Privacy Risk for Individual Basket Patterns

Roberto Pellungrini[1], Anna Monreale[1], and Riccardo Guidotti[1,2]

[1] University of Pisa, Largo B. Pontecorvo, 3, Pisa, Italy, name.surname@di.unipi.it,
[2] KDDLab, ISTI-CNR, Via G. Moruzzi, 1, Pisa, Italy, name.surname@isti.cnr.it

Abstract. Retail data are of fundamental importance for businesses and enterprises that want to understand the purchasing behaviour of their customers. Such data is also useful to develop analytical services and for marketing purposes, often based on individual purchasing patterns. However, retail data and extracted models may also provide very sensitive information to possible malicious third parties. Therefore, in this paper we propose a methodology for empirically assessing privacy risk in the releasing of individual purchasing data. The experiments on real-world retail data show that although individual patterns describe a summary of the customer activity, they may be successful used for the customer re-identifiation.

1 Introduction

Retail data are one of the most important source of information that enables commercial companies in understanding their customers behavior by analyzing their purchasing patterns. In the literature, many data mining methods have been proposed to extract customer patterns describing frequent itemsets [2], top-k frequent itemsets [29], regular itemsets [14]. All these individual purchasing models may enable not only the understanding of collective and individual behaviors, but also the development of data-driven services such as personal recommendation systems able to capture the customers' preferences.

Unfortunately, the analysis of retail data might lead to the inference of highly sensitive information about individuals. Thus, in the literature some works have addressed the problem of privacy issues in market basket data. Some of them proposed a methodology for the empirical privacy risk evaluation [20], while others proposed some approaches for guaranteeing privacy protection [15, 30]. However, all these works are focused on the study of the privacy issues in the released purchasing data, that is, they study the potential privacy risk related to the release of raw data collected from individuals. Instead, in this paper we propose to study the privacy risk assessment of individual purchasing models extracted from the purchasing data of individuals during analysis processes. Specifically, we identify two types of individual purchasing models: individual models composed by a single pattern and individual models composed by a set of patterns. Then, we define the privacy attack models and the methods for their simulation. Finally, we simulate these attacks on real-world retail data and we analyze the privacy risk distributions trying also to identify the properties of

bought items that can lead to customer re-identification by her patterns. The results show that, although individual patterns are models that abstract from the details of the raw data, they are able to capture peculiarities of the customer behavior which often lead to the customer re-identification.

The rest of the paper is organized as follows. In Section 2, we discuss the related work. Section 3 introduces the data models used for representing retail data. In Section 4, we define the privacy risk assessment methodology including the privacy attacks. Section 5 shows the results of our experiments and, finally, Section 6 concludes the paper.

2 Related Work

Customer profiling is a process widely used in economy since long time ago for direct marketing, site selection, and customer relationship management. The process of construction and extraction of a personal data model formed by personal patterns is generally referred to as *user profiling*. A user profile contains the systematic behaviors expressing the repetition of habitual actions, i.e., personal patterns. These patterns can be expressed as simple or complex indexes [10], behavioral rules [14], set of events [13], typical actions [28], etc. Profiles can be classified as individual or collective according to the subject they refer to [9,16]. An *individual* profile is built considering the data of a single person. This kind of profiling is used to discover the particular characteristics of a certain individual, to enable unique identification for the provision of personalized services. We talk about *collective* data models when personal data or individual models generated by individual profiling are aggregated without distinguishing the individuals.

With respect to market basket analysis, customer profiling can play today a very important role. Nowadays the market is characterized by being global, products and services are almost identical and there is an abundance of suppliers. Therefore, instead of targeting all the customers equally, a company can select only those customers who meet certain profitability criteria based on their individual needs and buying patterns [4]. To achieve this goal, the customers must be described by characteristics valuable for the business, like the demographic ones, the lifestyle, and the shopping habits. These targets can be reached through customer profiling. By knowing the profile of each customer, a company can treat a customer according to her individual needs and increase the lifetime value of the customer [4]. Furthermore, customer profiling is a key element which impacts into the decisions in product life cycle cost [7]. One of the first methodology proposed to analyzed shopping session is [3] where frequent patter mining rules are defined. In [1] is described a system exploiting these rules for building personal profiles on transactional histories. The profiles consists of a set of rules describing customers' behavior. However, this system requires a constant user feedback to assess the pattern validity and parameter setting. An automatic and parameter-free approach to derive personal patterns is proposed in [13]. An evolution of [13] that also consider the temporal dimension is described in [14]. In [31] the authors analyze customers' shopping behaviors with respect to both

on product profiles and customer profiles. The product profile is characterized by a set of features describing the product. The customer profile this time is an index expressing the level of interest in product features calculated using the product profiles. A two-stage clustering technique is used to find the group of customers that have similar interests and then extract rules from each cluster. In [10] the authors propose two indexes that consider the level of repetitiveness in both the basket composition and also in the temporal and spatial dimension of shopping purchases, i.e., when and where the customers go to the supermarket. Other forms of customer profiling on market basket data like those described in [11, 12] adopt ad vector based modeling.

