Assessing Privacy Risk in Retail Data

Roberto Pellungrini[1], Francesca Pratesi[1,2], and Luca Pappalardo[1,2]

[1] Department of Computer Science, University of Pisa, Italy
[2] ISTI-CNR, Pisa, Italy

Abstract. Retail data are one of the most requested commodities by commercial companies. Unfortunately, from this data it is possible to retrieve highly sensitive information about individuals. Thus, there exists the need for accurate individual privacy risk evaluation. In this paper, we propose a methodology for assessing privacy risk in retail data. We define the data formats for representing retail data, the privacy framework for calculating privacy risk and some possible privacy attacks for this kind of data. We perform experiments in a real-world retail dataset, and show the distribution of privacy risk for the various attacks.

1 Introduction

Retail data are a fundamental tool for commercial companies, as they can rely on data analysis to maximize their profit [7] and take care of their customers by designing proper recommendation systems [11]. Unfortunately, retail data are also very sensitive since a malicious third party might use them to violate an individual's privacy and infer personal information. An adversary can re-identify an individual from a portion of data and discover her complete purchase history, potentially revealing sensitive information about the subject. For example, if an individual buys only fat meat and precooked meal, an adversary may infer a risk to suffer from cardiovascular disease [4]. In order to prevent these issues, researchers have developed privacy preserving methodologies, in particular to extract association rules from retail data [3,10,5]. At the same time, frameworks for the management and the evaluation of privacy risk have been developed for various types of data [1,13,2,9,8].

We propose privacy risk assessment framework for retail data which is based on our previous work on human mobility data [9]. We first introduce a set of data structures to represent retail data and then present two re-identification attacks based on these data structures. Finally, we simulate these attacks on a real-world retail dataset. The simulation of re-identification attacks allows the data owner to identify individuals with the highest privacy risk and select suitable privacy preserving technique to mitigate the risk, such as k-anonymity [12].

The rest of the paper is organized as follows. In Section 2, we present the data structures which describe retail data. In Section 3, we define the privacy risk and the re-identification attacks. Section 4, shows the results of our experiments and, finally, Section 5 concludes the paper proposing some possible future works.

2 Data Definitions

Retail data are generally collected by retail companies in an automatic way: customers enlist in membership programs and, by means of a loyalty card, share informations about their purchases while at the same time receiving special offers and bonus gifts. Products purchased by customers are grouped into baskets. A basket contains all the goods purchased by a customer in a single shopping session.

Definition 1 (Shopping Basket). *A shopping basket S_j^u of an individual u is a list of products $S_j^u = \{i_1, i_2, \ldots, i_n\}$, where i_h $(h = 1, \ldots, n)$ is an item purchased by u during her j-th purchase.*

The sequence of an individual's baskets forms her shopping history related to a certain period of observation:

Definition 2 (History of Shopping Baskets). *The history of shopping baskets HS^u of an individual u is a time-ordered sequence of shopping baskets $HS^u = \{S_1^u, \ldots, S_m^u\}$.*

3 Privacy Risk Assessment Model

In this paper we start from the framework proposed in [9] and extended in [8], which allows for the assessment of the privacy risk in human mobility data. The framework requires the identification of the minimum data structure, the definition of a set of possible attacks that a malicious adversary might conduct on an individual, and the simulation of these attacks. An individual's privacy risk is related to her probability of re-identification in a dataset w.r.t. a set of re-identification attacks. The attacks assume that an adversary gets access to a retail dataset, then, using some previously obtained background knowledge, i.e., the knowledge of a portion of an individual's retail data, the adversary tries to re-identify all the records in the dataset regarding that individual. We use the definition of privacy risk (or re-identification risk) introduced in [12].

The background knowledge represents how the adversary tries to re-identify the individual in the dataset. It can be expressed as a hierarchy of categories, configurations and instances: there can be many background knowledge categories, each category may have several background knowledge configurations, each configuration may have many instances. A background knowledge category is an information known by the adversary about a specific set of dimensions of an individual's retail data. Typical dimensions in retail data are the items, their frequency of purchase, the time of purchase, etc. Examples of background knowledge categories are a subset of the items purchased by an individual, or a subset of items purchased with additional spatio-temporal information about the shopping session. The number k of the elements of a category known by the adversary gives the background knowledge configuration. This represents the fact that the quantity of information that an adversary has may vary in size. An

example is the knowledge of $k = 3$ items purchased by an individual. Finally, an instance of background knowledge is the specific information known, e.g., for $k = 3$ an instance could be eggs, milk and flour bought together. We formalize these concepts as follows.

