Quantifying Location Privacy Leakage from Transaction Prices

Arthur Gervais, Hubert Ritzdorf, Mario Lucic, Srdjan Čapkun
Institute of Information Security
ETH Zurich
{firstname.lastname}@inf.ethz.ch

Abstract

Large-scale datasets of consumer behavior might revolutionize the way we gain competitive advantages and increase our knowledge in the respective domains. At the same time, valuable datasets pose potential privacy risks that are difficult to foresee. In this paper we study the impact that the prices from consumers' purchase histories have on the consumers' location privacy. We show that using a small set of low-priced product prices from the consumers' purchase histories, an adversary can determine the country, city, and local retail store where the transaction occurred with high confidence. Our paper demonstrates that even when the product category, precise time of purchase, and currency are removed from the consumers' purchase history (e.g., for privacy reasons), information about the consumers' location is leaked. The results are based on three independent datasets containing thousands of low-priced and frequently-bought consumer products. In addition, we show how to identify the local currency, given only the total price of a consumer purchase in a global currency (e.g., in Bitcoin). The results show the existence of location privacy risks when releasing consumer purchase histories. As such, the results highlight the need for systems that hide transaction details in consumer purchase histories.

1 Introduction

Making data publicly available creates unexpected privacy risks. Recent examples include AOL's release of users' search keywords [33], which has led to the identification of users and their profiles [1]. Data released by Netflix was de-anonymized by leveraging IMDB and dates of user ratings [31], showing that the release of data cannot be analyzed in isolation. The privacy risks of combining different public records have led to several [39] de-anonymization attacks. Recent studies of anonymized mobility data showed that mobility traces can be de-anonymized by leveraging a few observations [21]. One source of consumer information involves their spending patterns. To date however, it was unclear to what extent consumer prices leak information about the respective purchase.

Consumer purchase histories are typically recorded by store chains with loyalty programs and are used to compute consumer spending profiles [7]. Banks, payment card issuers, and point-of-sale system providers collect this data at different levels of granularity. In a number of scenarios, it might be desirable to share this data within different departments of a company, across companies, or with the public [8]. Before disclosure, the data is sanitized so that it does not leak sensitive data, such as personally identifiable information and that it (partially or fully) hides location information. In new digital currency systems such as Bitcoin [36] and Ripple [11], transaction values are stored on a public ledger. Irrespective of whether transaction values are made available so that a system can fulfill its functions or are being disclosed for research purposes, it is important to understand the privacy implications of such disclosures.

In this paper we focus on quantifying location disclosure resulting from the release of prices from consumer's purchase histories. Intuitively, the price distribution for a product differs from country to country (cf. Figure 17 in appendix), which allows us to identify possible purchase locations. We focus on consumer products which are generally inexpensive ($\leq$ 25 USD) and frequently-bought. More precisely, based on global prices (leveraging the Numbeo dataset [10]), we show that given access to a few consumer prices (and even without the product categories, precise times of purchase or currency), an adversary can determine the country in which the purchase occurred. Similarly, given the country, the city can be determined and within a city (leveraging the Chicago dataset [13]), the *local store* can be identified. We further demonstrate that it is possible to distinguish purchases among store chains (leveraging the Kaggle dataset [8]).

