Attacker Models with a Variety of Background Knowledge of Payment History

Satoshi Ito[1], Hiroaki Kikuchi[1], Hiroshi Nakagawa[2]

[1] Meiji University Graduate School, Tokyo, 164-8525 Japan,
cs172032@meiji.ac.jp, kikn@meiji.ac.jp
[2] RIKEN Center for Advanced Intelligence Project, Tokyo, 103-0027, Japan,
hiroshi.nakagawa@riken.jp

Abstract. We consider the risks of attackers who try to infer personal identification information from given de-identified datasets. These risks depend on the background knowledge they have. However, it is not known what kinds of background knowledge most increase the risk of re-identification. In this paper, we model attackers of 10 representative types with varying amounts of background knowledge. We quantify the risk of these attackers as the probability of a record being identified using the publicly available open dataset, "Online Retail Data Set".

1 Introduction

In the era of big data, companies are required to assess the re-identification risks of de-identify personal identification information (PII) when employing big data extensively in their business. De-identification is a process of modifying PII to prevent individuals from being identified from the original data. However, there is a concern that attackers may try to re-identify individuals from available de-identified data using external background knowledge. It is unclear what kind of background knowledge is useful for these attackers.

Domingo-Ferrer [8] proposed a model for a maximum-knowledge attacker who is assumed to know both the original and the de-identified datasets. The attacker can use all the attributes as background knowledge to estimate the most likely linkages. However, this attacker is much stronger than that could be expected in a realistic environment and we should relax the assumptions to make the attacker model more realistic. It is well known that there is no universal attacker model because the impact greatly depends on a use case. For example, the impact of leakage of a cancer patients data is bigger than the impact of leakage of a purchase data of convenience store.

Therefore, we aim to study the risk of an attacker depending on their background knowledge. Instead of dealing with a universal model, we focus on smaller specific models, representing 10 types with distinct amounts of background knowledge. We evaluate the probability of a record being identified as the risk associated with these attackers using the Online Retail Data Set, which is available from the UCI Machine Learning Repository [17].

Our study makes two contributions: (1) we propose a theoretical risk model; (2) we conducted an experiment to estimate risk using real purchase history data. Our theoretical model allows us to estimate the mean probability of a record being identified by an attacker who has background knowledge of specific attributes without exactly computing risks. The probability can be determined with the number of unique values of the attributes. We demonstrate the efficiency of our model using empirical evidence.

The rest of this paper is organized as follows. In Section 2, we describe and analyze the Online Retail Data Set. In Section 3, we describe the attacker models with 10 representative types with distinct background knowledge. In Section 4, we evaluate the risk of these attackers. Section 5 describes related work, and Section 6 concludes the paper.

2 The Online Retail Data Set

The Online Retail Data Set from the UCI Machine Leaning Repository [17] contains one year of purchase history data from the UK. This data was used in the PWSCUP [13] data deidentification competition in Japan in 2016 and 2017. In this study, we use a subset of the Online Retail Data Set that consists of the purchase history data T for 400 customers. We show a summary of T, an example, and the statistics of T in Tables 1, 2, and 3, respectively. We define the following.

Definition 2.1 (m, n, ω_X) Let m and n be the number of records and the number of customers in T, respectively. Let ω_X be the unique number of kinds of values of attribute X in T.

Example 2.1 Table 2 is a table of $m = 10$ and $n = 5$. Suppose attribute $X =$ purchase date, which has distinct three values, i.e., 2010/12/1, 12/5, and 12/6, hence we have $\omega_X = 3$. For $Y =$Number, we have $\omega_Y = 6$.

The attackers in our model are classified into three classes in terms of attributes, the purchase date, the number of kinds of goods purchased in a day, and the goods purchased in a day. Note that the combination of these attributes constitutes background knowledge. Table 4 shows the statistics of these attributes of T. T has 373 days of purchase history records observed from December 10, 2010 to December 9, 2011. It contains 290 days when purchases were made; no purchases were made on 83 days. Figure 1 shows the distribution of the number of kinds of goods purchased in a day in T. The numbers of kinds of goods are distributed in a power-lour with mean of 13.75 at the top of distribution. Note that there is one more peak at the left most number 1, that is, just a single item, and occurred 73 times, which are the most frequent records in T. A total of 114 kinds of goods are stored in T. Figure 2 shows the frequency distribution of goods purchased in a day, where the X-axis indicates the rank. For example,