In existing literature, the privacy risk for the sharing of retail data or customer's profiles is not considered. This is especially interesting considering the high amount of privacy related literature.

A vastly used privacy-preserving model, and one of the models of our choosing for this paper, is k-anonymity [23], which requires that an individual should not be identifiable from a group of size smaller than k based on a subset of her own attributes used to univocally identify her, called quasi-identifiers. In [5] the authors present a set of attacks on the k-anonymity model to prove it's possible weaknesses while in [34] a graph-attack method based on k-anonymity to defend from possible privacy attacks is proposed. More recently, in [19] the k-anonymity model has been used as a base to propose a privacy framework for the systematic simulation of privacy attacks, then applied to mobility data. For retail data very little has been done in terms of privacy risk assessment. In [21] authors propose a framework for anonymizing transactional data, and in [33] and [32] the authors propose various methods for privacy preserving data publishing with transactional and retail data.

For privacy risk assessment, a fundamental work is the LINDDUN methodology, presented in [6]. The LINDDUN framework for privacy threats analysis is largely based on the privacy threat modeling framework STRIDE [25] used in software-based systems. Other methods for privacy risk evaluation have been published recently such as in [27], where the authors elaborate an entropy-based method to evaluate the disclosure risk of personal data, trying to manage quantitatively privacy risks.

In this paper we use a well known technique to match records of different data-set known as distance based record linkage. This technique was first introduced in [17], and allows for the matching of records from different data-sets based on a measure of distance between records. Records that have minimal distance between each other are considered to belong to the same individual and are matched. Different variations of this technique have been used in privacy literature such as in [26], where the Mahalanobis distance is used for distance based record linkage.

3 Retail Data

Retail data is generally collected through membership programs: customers who wish to do so, voluntarily agree to such programs in order to receive some benefits through the use of a specific membership card, the data about their purchases is subsequently collected. The raw data of each individual is represented by baskets. A basket is a set of items purchased by the individual during a shopping session. We consider baskets with no repetitions, i.e., proper sets where items can appear only once. Therefore, and individual may have multiple baskets associated to her.

Definition 1 (Basket). *We define a basket (or transactions) b as a subset of items such that $\emptyset \subset b_i \subseteq I$ where $I = \{i_1, \ldots, i_D\}$ is the set of all D items.*

Definition 2 (Basket History). *We define the basket history $B_u = \{b_1, \ldots, b_N\}$ as the set of N baskets (or transactions) belonging to the individual u.*

Such data is usually used to perform analysis of various kind, from association rule mining [2] to clustering [8]. In this paper we focus on transactional clustering, as performed with the state-of-the-art algorithm TX-Means [13]. TX-Means is a parameter-free clustering method that follows a clustering strategy similar to TX-Means [18] designed for finding clusters in the specific context of transactional data. TX-Means automatically estimates the number of clusters and it also provides the *representative basket* of each cluster, which summarizes the pattern captured by that cluster. The representative baskets correspond to the centroids of the sub-clusters and are calculated adopting the procedure described in [8]. Therefore, the output of TX-means, consisting in the representative baskets, is a set of typical patterns that represent recurring purchasing behavior of each individual. Note that, TX-means is only one of the algorithms able to discover purchasing patterns. We point out that different algorithms may discover purchasing patterns capturing different properties. For example, a standard pattern mining algorithm as Apriori [2] is able to extract frequent patterns that differ from recurrent patterns. However, it requires the minimum support as parameter that, from a personal data analytics perspective [9], should be personally tuned of each user. Another example of pattern can represent the top-k frequent items. However, in all these cases a *pattern* may be modelled similarly to a set of baskets.

Definition 3 (Patterns). *We define as $P_u = \{p_1, p_2, \ldots, p_M\}$ the sets of patterns of the individual u, where each $p_i \subseteq I$ and I is the set of all D items.*

4 Privacy Risk Assessment Methodology

In literature there are several notable methodologies proposed to assess privacy risks. The definition of privacy that we use was first introduced in [23]. To assess privacy risk we adopt the framework proposed in [22] that is also used in [19]. The basic assumption is that a malicious third party, commonly referred to as

the *adversary*, gathers some background knowledge about an individual, i.e., a subset of the information related to the individual. Then, the adversary tries to re-identify the individual in a published data-set using that background knowledge. If successful, the adversary could then be able to retrieve the complete information associated to the individual, i.e., the adversary could gain access to all the records regarding the individual. Thus, the general approach in applying this framework is to first determine the possible background knowledge of an adversary, then simulate an attack on the data using such background knowledge, empirically compute the privacy risk, and finally explore and analyze the results to assess privacy risk.