Definition 3 (Background knowledge configuration). *Given a background knowledge category $\mathcal{B}$, we denote by $B_k \in \mathcal{B} = \{B_1, B_2, \ldots, B_n\}$ a specific background knowledge configuration, where k represents the number of elements in $\mathcal{B}$ known by the adversary. We define an element $b \in B_k$ as an* instance *of background knowledge configuration.*

Let $\mathcal{D}$ be a database, D a retail dataset extracted from $\mathcal{D}$ (e.g., a data structure as defined in Section 2), and D_u the set of records representing individual u in D, we define the probability of re-identification as follows:

Definition 4 (Probability of re-identification). *The probability of re-identification $PR_D(d = u|b)$ of an individual u in a retail dataset D is the probability to associate a record $d \in \mathcal{D}$ with an individual u, given an instance of background knowledge configuration $b \in B_k$.*

If we denote by $M(D, b)$ the records in the dataset D compatible with the instance b, then since each individual is represented by a single History of Shopping Baskets, we can write the probability of re-identification of u in D as $PR_D(d = u|b) = \frac{1}{|M(D,b)|}$. Each attack has a matching function that indicates whether or not a record is compatible with a specific instance of background knowledge.

Note that $PR_D(d{=}u|b) = 0$ if the individual u is not represented in D. Since each instance $b \in B_k$ has its own probability of re-identification, we define the risk of re-identification of an individual as the maximum probability of re-identification over the set of instances of a background knowledge configuration:

Definition 5 (Risk of re-identification or Privacy risk). *The risk of re-identification (or privacy risk) of an individual u given a background knowledge configuration B_k is her maximum probability of re-identification $Risk(u, D) = \max PR_D(d{=}u|b)$ for $b \in B_k$. The risk of re-identification has the lower bound $\frac{|D_u|}{|D|}$ (a random choice in D), and $Risk(u, D) = 0$ if $u \notin D$.*

3.1 Privacy attacks on retail data

The attacks we consider in this paper consist of accessing the released data in the format of Definition (2) and identifying all users compatible with the background knowledge of the adversary.

Intra-Basket Background Knowledge. We assume that the adversary has as background knowledge a subset of products bought by her target in a certain shopping session. For example, the adversary once saw the subject at the workplace with some highly perishable food, that are likely bought together.

Definition 6 (Intra-Basket Attack). *Let k be the number of products of an individual w known by the adversary. An Intra-Basket background knowledge instance is $b = S'_i \in B_k$ and it is composed by a subset of purchase $S'_i \subseteq S^w_j$ of length k. The Intra-Basket background knowledge configuration based on k products is defined as $B_k = S^{w[k]}$. Here $S^{w[k]}$ denotes the set of all the possible k-combinations of the products in each shopping basket of the history.*

Since each instance $b = S'_i \in B_k$ is composed of a subset of purchase $S'_i \subseteq S^w_j$ of length k, given a record $d = HS^u \in D$ and the corresponding individual u, we define the matching function as:

$$matching(d, b) = \begin{cases} true & \exists\ S^d_j \mid S'_i \subseteq S^d_j \\ false & otherwise \end{cases} \tag{1}$$

Full Basket Background Knowledge. We suppose that the adversary knows the contents of a shopping basket of her target. For example, the adversary once gained access to a shopping receipt of her target. Note that in this case it is not necessary to establish k, i.e., the background knowledge configuration has a fixed length, given by the number of items of a specific shopping basket.

Definition 7 (Full Basket Attack). *A Full Basket background knowledge instance is $b = S^w_j \in B$ and it is composed of a shopping basket of the target w in all her history. The Full Basket background knowledge configuration is defined as $B = S^w_i \in HS^w$.*

Since each instance $b = S^w_i \in B$ is composed of a shopping basket S^w_i, given a record $d = HS^u \in D$ and the corresponding individual u, we define the matching function as:

$$matching(d, b) = \begin{cases} true & \exists\ S^d_j \mid S^w_i = S^d_j \\ false & otherwise \end{cases} \tag{2}$$

4 Experiments

For the Intra-basket attack we consider two sets of background knowledge configuration B_k with $k = 2, 3$, while for the Full Basket attack we have just one possible background knowledge configuration, where the adversary knows an entire basket of an individual. We use a retail dataset provided by Unicoop[3] storing the purchases of 1000 individuals in the city of Leghorn during 2013, corresponding to 659,761 items and 61,325 baskets. We consider each item at the category level, representing a more general description of a specific item, e.g., "Coop-brand Vanilla Yogurt" belongs to category "Yogurt".

We performed a simulation of the attacks for all B_k. We show in Fig. 1 the cumulative distributions of privacy risks. For the Intra-basket attack, with $k = 2$ we have almost 75% of customers for which privacy risk is to equal 1.

[3] `https://www.unicooptirreno.it/`

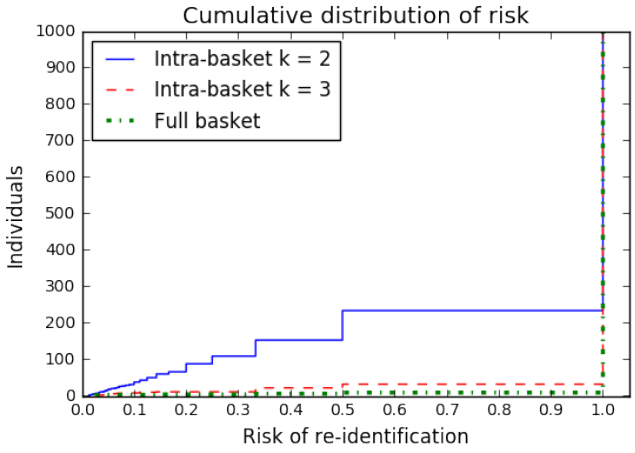

Fig. 1. Cumulative distributions for privacy attacks.