Table 1: Summary of purchase data T

Attribute	Detail
User ID	ID of user (5 digit number)
Receipt ID	ID of receipt (6 digit number)
Date	Purchase date (yyyy/mm/dd)
Time	Purchase time (hh:mm)
Goods	ID of purchased goods (number and character)
Price	Price of purchased goods (Pound sterling)
Number	Quantity of purchased goods (number)

Table 2: Example of T

User ID	Receipt ID	Date	Time	Goods	Price	Number
12583	536370	2010/12/1	8:45	22728	3.75	24
12583	536370	2010/12/1	8:45	22727	3.75	24
12431	536389	2010/12/1	10:03	22941	8.5	6
12431	536389	2010/12/1	10:03	21622	4.95	8
12431	536389	2010/12/1	10:03	21791	1.25	12
12838	536415	2010/12/1	11:57	22952	0.55	10
12567	537065	2010/12/5	11:57	22837	4.65	8
12567	537065	2010/12/5	11:57	22846	16.95	1
12748	537429	2010/12/6	15:54	84970S	0.85	12
12748	537429	2010/12/6	15:54	22549	1.45	8

there are 2781 kinds of goods in T and the most frequent item was purchased more than 1000 times.

3 Ten Types of Attackers

We consider 10 typical types of attackers who have background knowledge of T. Let $T(id, day)$ be a table transformed from T containing records with user ID "id", purchase date "day", and purchased items. This table is a data of the "Goods" attribute of T about two attributes, "User ID" and "Date". Suppose that an attacker obtains background knowledge about target user u and tries to identify u from T. For simplification, we consider the small dataset T_2 and the sample transformed table $T_2(id, day)$ shown in Tables 5 and 6, respectively. We will describe the background knowledge about T in Section 3.1 and the 10 types of attackers using T_2 and $T_2(id, day)$ in Section 3.2.

Table 3: Statistics of T and T_2

	Online retail T	Sample data T_2
Number of records m	38087	10
Number of users n	400	3
Number of goods ω_{goods}	2871	4
Purchase date ω_{day}	2010/12/1–2011/12/9	2010/12/1–2010/12/3

Table 4: Statistics of three kinds of knowledge in T

	Purchase date	Goods/day	Goods
Unique values ω_X	290	114	2781
Mean frequency	5.4	13.75	13.7
Max frequency	21	71	1012
Mode of frequency	5	1	1

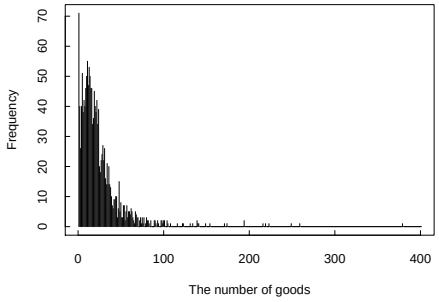

Fig. 1: Frequency of numbers of kinds of goods in T

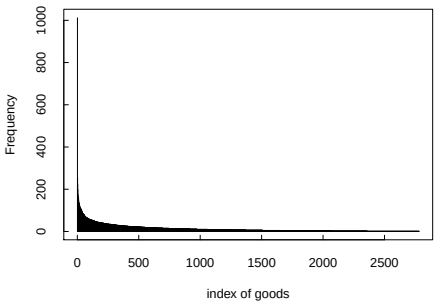

Fig. 2: Frequency of index of goods in T

3.1 Background Knowledge in T

Suppose that an attacker can gain background knowledge from T. Seven attributes in T are classified into three types: (1) *who* (user ID); (2) *when* (purchase date and time); and (3) *what* (goods, price, and quantities). However, it is not likely that attackers can learn "*who*" purchased as background knowledge. Therefore, we assume that attackers learn "*when*" or "*what*" information as background knowledge. In this study, we consider the attributes, "purchase date" and "goods" because these are representative attributes of the class of "*when*" and "*what*" as background knowledge. In addition, we also consider background knowledge of the number of kinds of goods purchased in a day.

Questions "when did u purchase?" and "how many kinds of goods did u purchase?" take Boolean values, yes or no. We suppose that question that "what did u purchase?" has three possible answers; yes (attacker knows all goods), yes (knows one of goods), or no (no knowledge). By combining these questions, we have possible $2 \times 2 \times 3 = 12$ types of attackers. Note that we omit two extreme attackers who learn all the goods purchased in a day without knowing the number of kinds of goods purchased with/without knowing date, as the assumptions contradict each other. Finally, we have the 10 types of attackers shown in Table 7.