In order to understand the nature of privacy risk in retail data we define a set of attacks based upon the above framework to explore the privacy risk in this kind of data.

Patterns Against Patterns In the first attack we consider an adversary who tries to understand how unique the individual patterns extracted by clustering algorithms are. To this end, we conducted our study on two types of individual purchasing patterns, extracted by using two different clustering algorithms. The first one is a very simple baseline approach that for each individual u extracts a single pattern consisting in the set of her most frequent k items. In other words, for each individual u we have only one pattern in P_u, i.e., $p = \{i_1, i_2, \ldots, i_k\}$. In the rest of the paper we refer to this patterns as simple patterns. The second approach is the state-of-the-art clustering algorithm, TX-Means [13]. Using this more complex approach every customer can be characterized by a different number of patterns. Every pattern $p_j \in P_u$ corresponds to a *representative basket* extracted by TX-Means. In the rest of the paper we refer to this patterns as TX-means patterns. A representative basket is a virtual transaction that approximates a set of similar baskets, therefore capturing the items that best characterize it, i.e., the typical combination of items expected to appear in any of its baskets. Then, we define an attack where an adversary gathers a certain number of the patterns for each individual and tries to re-identify the individual in the whole set of published patterns.

For the first approach, the privacy risk of an individual is given by the number of other individuals sharing the same pattern.

Definition 4 (Single Pattern Risk). *Given an individual u with a single pattern in P_u, we define her privacy risk as: $Risk_u = \frac{1}{|M_{P_u}|}$, where $|M_{P_u}|$ is the cardinality of the set of individuals having the same pattern in P_u. This measure ranges from 0 to 1.*

For the second approach, where multiple patterns belong to the same individual, we relied on a systematic exploration of all the possible background knowledge of a certain length h. For instance, if a customer has 3 patterns $\{p_1, p_2, p_3\}$ and we assume an adversary knows 2 of them, we calculate the privacy risk exploring all the possible combinations of the 3 patterns with length 2. In the above example, the following three background knowledge would be used: *(i)* $\{p_1, p_2\}$, *(ii)* $\{p_1, p_3\}$, *(iii)* $\{p_2, p_3\}$. Each combination is compared with all the

patterns in the published dataset, i.e., we check how many customers have the same patterns in the data.

Definition 5 (Multiple Patterns Risk). *Let u an individual with multiple patterns in P_u and let C_h be the set of possible combinations of patterns with length h. The customer privacy risk is defined as: $Risk_u = \frac{1}{min_c(|M_c|)}$, where M_c is the set of customers having a particular combination of patterns $c \in C_h$. This measure ranges from 0 to 1.*

This is a worst-case based approach, as we use the most unique patterns to calculate the risk, given by the use of minimum value of $|M_c|$.

Patterns Against Baskets In the definition of the second attack we assume that an adversary might get access to the patterns dataset $\mathcal{P} = \{P_{u_1}, \dots P_{u_U}\}$ and use it to attack the basket history data $\mathcal{B} = \{B_{u_1}, \dots B_{u_U}\}$, where U is the number of different customers. This could happen for example in the case when the patterns are publicly made available because considered safe, and the adversary gets access to the anonymized basket history data. In this case, we cannot directly compare the pattern of an individual with the customer baskets to find a match, but we need to identify the possible basket history $B_i \in \mathcal{B}$ that could have generated the known pattern $P_i \in \mathcal{P}$. Thus, we should link the different basket histories in $\mathcal{B}$ with each pattern in $\mathcal{P}$ through the use of a distance measure. In particular, we propose to use the distance function introduced in [17]. The adversary will match each pattern in $\mathcal{P}$ with the closest basket history in $\mathcal{B}$. Clearly, if the distance between the pattern of the customer u in $\mathcal{P}$ and the basket history of u is the minimum, then the two records of that customer are correctly matched.

We recall that the set of the representative patterns of each individual is computed with either TX-means or the baseline approach. To calculate the distance between this the records in the data to be matched we propose to use a modified version of the Jaccard distance.