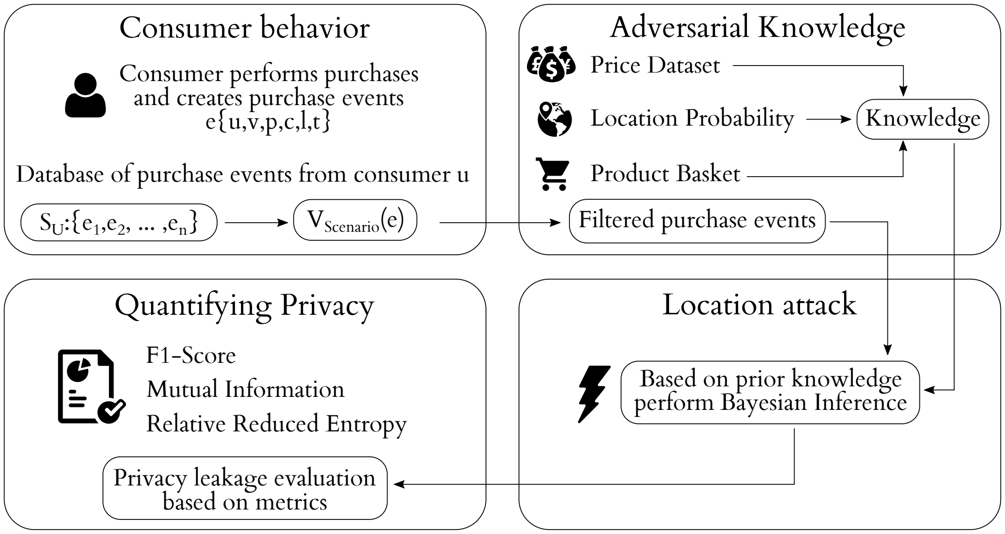

Figure 1: Overview of our framework for quantifying location privacy leakage from consumer price datasets.

We present a generic framework (cf. Figure 1) that allows the modeling and quantitative evaluation of location leakage from consumer price datasets. In our framework we model the adversarial knowledge, composed of a public dataset of consumer prices and location-specific information. We assume that the adversary has access to the individual product prices of a purchase (similar to the Kaggle dataset) and a coarse-grained value of the purchase time. In order to make the framework more flexible, our model supports different prior knowledge scenarios, e.g., the adversary additionally has access to the merchant category (e.g., knowledge that the product was bought in a market or a restaurant) or the product category (e.g., apples). Furthermore, we model the adversarial attack by detailing the corresponding probability functions. In particular, we point out how the adversary leverages multiple product prices in order to increase the probability of identifying the correct location.

Within our framework, we quantify the location privacy of consumer purchases in relation to different dimensions. For example, we measure how well the adversary estimates the location probability of the purchases with the F_1-score [38], capturing the test's accuracy. Furthermore, we use mutual information [20] to quantify the absolute location privacy loss of consumers, based on the considered price dataset. In addition, we capture the relative privacy loss by measuring the reduction in entropy. The proposed metrics are independent of the choice of adversarial strategy and therefore allow us to quantitatively measure the privacy loss induced from any price dataset known to the adversary.

We apply our framework to three real-world datasets: (i) the *Numbeo dataset* [10] contains, after outlier filtering, crowd-sourced real-world consumer prices from 112 countries and 23 US cities for 23 distinct product categories; (ii) the *Chicago dataset* [13] contains 24 million prices for 28 product categories capturing on average of 6304 products sold in Dominick's stores within the Chicago metropolitan area; finally, (iii) the *Kaggle dataset* [8] contains 350 million purchases from 311,541 consumer across 134 store chains.

Our evaluation shows that in order to infer the country based on a vector of purchases, an adversary often needs to observe less than 30 prices. Similarly, after having identified the country of the purchases and given roughly 30 prices, we show that we can reliably predict among 23 major cities within the United States. Finally, when the adversary narrowed down the coarse location, such as the Chicago metropolitan area, we show that based on a regional price dataset, and given a vector of purchases, an adversary can distinguish with high confidence *among local stores* using 100 purchases. For comparison, a weaker adversary with access only to coarse-grained time, i.e., the day of the purchase and price information, requires 50 purchases to identify the country. Furthermore, to establish practical utility of our methodology, we evaluate it on a dataset of purchase records (Kaggle [8]) and show that an adversary requires approximately 250 purchases to distinguish with high confidence among 134 store chains.

In addition to our proposed framework, we show that the local currency of a purchase can be identified, when only the total purchase price is given (e.g., in Bitcoin). The currency is an indicator of the location of the purchase. When estimating the currency among 155 local currencies, based on purchases with up to 20 consumer products, we achieve an average F_1-score of over 0.45, while the baseline of random guessing is near zero.

The main contributions of this paper are as follows:

- We propose a generic quantitative framework for evaluating attacks against the location privacy of consumer purchases. We validate our framework on three independent price datasets of real-world consumer prices and show that location information can be extracted reliably.

- We introduce three privacy metrics to capture the performance of the adversary in the attack as well as the extent to which location privacy of consumers is reduced when the adversary has access to a specific dataset of purchases.

- We show that given the total price of a consumer purchase in a global currency (e.g., Bitcoin), we can extract knowledge about the native currency of the purchase.

To the best of our knowledge, this is the first work to infer the location and currency of a purchase based on the price value in consumer purchases.

The remainder of this paper is organized as follows. In Section 2, we model purchase history and describe the adversarial model. In Section 3, we present the datasets selected for our evaluation in Section 4. In Section 5, we present our methodology for identifying the currency based on the purchase price. We survey the related work in Section 6 and conclude the paper in Section 7.