Switching to $k = 3$ causes a sharp increase in the overall risk: more than 98% of individuals have maximum privacy risk (e.g., 1). The difference between the two configurations is remarkable, showing how effective an attack could be with just 3 items. Since most of customers are already re-identified, further increasing the quantity of knowledge (e.g., exploiting higher k or the Full Basket attack) does not offer additional gain. Similar results were obtained for movie rating dataset in [6] and mobility data in [9], suggesting the existence of a possible general pattern in the behavior of privacy risk.

5 Conclusion

In this paper we proposed a framework to assess privacy risk in retail data. We explored a set of re-identification attacks conducted on retail data structures, analyzing empirical privacy risk of a real-world dataset. We found, on average, a high privacy risk across the considered attacks. Our approach can be extended in several directions. First, we can expand the repertoire of attacks by extending the data structures, i.e., distinguishing among shopping sessions and obtaining a proper transaction dataset, or considering different dimensions for retail data, e.g., integrating spatio-temporal informations about the purchases. Second, it would be interesting to compare the distributions of privacy risk of different attacks through some similarity measures, such as the Kolmogorov-Smirnov test. A more general and thorough approach to privacy risk estimation can be found in [14] and it would be interesting to extend our framework with it's approaches. Another possible development is to compute a set of measures commonly used in retail data analysis and investigate how they relate to privacy risk. Finally, it would be interesting to generalize the privacy risk computation framework to

data of different kinds, from retail to mobility and social media data, studying sparse relation spaces across different domains.

Acknowledgment

Funded by the European project SoBigData (Grant Agreement 654024).

References

1. C. Alberts, S. Behrens, R. Pethia, W. Wilson. 1999. *Operationally Critical Threat, Asset, and Vulnerability Evaluation (OCTAVE) Framework, Version 1.0.* CMU/SEI-99-TR-017. Software Engineering Institute, Carnegie Mellon University.
2. M. Deng, K. Wuyts, R. Scandariato, B. Preneel, W. Joosen. *A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements.* Requir. Eng. 16, 1. 2011.
3. F. Giannotti, L. V. Lakshmanan, A. Monreale, D. Pedreschi, H. Wang. *Privacy-preserving mining of association rules from outsourced transaction databases.* IEEE Systems Journal, Volume 7, Issue 3. 2013.
4. A. K. Kant. *Dietary patterns and health outcomes.* Journal of the American Dietetic Association, Volume 104, Issue 4, 2004.
5. H. Q. Le, S. Arch-int, H. X. Nguyen, N. Arch-int. *Association rule hiding in risk management for retail supply chain collaboration.* Computers in Industry, Volume 64, Issue 7, 2013.
6. Narayanan A, Shmatikov V. *Robust de-anonymization of large sparse datasets.* IEEE Security and Privacy, 2008
7. G. Pauler, A. Dick. *Maximizing profit of a food retailing chain by targeting and promoting valuable customers using Loyalty Card and Scanner Data.* European Journal of Operational Research, Volume 174, Issue 2, 2006.
8. R. Pellungrini, L. Pappalardo, F. Pratesi, A. Monreale. *A data mining approach to assess privacy risk in human mobility data.* Accepted for pubblication in ACM TIST Special Issue on Urban Computing.
9. F. Pratesi, A. Monreale, R. Trasarti, F. Giannotti, D. Pedreschi, T. Yanagihara. *PRISQUIT: a System for Assessing Privacy Risk versus Quality in Data Sharing.* Technical Report 2016-TR-043. ISTI - CNR, Pisa, Italy.
10. S. J. Rizvi, J. R. Haritsa. *Maintaining data privacy in association rule mining.* VLDB 2002.
11. C. Rygielski, J.-C. Wang, D. C. Yen. *Data mining techniques for customer relationship management.* Technology in society 24.4 2002.
12. P. Samarati, L. Sweeney. *Generalizing Data to Provide Anonymity when Disclosing Information (Abstract).* PODS 1998.
13. G. Stoneburner, A. Goguen, A. Feringa. *Risk Management Guide for Information Technology Systems: Recommendations of the National Institute of Standards and Technology.* NIST special publication, Vol. 800. 2002.
14. V. Torra *Data Privacy: Foundations, New Developments and the Big Data Challenge.* Springer 2017.
15. Y. Xu, B. C. M. Fung, K. Wang, A. W. C. Fu and J. Pei. *Publishing Sensitive Transactions for Itemset Utility*, ICDM 2008.