3.2 Detail of the Ten Types of Attackers

Among ten type attackers, we describe representative two of them. Attacker 1 learns only one of the items that customer u purchased on the day. For example,

Table 5: Sample purchase data T_2

User ID	Receipt ID	Date	Time	Goods	Price	Number
1	100	2010/12/1	8:45	Bread	1.45	2
1	100	2010/12/1	8:45	Book	3.75	1
1	200	2010/12/1	20:10	Tea	0.85	2
2	300	2010/12/1	10:03	Bread	1.45	3
1	400	2010/12/2	15:07	Tea	0.85	3
3	500	2010/12/2	11:57	Bread	1.45	4
3	500	2010/12/2	11:57	Juice	1.25	4
3	600	2010/12/3	15:54	Book	3.75	1
3	600	2010/12/3	15:54	Tea	0.85	10
3	600	2010/12/3	15:54	Juice	1.45	10

Table 6: Transformed T_2 with user ID and purchase date, $T_{2(id,day)}$

User ID \Date	2010/12/1	2010/12/2	2010/12/3
1	Bread, book, tea	Tea	
2	Bread		
3		Bread, juice	Tea, tea, juice

when this attacker knows background knowledge that "u purchased tea one day", he finds two users (user 1 and 3) who purchased tea in one day in T_2. Then, he can identify u with the probability of $\frac{1}{2}$. Attacker 6 learns one purchase date when customer u purchased, and only one of the items purchased on that day. For example, when this attacker knows the background knowledge that "u purchased tea on December 1, 2010", he finds that only user 1 satisfies these requirements in T_2. Therefore, the attacker can identify u uniquely.

Similarly, the other attackers 0, 2, 3,..., 9, are defined from background knowledge. Table 7 summarizes the relationship between attacker types and their given knowledge.

Table 7: Ten types of attacker

Attacker	Date	Number of kinds	Goods
0	–	–	–
1	–	–	One
2	–	✓	–
3	–	✓	One
4	–	✓	All
5	✓	–	–
6	✓	–	One
7	✓	✓	–
8	✓	✓	One
9	✓	✓	All

4 Evaluating Risk of Attackers

4.1 Mean Probability of Identification

In this study, we define a mean probability of identification of u in T, to be identified by an attacker who has background knowledge (attribute) X. We assume that the probability of attackers gaining background knowledge is proportional to the frequency of the knowledge appearance in T.

For instance, from T_2, Table 8 shows the probabilities $T_2(id, day)$ and the mean identification probabilities $Pr(\text{identify}|X)$ for Attacker 5 to gain background knowledge. In this case, Attacker 5 learns one piece of background knowledge X about u from three candidates, "u purchased on December 1, 2010", "December 2, 2010", and "December 3, 2010". The probabilities of these events depend on the frequencies of records in T_2. Table 8 shows the number of users who satisfy these requirements, X and days, and the probability of identifying u given X. For example, if Attacker 5 knows that "u purchased on December 3, 2010", he can identify u uniquely as the only customer who satisfies all the requirements. In this case, Attacker 5 has this background knowledge with probability of 0.3 and can identify u uniquely (with probability 1). Therefore, the probability of identifying u given this background knowledge X is

$$Pr(\text{identify}, X) = Pr(X) \cdot Pr(\text{identify}|X) = 0.3 \cdot 1 = 0.3.$$

We give the expected value of probability, i.e., the sum of the probabilities to identify u with all background knowledge. In this case, the risk associated with Attacker 5 is

$$Pr(\text{identify by Attacker 5}) = \sum_X Pr(\text{identify}, X) = 0.2 + 0.15 + 0.3 = 0.65,$$

which means that u is identified by attacker 5 with probability 65%.