2 Model

In this section we introduce our system and adversarial model. We present the privacy metrics that quantify the probability of location disclosure based on the assumption that the adversary has access to a part of a consumer's purchase history.

2.1 System Model

A consumer interacts with merchants and performs purchases of one or more products. This interaction leaves a trace of purchase activity as a sequence of *purchase events*. We model each of the consumer's purchase events together with their contextual information as e: {consumer u, value v, product p, product category c, location l, time t}, where v is the price value spent on product p of product category c at location l and time t. In our model, one purchase event is limited to one product, similar to the data contained in the Kaggle dataset. In addition, the price value is given in a global currency, which usually is different from the local currency of the purchase (e.g., the original price is SEK, but recorded in USD). The trace of purchases performed by the target consumer U, given as a series of purchase events, is denoted by S_U:$\{e_1, e_2, \dots, e_n\}$. We define the following functions to represent the adversarial knowledge:

LOCATION PROBABILITY: It describes the prior probability of a purchase event taking place in a specific location, e.g., $P(\text{USA})$ is the prior probability with which a random purchase event e has $e.l = \text{USA}$. We define $\mathbb{L}$ as the set of all considered locations.

CATEGORY PROBABILITY Given location l, $P(c \mid l)$ describes the conditional probability of a purchase event to belong to a certain product category, e.g., $P(\text{Milk} \mid \text{USA})$ is the conditional probability with which a random event e from the USA has $e.c = \text{milk}$. This conditional probability models the product category preferences in a location. We define $\mathbb{C}$ as the set of all considered product categories.

VALUE PROBABILITY: Given location l and product category c, $P(v \mid l, c)$ describes the conditional probability of a purchase event at a given price value. It models the price distributions for different product categories in different locations, e.g., $P(1.5 \mid \text{USA}, \text{Milk})$ is the conditional probability with which milk can be bought in the USA for 1.5 worth of a global currency.

The adversary can now model the spending behavior and identify likely candidate locations. Specifically, the adversary computes the posterior probability that a single price value v for a product category c originated from a location l. The computation involves the prior and the conditional probabilities described above and the application of Bayes' theorem:

$$P(l \mid c, v) = \frac{P(l) \cdot P(c, v \mid l)}{P(c, v)} \tag{1}$$

In order to infer the location without knowing the product category, the adversary computes the probability that a price value v originates from location l:

$$P(l \mid v) = \frac{P(l) \cdot P(v \mid l)}{P(v)} \tag{2}$$

2.2 Adversarial Model

The adversary's goal is to identify the location of the events in S_U. In this section we present two different adversaries: (1) an adversary with complete knowledge and (2) an adversary with only public knowledge.

2.2.1 Adversary with Complete Knowledge

The ideal adversary represents a strong adversary with complete access to global purchase events. In particular, the adversary has access to the following prior knowledge:

GLOBAL PURCHASE HISTORY: The complete series of purchase events in the history of global purchases[1], denoted by $\mathscr{H}_G$. The adversary computes the posterior probability of a location based on $\mathscr{H}_G$.

HISTORY FOR TARGET CONSUMER: The adversary might have access to prior information about the target consumer's purchase history, denoted by $\mathscr{H}_U$. This could help the adversary to optimize the model for the target consumer[2].

Based on this knowledge, the ideal adversary computes the probabilities in Equations 1 and 2.[3]

2.2.2 Adversary with Public Knowledge

Our second adversarial model is a more realistic one, where the adversary only makes use of public information.

POPULATION: Given the population at each location, the adversary estimates the location probability $P(l)$.

PRODUCT BASKET: A product basket indicates which products an average consumer purchases during a year, both in terms of quantity and monetary amount. We leverage the product basket in order to estimate the probability of a product category given the location ($P(c \mid l)$)[4].

PRICE DATASET: For each location and product category combination, a price value distribution D is available, e.g., the Numbeo or the Chicago dataset. The adversary can use the distribution to estimate $P(v \mid l, c)$. We define $D(l, c, v)$ as the number of occurrences of price value v for product category c in location l and $D(l, c)$ as the number of price values for product category c and location l.

Since D might be imperfect, the adversary can have incomplete or incorrect knowledge about the price value probabilities (i.e. unknown or rounded product prices). In this case the adversary should perform additive smoothing, which assigns a small probability α to each event [29]. On the contrary, if the adversary has or assumes complete knowledge of the price value probabilities, additive smoothing is not required.