Definition 6 (Jaccard Distance). *Let A and B be two sets. The Jaccard distance is defined as: $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$.*

Definition 7 (Minimum Jaccard). *Let A and $Y = \langle b_1, b_2, \dots, b_m \rangle$ be a set and a set of sets respectively. The Minimum Jaccard distance is defined as: $MJ(A, Y) = min_{i=1,2,\dots,m}(J(A, b_i))$*

Definition 8 (Best Jaccard). *Let $X = \langle a_1, a_2, \dots, a_n \rangle$ and $Y = \langle b_1, b_2, \dots, b_m \rangle$ be two set of sets, with $n \leq m$. The Best Jaccard distance is defined as: $BJ(A, Y) = \sum_{i=1}^{n} MJ(a_i, Y)$*

Using the Best Jaccard distance, we can calculate the number of correct matches that an adversary could make using the pattern dataset to attack the basket history dataset. Now, we are ready to introduce the definition of the privacy risk in this particular setting.

Definition 9 (Patterns Against Baskets Risk). *Let U be the set of all individuals and M be the set of individuals for whom $BJ(P_u, B_u)$ has the minimum value. Then, we define the privacy of the dataset as: $Risk = \frac{|M|}{|U|}$. This measure ranges from 0 to 1.*

This approach dates back to [24]. Note that, in this case, we cannot directly express a measure for individual risk, since an adversary either correctly matches two records of the same individual or doesn't.

5 Experiments

We performed experiments on real world dataset provided by UniCoop Tirreno, a large Italian supermarket chain. Customers are provided with a loyalty card which allows to link different shopping sessions, and therefore reconstruct their personal shopping history. We analyzed a dataset of 2,021,414 shopping sessions, i.e., baskets, performed by 8564 individuals between the 2010 and 2012 in Leghorn province. These customers are "loyal customers", i.e., customers active in at least ten months every year. For each customer we have on average 240 baskets, containing 100 different items, and the average basket length is 8 items.

For each customer we extracted her typical patterns using the two approaches discussed previously in Section 4. Using the baseline approach for the patterns extraction, we obtained patterns considering the k-most frequent items for each person, with k ranging from 1 to 10. Applying TX-Means we extracted a total of 38,068 patterns, more than 4 patterns per individual on average.

5.1 Patterns Against Patterns

In this section we analyze the empirical results related to the privacy risk for the patterns against patterns attack.

Simple Patterns Against Simple Patterns Risk The first experiment that we performed is the simulation of the patterns against patterns attack using simple patterns, i.e., the top k items by frequency for each individual.

In Figure 1 we show the distribution of privacy risk for this attack using the baseline approach, by increasing the value of k, i.e., increasing the number of items in the k-most frequent patterns. We observe that, with 2 items (Figure 1 (a)), we have a lower distribution of the privacy risk. But increasing the number of known items, the level of risk increases rapidly. With 4 items (Figure 1 (c)), more than half of the population shows risk 1, i.e. maximum risk. Beyond $k = 5$ the risk becomes 1 for more than 95% of the population. Starting from the different top-k items of each individual for any value of k, we analyzed the length of the shortest simple pattern of each individual that yields privacy risk 1. The idea is to understand for the customers the distribution of risky k values.

Figure 2 reports the result of this analysis. We found a rather classical Gaussian distribution, with a peak around 4 as expected. Moreover, we also tried to

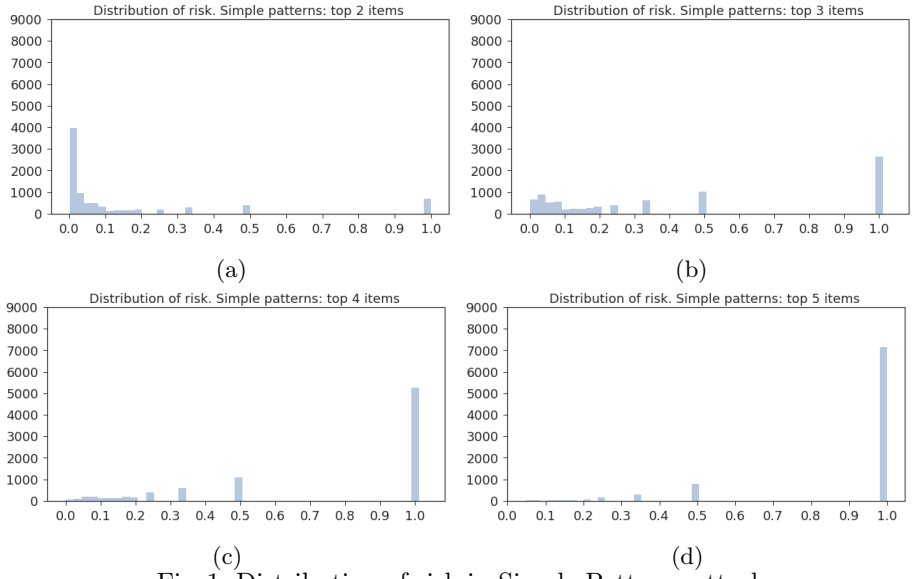
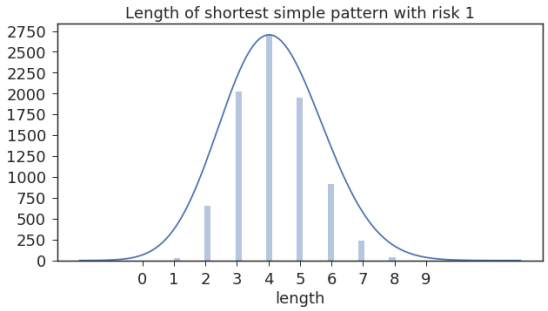