Table 9 shows the probability that Attacker 3 has background knowledge X =(# kinds, one good) in $T_2(id, day)$ and the identification probabilities. For example, the event that "someone purchased three kinds of goods and one of these was a book" appears twice and hence the probability of appearance $Pr(X)$ is 0.2 ($= \frac{2}{10}$). In this case, two users satisfy this requirement so the probability of identifying a user is $Pr(\text{identify}|\text{X}) = \frac{1}{\#\text{users with } X} = \frac{1}{2} = 0.5$ and the total risk associated with Attacker 3 is 0.8. Note that the number of events (records) that appear is not necessarily the same as the number of distinct users who satisfy requirements, #users. For instance, when user 1 purchased three books one by one for three days, the number of events that "someone purchased one kind of goods, a book" is three but #users is 1 because only one same user satisfies the requirement.

4.2 Properties of Mean Identification Probability

Definition 4.1 (R_X, U_X) Let X be an element of the set of background knowledge $D(X)$ in table T. Let R_X and U_X be sets of records of T and users of T, respectively, that satisfy restriction X.

Table 8: Risk of Attacker 5 in T_2

Knowledge X	Frequency	Prob. of occur. $Pr(X)$	#users	Prob. of identification $Pr(\text{id}\|X)$	Risk $Pr(\text{id}, X)$
2010/12/1	2	0.4	2	0.5	0.2
2010/12/2	2	0.3	2	0.5	0.15
2010/12/3	1	0.3	1	1	0.3
Sum	5	1			0.65

Table 9: Background knowledge of Attacker 3 in T_2

Knowledge X	Frequency	Prob. of occur. $Pr(X)$	#users	Prob. of identification $Pr(\text{id}\|X)$	Risk $Pr(\text{id}, X)$
1, bread	1	0.1	1	1	0.1
1, tea	1	0.1	1	1	0.1
2, bread	1	0.1	1	1	0.1
2, juice	1	0.1	1	1	0.1
3, bread	1	0.1	1	1	0.1
3, book	2	0.2	2	0.5	0.1
3, tea	2	0.2	2	0.5	0.1
3, juice	1	0.1	1	1	0.1
Sum	10	1			0.8

Example 4.1 For T_2, $D(X) = \{2010/12/1, 2010/12/2, 2010/12/3\}$, for element $X =$"2010/12/1" of $D(X)$, we have $R_X = \{1, 2, 3, 4\}$ and $U_X = \{1, 2\}$.

We define the mean identification probability for an attacker who learns only one kind of background knowledge (Attackers 1, 2, or 5) in the following theorem.

Theorem 4.1 When $|R_X| = |U_X|$, the mean identification probability $Pr(\text{attacked with } X)$ for an attacker who knows only one kind of background knowledge X about data T is

$$Pr(\text{attacked with } X) = \frac{\omega_X}{m}.$$

Proof According to Definition 4.1, the mean identification probability is

$$Pr(\text{attacked with } X) = \sum_{X \in D(X)} \frac{|R_X|}{m} \frac{1}{|U_X|} = \frac{1}{m}$$

with the assumption of $|R_X| = |U_X|$. Hence,

$$Pr(\text{attacked with } X) = \sum_{X \in D(X)} \frac{1}{m} = \frac{\omega_X}{m}.$$

Theorem 4.1 is proved.

Example 4.3 When $D(X) = \{2010/12/1, 2010/12/2, 2010/12/3\}$ and $X =$"2010/12/1", $R_X = \{1, 2, 3, 4\}$ and $U_X = \{1, 2\}$ because $m = 10$ and $\omega_X = 4$, the conditional mean probability of identification given X is $Pr(\text{identify}|X) = \frac{1}{|U_X|} = \frac{1}{2}$

and the probability of gaining X is $Pr(X) = \frac{|R_X|}{m} = \frac{4}{10}$. Therefore, the mean identification probability of an attacker who knows one element of $D(X)$ is

$$Pr(\text{identify}, X) = \sum_{X \in D(X)} \frac{|R_X|}{m} \frac{1}{|U_X|} = \frac{4+3+6}{20} = 0.65,$$

when $|R_X| = |U_X|$. Hence, the overall risk (expected probability of identification) from attacker who knows X is

$$Pr(\text{attacked with } X) = \sum_{X \in D(X)} \frac{1}{m} = \frac{\omega_X}{m} = 0.3.$$