The adversary with public knowledge computes the following probabilities:

$$P(l) = \frac{\text{Population}(l)}{\sum_{l' \in \mathbb{L}} \text{Population}(l')} \tag{3}$$

$$P(c \mid l) = \frac{\text{Basket}(l, c)}{\sum_{c' \in \mathbb{C}} \text{Basket}(l, c')} \tag{4}$$

$$P(v \mid l, c) = \frac{D(l, c, v) + \alpha}{D(l, c) + \alpha \cdot |S_U|} \tag{5}$$

In order to compute the probabilities defined earlier in Equations 1 and 2, the adversary requires access to either $P(l \mid c, v)$ or $P(l \mid v)$. Next, we describe how the adversary computes these probabilities and we define the adversary's knowledge.

2.3 Knowledge Scenarios

As mentioned, the adversary's objective is to identify the location of the events in S_U. The adversary is given a finite set of events S_U on which the attack is executed—the adversary is not allowed to choose or request new purchase events e. We consider an adversary with public knowledge and distinguish among three distinct adversarial knowledge scenarios, each consisting of a subset of the public knowledge. Depending on the knowledge scenario, the adversary might not have access to all information from a purchase event e. Therefore, we define a family of functions $V_{\text{scenario}}(e) = V(e)$ that filter, depending on the given scenario, the public knowledge accessible to the adversary.

PRICE: This scenario corresponds to an adversary that has access to multiple purchase events e, *only* the corresponding *price value* and a notion of the purchase time $e.t$. The adversary is not aware of the product $e.p$ or the product category $e.c$. The precision of the purchase time depends on further specifications of the scenario. More formally, $V_{\text{price}}(e) = \{e.v, e.t\}$. Given the public knowledge modeled by Equations 3, 4 and 5, the adversary computes the posterior probability $P(l \mid v)$ of a price value v from location l. The intermediate steps for computing $P(v \mid l)$ and $P(v)$ are detailed in the appendix A in Equations 12 and 10.

PRICE_MERCHANT: Similar to the former knowledge scenario, the adversary here has access to S_U, a series of multiple purchase events. In this scenario, however, the adversary knows the price value $e.v$ of the event as well as which merchant category m sold the product. Formally, for each purchase event e, $V_{\text{price\_merchant}}(e) = \{e.v, e.t, m\}$, where $V_{price\_merchant}$ requires a function $M(e) = m$. We consider three merchant categories: restaurant, market and local transportation.

[1] The area of the attacker's interest can be restricted, e.g., when the adversary knows that its victim is somewhere in that restricted area.

[2] For example, by only considering the locations of previous purchases.

[3] The intermediate steps are given in the appendix A.

[4] We currently use a single product basket for all locations.

The $V_{\text{price\_merchant}}(e)$ function estimates the merchant category m from the product category $e.c$ of the respective event (cf. Table 8 for an overview).[5] Analogously, using Equation 1, the adversary computes the probability of a location, based on the merchant and the price value:

$$P(l \mid m, v) = \frac{P(l) \cdot P(m, v \mid l)}{P(m, v)} \tag{6}$$

where $P(m, v \mid l)$ is computed as follows:

$$P(m, v \mid l) = \sum_{c \in M^{-1}(m)} P(c, v \mid l) \tag{7}$$

PRICE_PRODUCT-CATEGORY: This scenario corresponds to the most knowledgeable adversary with public knowledge. Similarly to the former scenarios, the adversary receives multiple purchase events S_U. In addition, the adversary has access to the product category $e.c$ as well as the price value $e.v$. Note that $e.c$ implicitly assumes knowledge of the merchant. Formally, $V_{\text{price\_product-category}}(e) = \{e.v, e.t, e.c\}$.

Given the public knowledge described in Section 2.2, the adversary computes the probability $P(l \mid c, v)$ of a purchase event with product category c and price value v originating in location l. The intermediate steps for computing $P(c, v \mid l)$ and $P(c, v)$ are detailed in the appendix in Equations 13 and 11.

In the following section we provide an intuitive perspective on the probabilities $P(l \mid v)$ and $P(l \mid c, v)$.