Fig. 1: Distribution of risk in Simple Patterns attack



	length
mean	4.150514
std	1.243128
min	1.000000
max	9.000000

Fig. 2: Distribution of the length of the shortest simple patterns that yield risk 1

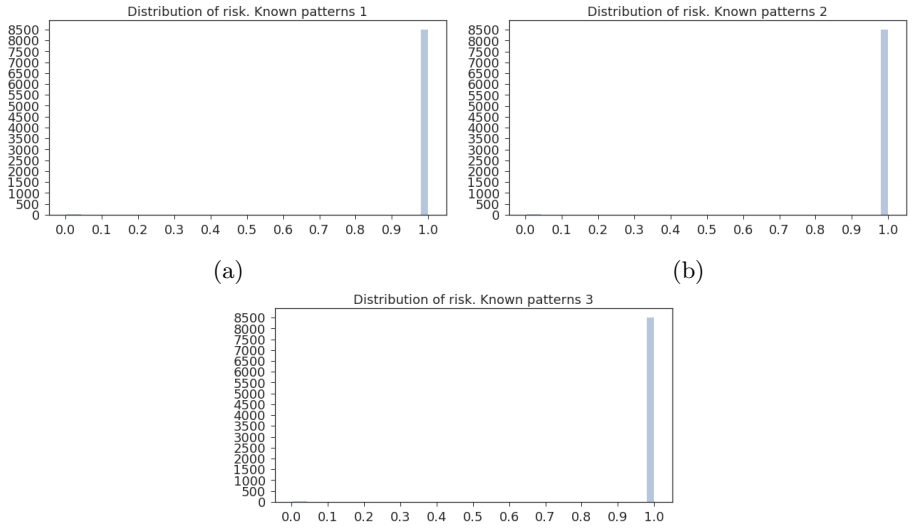
characterized the risky top-k items. To this end, for each customer we selected the shortest pattern that yield risk 1 and among the item composing them we identify those having the lowest global frequency in the basket history data and the lowest frequency in the set of top-k patterns. In practice, these items are bought by very few customers but are very frequent in the basket history of their customers. Given this property they probably are the cause of the customer high privacy risk. In Table 1 we report the list of the 10 items with lowest global frequency that appear in a low number of simple patterns. We observe that they are very particular items and most of them are not food items.

TX-means Patterns Against TX-means Patterns Risk The second experiment is focused on the simulation of a patterns-against-patterns attack using

Macro-sector	Category
No Food	Deodorants for environments
Grocery	Honey
No Food	Hardware
No Food	Anniversary card
Fresh food	Fruit beverages
No Food	Woman's socks
No Food	Sandpaper
Fresh food	Sheep meat
No Food	Flowers
No Food	Chemical products

Table 1: Infrequent items within the simple patterns

the individual models extracted with the TX-means algorithm. Each individual is hence represented by multiple patterns. To compute the privacy risk we checked all possible combinations of patterns of length h, with h values ranging from 1 to 3. We report the results in Figure 3. We can see that changing the value



(a) (b)

(c)

Fig. 3: Distribution of risk in TX-means patterns attack

of h does not impact on the level of risk as with just one pattern (Figure 3 (a)), it is possible to correctly re-identify more than 99% of the individuals. This means that almost every individual has at least one unique pattern that represents him. This is not surprising, since TX-means is an advanced algorithm for

personal data analytics and yields highly personalized results. We can further explore the results by looking at the length of the patterns and the privacy risk that they yield.