Theorem 4.2 When $|R_X| = |U_X|$, the mean identification probability from attacker who knows $Pr(\text{attacked with } X, Y)$ independent background knowledges X and Y is

$$Pr(\text{attacked with } X, Y) = \frac{\omega_X \omega_Y}{m}.$$

Proof From the premise of independence of X and Y, when Y is an element of set $D(Y)$, the probability of gaining both X and Y is $Pr(X, Y) = Pr(X)Pr(Y) = \frac{|R_X|}{m}\frac{|R_Y|}{m}$. From the assumption that $|R_X| = |U_X|$, the number of users who meet X and Y is: $m \cdot \frac{|R_X|}{m}\frac{|R_Y|}{m}$ and the risk for an attacker who knows both X and Y is $1/m(|R_X|/m)(|R_Y|/m)$. Therefore, the mean identification probability for an attacker who knows X and Y is:

$$\begin{aligned} Pr(\text{attacked with } X, Y) &= \sum_{X \in D(X)} \sum_{Y \in D(Y)} \frac{\frac{|R_X|}{m}\frac{|R_Y|}{m}}{m\frac{|R_X|}{m}\frac{|R_Y|}{m}} \\ &= \sum_{X \in D(X)} \sum_{Y \in D(Y)} \frac{1}{m} = \frac{\omega_X \omega_Y}{m}. \end{aligned}$$

Theorem 4.2 is proved.

Example 4.4 When $D(X) = \{2010/12/1, 2010/12/2, 2010/12/3\}$, $D(Y) = \{1, 2, 3\}$, $X = \{2010/12/1\}$, $Y = \{1\}$, $R_X = \{1, 2, 3, 4\}$, $U_X = \{1, 2\}$, $R_Y = \{4, 5\}$, and $U_Y = \{1, 2\}$ because $m = 10$ and $\omega_X = 4$ and $\omega_Y = 3$. Under the assumption that $|R_X| = |U_X|$ and X is independent from Y, the mean identification probability $Pr(\text{identify}, X, Y)$ for an attacker who knows X and Y is $1/m(|R_X|/m)(|R_Y|/m) = \frac{10}{8}$ and the probability of gaining X and Y is $Pr(X, Y) = \frac{|R_X|}{m}\frac{|R_Y|}{m} = \frac{8}{100}$. Therefore, the mean identification probability for an attacker who knows X and Y is $Pr(\text{attacked with } X, Y) = \frac{\omega_X \omega_Y}{m} = 0.9$.

Corollary 4.3 When $|R_X| = |U_X|$, the mean identification probability $Pr(\text{identify}|X_1, X_2, ..., X_k)$ for an attacker who knows k kinds of background knowledge X in T is

$$Pr(\text{identify}|X_1, X_2, \ldots, X_k) = \frac{\omega_{X_1} \omega_{X_2} \cdots \omega_{X_k}}{m}.$$

4.3 Experimental Results with the Online Retail Data Set

We investigated the Online Retail Data Set to compute the exact probabilities of identifying customers using this data. Table 10 shows the experimental results; the risks associated with attackers in T. These exact values are computed in the same way as the examples in Section 4.1 and the theoretical values are estimated using the theorems in Section 4.2. In this case, we have $\omega_{day} = 290$, $\omega_{num} = 114$, and $\omega_{goods} = 2781$. Attackers 1, 2, and 5 know only one kind of background knowledge (purchase date, number of kinds, or one of the goods purchased, respectively) about purchase data T.

Based on the experiment, we found that Attacker 5 is the most serious one (0.1851 risk) among these three attackers. The risk of attackers increases monotonically with the number of kinds of background knowledge.

Table 10: Risks for 10 attackers in T

Attacker	Actual value	Theoretical value
0	0.0025	0.0025
1	0.0965	0.0730
2	0.0807	0.0030
3	0.7974	8.3239
4	0.9788	4.5436
5	0.1851	0.0076
6	0.8945	21.1749
7	0.9400	0.8680
8	0.9750	2413.9433
9	0.9994	1317.6433

Table 11: Ranks of risks for attackers

Rank	Actual value	Theoretical value
1	**9**	**8**
2	4	**9**
3	**8**	6
4	7	3
5	6	4
6	3	7
7	5	1
8	1	5
9	2	2
10	0	0

4.4 Discussion

Table 11 compares the exact ranks of attackers and the risks estimated in our model. In either case, Attacker 0 is the weakest but the riskiest attackers are inconsistent between the two models. In the estimated risks, Attacker 8 with knowledge of "purchase date", "number of kinds", and "one of goods" is ranked as the most threatening. However, in the exactly computed result, the background knowledge "purchase date", "number of kinds", and "all goods" appear to give attackers the highest risk. However, Attacker 8 knows obviously less useful background knowledge than Attacker 9. Therefore, this result does not make sense. Furthermore, the distance between the exactly computed and theoretically estimated mean identification probabilities is very large, and some estimates are much greater than 1.0 (note that these are probabilities).

