi-identifiers that need to be generalized or even suppressed, the more uniquely characterizing is the information that can be used for fingerprinting. Moreover, the number of possible generalization steps of quasi-identifiers strongly affects the total number of unique generalization patterns used for fingerprinting. To thwart detection, a malicious recipient could reduce the granularity of several quasi-identifiers to make the origin of the leaked data undetectable (still, the recipient would need to know the generalization strategies in use for other recipients), thus (drastically) reducing the value of the data leaked.

Let N denote the number of quasi-identifiers and n_i the number of generalization levels of the i-th QI. Then the number of possible fingerprints is given by

$$F = \prod_{N} n_i.$$

However, depending on the size of k as well as the structure of the actual data, not all of them are k-anonymous, and most of them do not share the same data precision level. Therefore, they cannot be considered as equally useful for analysis, etc. A general calculation of the number of possible fingerprints with respect to a given k without taking into account the actual data is obviously not possible.

Another consideration is the susceptibility to collusion attacks aimed at both, compromising privacy as well as removing the fingerprint. A well known attack scenario on k-anonymity is the complementary release attack described by Sweeney in 2002 [9]: different releases of the same source set are linked in order to compromise k-anonymity and hence to de-anonymize the data records. This attack exploits a fundamental problem with the concept of k-anonymity that we do not address in our approach. Knowing the generalization patterns and actual data of your released versions of anonymized microdata can help to mitigate the

risk of de-anonymization but does not entirely solve the problem as other data holders might also release some data that could be used for linking. Our proposed fingerprinting, however, is robust against a complementary release attack. Since k-anonymity is based on irreversible generalization operations, even if two or more microdata receivers should combine their data sets in order to de-anonymize them, it is possible to partially reconstruct who contributed which parts of the combined set because specific values for QI attributes can only be retrieved from a data set that has at most a generalization level for that specific QI that equals the level of the leaked data.

C. Example

A simple example is given here to illustrate how a complementary release attack can be used to thwart k-anonymity and how our fingerprinting approach can help to identify the involved parties. The example data set includes two quasi-identifiers:

- Birthdate (a), granulated as in previous example, and
- ZIP code (b), granulated on three levels

Table V shows the original microdata set as well as three different generalizations that fulfill the 2-anonymity criterion.

Table V
ORIGINAL DATA AND THREE GENERALIZATIONS.

Set	Name	Gender	Birthdate	Zip	Complaint
original data set	Bob	m	18.03.1970	1004	diabetic
	Dave	m	19.03.1970	1015	disability
	Alice	f	20.04.1970	1004	blind
	Eve	f	21.04.1970	1015	disability
anonymized set 1	-	p	1970	1004	diabetic
	-	p	1970	1015	disability
	-	p	1970	1004	blind
	-	p	1970	1015	disability
anonymized set 2	-	p	03.1970	100X	diabetic
	-	p	03.1970	101X	disability
	-	p	04.1970	100X	blind
	-	p	04.1970	101X	disability
anonymized set 3	-	m	1970	100X	diabetic
	-	m	1970	101X	disability
	-	f	1970	100X	blind
	-	f	1970	101X	disability

In our example, the local fireguard is provided with data set 1, the police units with data set 2 and the local ambulance with data set 3. After the disaster is overcome, the user rights to the three data sets are revoked. Still, after some time, the record {03.1970, p, 101X,disability} is encountered by the holder of the original data. Due to its form, it is obvious that the data leak must lie inside the local police units, since neither the fireguard (data set 1), nor the ambulance (data set 3) had knowledge on the exact month of birth.

Furthermore, if a data record of the form {03.1970, p, 1015,disability} is encountered, it can be deduced by the same means that the owners of the data sets one and two must have information leaks inside their respective organizations.

IV. CONCLUSION

In this paper, we proposed an approach for the identification of microdata data sets and even subsets based on the unique generalization patterns of the quasi-identifiers used for anonymization. We have shown, that our idea can support the analysis of data breaches as evidence on the source data set can be gained.

Further research in this area includes more in-depth research on data precision metrics, especially their effects on generalization clustering as well as on the clustering itself, focusing on good estimations for the expected cluster sizes. Contrary to the idea of finding one optimal solution, which is aimed in k-anonymity algorithms, we need classes of equivalent solutions, i.e. we mainly need to classify equivalency and be able to discern several solutions that are more or less equally good, not the single best solution. Thus, we aim at developing an efficient algorithm for our approach.

REFERENCES

[1] R. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. 2005.

[2] K. El Emam, F. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, et al. A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, 16(5):670, 2009.

[3] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106–115. IEEE, 2007.

[4] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.

[5] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, pages 1010–1027, 2001.

[6] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, 1998.

[7] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, 10(5):571–588, 2002.

[8] L. Sweeney. Comments to the department of health and human services on "standards of privacy of individually identifiable health information". 2002.

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