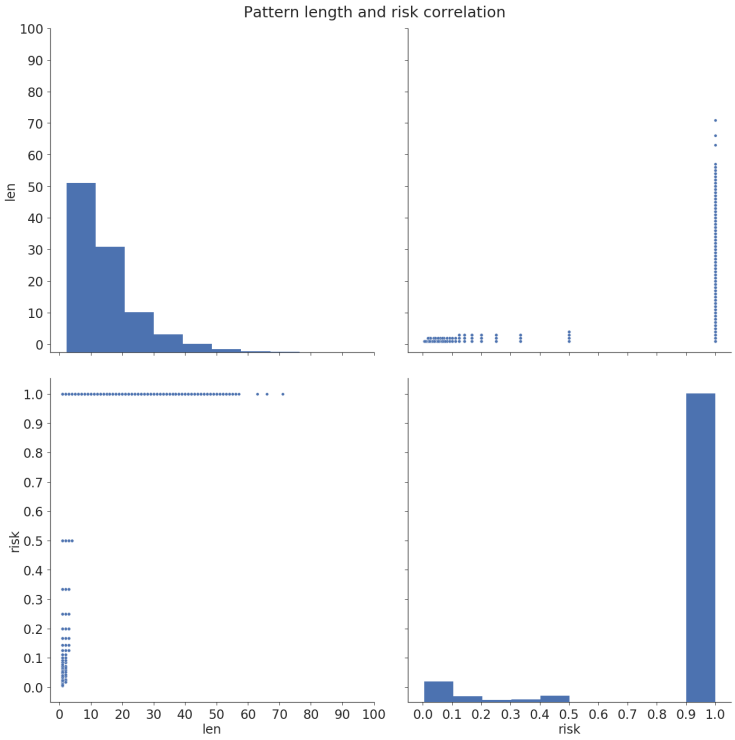
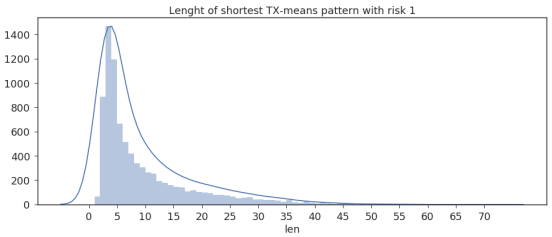


Fig. 4: Correlation between pattern length and privacy risk

Figure 4 highlights that there is no clear correlation between privacy risk values and pattern length. However, we observe that there is no pattern with length greater than 5 that yield a risk lower than 1. As for simple patterns, this suggests that longer and more complex patterns are more unique and personal; as a consequence, they lead to the identification of the individuals. For the TX-means patterns we performed the same analysis already presented for simple patterns; in other words, we analyzed the distribution of the length of the shortest pattern that for each individual yields the maximum risk.

We observe that TX-means provides longer patterns on average and the distribution presents a typical long tail shape. In Table 2 we report the list of the 10 items with lowest global frequency that appear in a low number of TX-means patterns. As for the simple patterns, we highlight that most of them are not food items but their categories are more common with respect to the simple patterns. Overall these experiments suggests that representative patterns extracted with either naive or advanced techniques are inherently unique. An individual may



	length
mean	9.046007
std	8.473491
min	1.000000
max	71.000000

Fig. 5: Distribution of the length of the shortest TX-means patterns that yield maximum risk

Macro-sector	Category
No Food	Manual tools
Fresh Food	Frozen meat
Fresh Food	Poultry for birds and rabbits
Fresh Food	Milk
No Food	Christmas decoration
No Food	Underwater Gear
No Food	Electrical equipment
No Food	House carpets
No Food	House decoration
No Food	Glasses

Table 2: Infrequent items within the TX-means patterns

be easily re-identifiable using these patterns even with a small number of items. As for the items themselves we see a fairly broad characterization, however, we can conclude that non-food related items are much more distinctive and may lead to higher chances of re-identification.

5.2 Patterns Against Baskets

In this section we analyze the empirical privacy risk in case of the patterns against baskets attack.

Simple Patterns Against Baskets The first experiment is based on the simulation of a patterns against baskets attack using simple patterns. We recall that for this attack risk is evaluated globally for the entire data-set and not individually. We performed distance based record linkage with simple patterns of 2, 4 and 5 items. For simple patterns of length 2 we have only 27 correct matches out of the total population of 8,564 customers. This yields a risk of 0.003. For simple patterns of length 4 we have 298 correct matches, yielding a risk of 0.034. For patterns of length 5 we have 388 correct matches, yielding a risk of 0.045. These low values are probably due to several factors: while we have shown previously that simple patterns are quite unique, they are not particularly representative

of the individual's baskets. Also, having only one pattern significantly diminishes the information used for the linkage. Because of how we compute distance, having only one simple pattern implies that such distance fall in the range 0 to 1. This leads to a high number of individuals with minimum distance, therefore impeding a univocal matching. We can conclude that simple patterns pose a relatively low threat when used to attack the raw data.