We believe that the reasons for these problems are the way to compute and the assumptions. The actual values and the theoretical values are computed in different ways. We compute the theoretical values as $p_X = \frac{\text{frequency of } X}{\text{sum of appearance of events}}$ as the probability p_X to gain X. However, we compute the exact values as $p'_X = \frac{\text{number of records about } X}{\text{number of records in } T} = \frac{|R_X|}{m}$ as the probability p_X to gain X. We further

make two assumptions, $|U_X| = |R_X|$ and independence of background knowledge in our theorems. Figures 3, 4, and 5 show scatter diagrams for $|U_X|$ and $|R_X|$ for purchase date, number of kinds, and purchase of one good. The red lines in these figures show $|U_X| = |R_X|$ and most points are far from this line in any figure. Therefore, this assumption is not realistic. The assumption of independence is also incorrect. Under this assumption, the probability of gaining background knowledge X and Y is $p_X p_Y$, and this value is greater than zero. However, few combinations of X and Y are observed (the probability of learning X and Y is zero) in the real data. For example, for purchase date and number of kinds in T, the number of combinations of these events is 33,060 ($= \omega_{day}\omega_{num} = 290 \cdot 114$) but only 1473 combinations (about 0.46%) appear in T. Therefore, this assumption need to be improved.

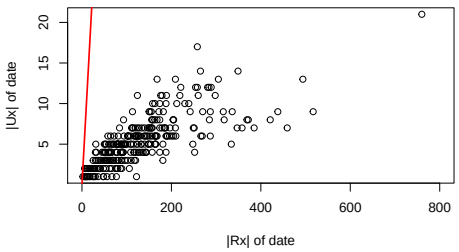

Fig. 3: Scatter plot of $|Ux|$ and $|Rx|$ for purchase date in T

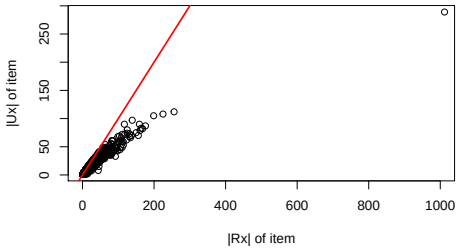

Fig. 4: Scatter plot of $|Ux|$ and $|Rx|$ for number of kinds in T

5 Related Work

There are two representative methods to evaluate the privacy level of data, k-anonymity [1] and differential privacy [2]. The k-anonymity was proposed by Sweeney in 2006. It evaluates the privacy level of data according to whether the data have at least k indistinguishable records in terms of quasi-identifiers.

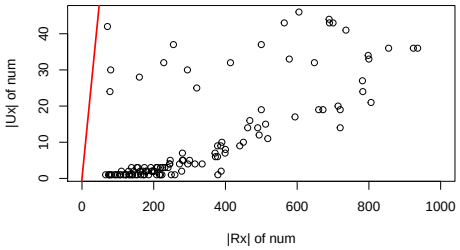

Fig. 5: Scatter plot of $|Ux|$ and $|Rx|$ for goods in T

Differential privacy was introduced by Dwork in 2006. It evaluates the privacy level of data according to whether the possibility of restoring personal data from differences in analysis results of the data is high.

Technical Specification ISO/TS 25237 [4] defines *anonymization* as "a process that removes the association between the identifying data and the data subject." The ISO definition classifies anonymization techniques into masking and deidentification, and has been considered favorably [5]. Many anonymization algorithms have been proposed to preserve privacy while retaining the utility of the data that have been *anonymized*. That is, the data are made less specific so that a particular individual cannot be identified. Anonymization algorithms employ various operations, including *suppression* of attributes or records, *generalization* of values, replacing values with *pseudonyms*, *perturbation* with random noise, sampling, rounding, swapping, top/bottom coding, and microaggregation [3, 6].