TX-means Patterns Against Baskets The second experiment is based on the simulation of a patterns against baskets attack using the patterns extracted with the TX-means clustering algorithm. As for the previous case, the risk is calculated for the entire data-set. With the TX-means patterns we have that 5,781 individuals out of the total population of 8,564 customers are correctly matched, i.e., the distance between the TX-means patterns of those individuals and their basket data is minimal. This yields a risk of 0.675. We can now characterize the individuals correctly matched, by looking at their patterns and baskets.

	Patterns: std of length	Patterns: mean length	Number of patterns	Number of baskets	Baskets: std of length	Baskets: mean length
mean	4.811004	13.049558	4.820446	244.230064	6.002396	10.897940
std	3.996948	7.899513	3.453788	201.790281	2.873166	5.264362
min	0.000000	2.200000	1.000000	10.000000	0.708363	1.744063
max	26.051631	71.000000	25.000000	1646.000000	26.411782	43.282051

Table 3: Characterization of matched individuals in the TX-means patterns against baskets attack

	Patterns: std of length	Patterns: mean length	Number of patterns	Number of baskets	Baskets: std of length	Baskets: mean length
mean	2.884773	10.653819	3.665469	219.015451	4.884745	8.000338
std	3.385043	7.122223	3.735964	220.721776	2.372840	3.951333
min	0.000000	1.000000	1.000000	10.000000	0.535428	1.221429
max	25.500000	53.000000	26.000000	2025.000000	16.146130	31.976744

Table 4: Characterization of non matched individuals in the TX-means patterns against baskets attack

In Table 3 and Table 4 we gathered some statistics for the individuals correctly matched and those who were not matched. For each individual, we gathered the mean length of her patterns and her baskets as well as the standard deviation for such lengths and the number of patterns and baskets. In the tables

we show mean, standard deviation, min value and max value for the aforementioned measures. If we compare the statistics in the two table we can see that there are not many differences. However, we observe that, for the individuals that were not re-identified by the attack, we have fewer, shorter patterns and baskets on average, again, confirming that higher risk is related to lengthier baskets and/or patterns.

6 Conclusion

In this paper we have studied the privacy risk assessment of individual purchasing patterns. In the study we have taken into consideration two different individual patterns: the top-k items of an individual and the representative patterns extracted by TX-means. After defining, two possible attacks that exploit individual patterns for customers re-identification, we have performed their simulation on real-world data. The empirical results on the privacy risk distributions show that individual patterns often lead to the re-identification of most of the customers because they accurately describe some customer habits that make him unique. This preliminary study suggests the need of the application of privacy-preserving methods for guaranteeing the privacy protection during the analysis and publishing of individual patterns. An interesting future work would involve the study of privacy methods that exploit the knowledge provided by the risk assessment methodology for reducing the model perturbations.

Acknowledgments. Work partially supported by the EU H2020 Program under the funding scheme "INFRAIA-1-2014-2015: Research Infrastructures", grant agreement 654024 *"SoBigData"* (`http://www.sobigdata.eu`).

References

1. G. Adomavicius and A. Tuzhilin. Using data mining methods to build customer profiles. *Computer*, 34(2):74–82, 2001.
2. R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, SIGMOD '93, pages 207–216, New York, NY, USA, 1993. ACM.
3. R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.
4. H. Andersen, M. Andreasen, and P. Jacobsen. The crm handbook: From group to multi-individual. *Norhaven: PricewaterhouseCoopers*, 1999.
5. S. De Capitani Di Vimercati, S. Foresti, G. Livraga, and P. Samarati. Data privacy: Definitions and techniques. 20:793–817, 12 2012.
6. M. Deng, K. Wuyts, R. Scandariato, B. Preneel, and W. Joosen. A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. *Requir. Eng.*, 16(1):pp 3–32, March 2011.
7. A. S. Dunk. Product life cycle cost analysis: the impact of customer profiling, competitive advantage, and quality of is information. *Management Accounting Research*, 15(4):401–414, 2004.