Domingo [8] proposed a model for a maximum-knowledge attacker who knows both the original dataset and the anonymized dataset. The attacker can use all the attributes to estimate the best possible linkages. Koot et al. proposed a method to quantify anonymity via an approximation of the uniqueness probability using a measure of the Kullback–Leibler distance [9]

Monreale et al. proposed a framework for the anonymization of semantic trajectory data, called c-safety [10]. Based on this framework, Basu et al. presented an empirical risk model for privacy based on k-anonymous data release [11]. Their experiment using car trajectory data gathered in the Italian cities of Pisa and Florence allowed the empirical evaluation of the protection of anonymization of real-world data. Stokes et al. defined n-confusion [14] that is a generalization of k-anonymity.

In 2017, Torra presented a general introduction on data privacy [15]. Zhizhou and Lai proposed a definition of a new δ-privacy model that requires that no adversary could improve more than δ privacy degree [16].

6 Conclusions

We have proposed 10 types of attackers with background knowledge about 400 users from the Online Retail Data Set and evaluated the risk (mean identification probability) associated with these attackers. We found that information about purchase date is the most useful for attackers among three kinds of background knowledge: purchase date, number of kinds, and knowledge of one good purchased. The risk associated with attackers rises greatly when they have multiple kinds of background knowledge.

We demonstrated that the risk theoretically can be estimated without computing it exactly under two assumptions ($|U_x| = |R_x|$ and independency). Unfortunately, the assumption of independency does not always hold, and we need to improve our models in precision.

We plan to create more accurate models for evaluating risks and to investigate the change in risks when data are anonymized.

References

1. L. Sweeny, "k-anonymity: a model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557-570, 2006.
2. C. Dwork, "Differential privacy", Proceedings of ICALP 2006, LNCS vol.4052, pp.1-12, 2006.
3. Information Commissioner's Office (ICO), Anonymisation: managing data protection risk code of practice, 2012.
4. "Health informatics – Pseudonymization", ISO Technical Specification ISO/TS 25237.
5. Khaled El Emam, Luk Arbuckle, "Anonymizing Health Data Case Studies and Methods to Get You Started", *O'Reilly*, 2013.
6. C.C. Aggarwal and P.S. Yu., "A General Survey of Privacy-Preserving Data Mining, Models and Algorithms", *Privacy-preserving data mining*, Springer, pp. 11-52, 2008.
7. J. Domingo-Ferrer and V. Torra, "A quantitative comparison of disclosure control methods for microdata", Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 111-133, 2001.
8. Josep Domingo-Ferrer, Sara Ricci and Jordi Soria-Comas, "Disclosure Risk Assessment via Record Linkage by a Maximum-Knowledge Attacker", 2015 Thirteenth Annual Conference on Privacy, Security and Trust (PST), *IEEE*, 2015.
9. Koot, M. R., Mandjes, M., van't Noordende, G., and de Laat, C., "Efficient probabilistic estimation of quasi-identifier uniqueness", In Proceedings of ICT OPEN 2011, 14-15, pp. 119-126, 2011.
10. A Monreale, R Trasarti, D Pedreschi, C Renso and V Bogorny, "C-safety: a framework for the anonymization of semantic trajectories", Transactions on Data Privacy, Vol. 4 (2), pp. 73-101, 2011.
11. A. Basu, A. Monreale, R. Trasarti, J. C. Corena, F. Giannotti, D. Pedreschi, S. Kiyomoto, Y. Miyake and T. Yanagihara, "A risk model for privacy in trajectory data", Journal of Trust Management, 2:9, 2015.
12. Daqing Chen, Sai Liang Sain, and Kun Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3, pp. 197–208, 2012.
13. H. Kikuchi, T. Yamaguchi, K. Hamada, Y. Yamaoka, H. Oguri and J. Sakuma, "What is the Best Anonymization Method? - a Study from the Data Anonymization Competition Pwscup 2015", Data Privacy Management Security Assurance (DPM2016), LNCS 9963, pp. 230 - 237, 2016.
14. Klara Stokes, Vicenç Torra, n-confusion: a generalization of k-anonymity, EDBT/ICDT Workshops 2012: 211-215.
15. V. Torra, "Data Privacy: Foundations, New Developments and the Big Data Challenge", Studies in Big Data 28, Springer, 2017.
16. Zhizhou Li, Ten H. Lai, δ-privacy: Bounding Privacy Leaks in Privacy Preserving Data Mining, DPM/CBT 2017, LNCS 10436, pp. 124–142, Springer, 2017.
17. UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/index.php, April 28, 2018.