8. F. Giannotti, C. Gozzi, and G. Manco. Clustering transactional data. In *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, PKDD '02, pages 175–187, London, UK, UK, 2002. Springer-Verlag.
9. R. Guidotti. Personal data analytics: capturing human behavior to improve self-awareness and personal services through individual and collective knowledge. 2017.
10. R. Guidotti, M. Coscia, D. Pedreschi, and D. Pennacchioli. Behavioral entropy and profitability in retail. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–10. IEEE, 2015.
11. R. Guidotti and L. Gabrielli. Recognizing residents and tourists with retail data using shopping profiles. In *International Conference on Smart Objects and Technologies for Social Good*, pages 353–363. Springer, 2017.
12. R. Guidotti, L. Gabrielli, A. Monreale, D. Pedreschi, and F. Giannotti. Discovering temporal regularities in retail customers' shopping behavior. *EPJ Data Science*, 7(1):6, 2018.
13. R. Guidotti, A. Monreale, M. Nanni, F. Giannotti, and D. Pedreschi. Clustering individual transactional data for masses of users. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 195–204, New York, NY, USA, 2017. ACM.
14. R. Guidotti, G. Rossetti, L. Pappalardo, F. Giannotti, and D. Pedreschi. Market basket prediction using user-centric temporal annotated recurring sequences. In *Data Mining (ICDM), 2017 IEEE International Conference on*, pages 895–900. IEEE, 2017.
15. L. Guo, S. Guo, and X. Wu. Privacy preserving market basket data analysis. In J. N. Kok, J. Koronacki, R. Lopez de Mantaras, S. Matwin, D. Mladenič, and A. Skowron, editors, *Knowledge Discovery in Databases: PKDD 2007*, pages 103–114. Springer Berlin Heidelberg, 2007.
16. M. Hildebrandt. Defining profiling: a new type of knowledge? In *Profiling the European citizen*, pages 17–45. Springer, 2008.
17. D. Pagliuca and G. Seri. Some results of individual ranking method on the system of enterprise accounts annual survey. *Esprit SDC Project, Deliverable MI-3 D*, 2:1999, 1999.
18. D. Pelleg, A. W. Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In *Icml*, volume 1, pages 727–734, 2000.
19. R. Pellungrini, L. Pappalardo, F. Pratesi, and A. Monreale. A data mining approach to assess privacy risk in human mobility data. *ACM Trans. Intell. Syst. Technol.*, 9(3):31:1–31:27, Dec. 2017.
20. R. Pellungrini, F. Pratesi, and L. Pappalardo. Assessing privacy risk in retail data. In *Personal Analytics and Privacy. An Individual and Collective Perspective - First International Workshop, PAP 2017, Held in Conjunction with ECML PKDD 2017, Skopje, Macedonia, September 18, 2017, Revised Selected Papers*, pages 17–22, 2017.
21. G. Poulis, G. Loukides, A. Gkoulalas-Divanis, and S. Skiadopoulos. Anonymizing data with relational and transaction attributes. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III*, pages 353–369, 2013.
22. F. Pratesi, A. Monreale, R. Trasarti, F. Giannotti, D. Pedreschi, and T. Yanagihara. Prisquit: a system for assessing privacy risk versus quality in data sharing. Technical report.
23. P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In *Proceedings of the Seventeenth ACM SIGACT-*

SIGMOD-SIGART Symposium on Principles of Database Systems, PODS '98, pages 188–, New York, NY, USA, 1998. ACM.
24. N. Spruill. The confidentiality and analytic usefulness of masked business microdata. *Proceedings of the Section on Survey Research Methods, 1983*, pages 602–607, 1983.
25. F. Swiderski and W. Snyder. *Threat Modeling*. O'Reilly Media, 2004.
26. V. Torra, J. M. Abowd, and J. Domingo-Ferrer. Using mahalanobis distance-based record linkage for disclosure risk assessment. In J. Domingo-Ferrer and L. Franconi, editors, *Privacy in Statistical Databases*, pages 233–242, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.
27. S. Trabelsi, V. Salzgeber, M. Bezzi, and G. Montagnon. Data disclosure risk evaluation. In *CRiSIS '09*, pages 35–72.
28. R. Trasarti, R. Guidotti, A. Monreale, and F. Giannotti. Myway: Location prediction via mobility profiling. *Information Systems*, 64:350–367, 2017.
29. V. S. Tseng, C. Wu, P. Fournier-Viger, and P. S. Yu. Efficient algorithms for mining top-k high utility itemsets. *IEEE Trans. Knowl. Data Eng.*, 28(1):54–67, 2016.
30. L. Wang and X. Li. Personalized privacy protection for transactional data. In *Advanced Data Mining and Applications - 10th International Conference, ADMA 2014, Guilin, China, December 19-21, 2014. Proceedings*, pages 253–266, 2014.
31. S.-S. Weng and M.-J. Liu. Feature-based recommendations for one-to-one marketing. *Expert Systems with Applications*, 26(4):493–508, 2004.
32. Y. Xu, B. C. M. Fung, K. Wang, A. W. Fu, and J. Pei. Publishing sensitive transactions for itemset utility. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, pages 1109–1114, 2008.
33. Y. Xu, K. Wang, A. W. Fu, and P. S. Yu. Anonymizing transaction databases for publication. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, pages 767–775, 2008.
34. R. Yarovoy, F. Bonchi, L. V. S. Lakshmanan, and W. H. Wang. Anonymizing moving objects: how to hide a mob in a crowd? In *EDBT*, pages 72–83, 2009.