2.4 Conditional probability intuition

$P(l \mid v)$ is the probability of a location, given a price value in a purchase event. An example plot based on our evaluation can be found in Figure 2. We have chosen the purchase event e with a price value of $e.v = 1$ Euro and estimated the location of the price. The figure shows that the most likely location for 1 Euro is France, closely followed by Germany, Italy and Spain. The plot also shows $P(l \mid c, v)$ for a purchase event with $e.v = 1$ Euro and the product category is milk. The most likely country is again France, followed by Germany and Italy. Surprisingly, China ranks as 5^{th}. This can be explained by the fact that (i) some prices from China in the dataset were erroneously reported in Euros and (ii) that the location probability $P(l)$ influences the overall outcome, and, since China's population is considerable, there is an increased probability of purchases occurring there. Overall we observed that the probability distribution changes when the product category is known, i.e., France is more likely to have a 1 Euro price for milk, than a 1 Euro price in general.

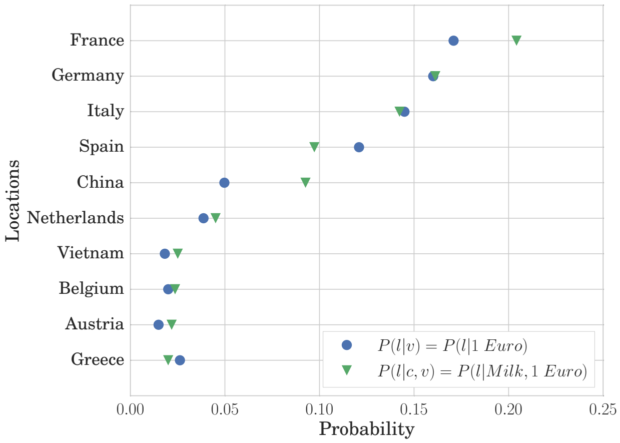

Figure 2: Probability distribution of $P(l \mid v)$ and $P(l \mid c, v)$, given 1 Euro and milk. France is the most likely location.

2.5 Multiple purchase events

Up to this point, the analysis has been based on a single purchase event. To naturally combine multiple purchase events, we assume that the purchase events are conditionally independent, given the location l. Therefore, the probability of a location l, given a set of purchase events S_U, is calculated as follows:

$$\begin{aligned} P(l \mid S_U) &= P(l \mid V(e_1), V(e_2), \ldots, V(e_n)) \\ &= \frac{P(l) \cdot \prod_{e \in S_U} P(V(e) \mid l)}{P(V(e_1), \ldots, V(e_n))} \end{aligned} \tag{8}$$

The intermediate steps for computing $P(l \mid S_U)$ can be found in the appendix in Equation 18. We experimentally verified the conditional independence of $V(e)$ given l for the three knowledge scenarios and therefore Equation 8 applies equally to the different adversarial knowledge scenarios. Note that we effectively weaken the adversary by considering the products of different purchases independent from each other.

2.6 Privacy Metrics

We introduce three privacy metrics in order to capture the privacy of consumers revealing their purchase histories across different dimensions: We (i) measure the performance of the adversary in identifying the true location with the F_1-score. Then, (ii) using the notion of mutual information [20], we quantify the absolute privacy loss of the consumer due to the adversary's knowledge of a price dataset. Finally, (iii) we use the relative reduced

[5] In the following we refer to the merchant category as merchant.

entropy as a relative privacy metric[6].

F_1-SCORE: The objective of the adversary is to assign the purchase events to the correct location. In the worst case, the adversary is forced to randomly guess among all possible locations. If the adversary, however, can estimate location probabilities more accurately, location privacy is reduced. Our problem corresponds to a multi-class classification problem and we therefore quantify the adversarial performance by averaging the F_1-score [38] of each individual class. The F_1-score corresponds to the harmonic mean of recall and precision, measuring the test's accuracy.

MUTUAL INFORMATION: A purchase event dataset enables the adversary to infer the distribution of prices among locations. Therefore, we want to measure how much privacy consumers lose when their purchase events are revealed and when the adversary has access to a dataset of purchase events. We quantify this privacy objective by measuring the *absolute* reduced location entropy given the purchase events. To this extent, we use the Mutual Information [20], denoted by $I(l,V(e))$, which measures how much the entropy of the locations is reduced given the purchase events (cf. Equation 9).

$$I(l,V(e)) = \sum_{l\in\mathbb{L},e\in S_U} P(l,V(e))\cdot\log_2\frac{P(l,V(e))}{P(l)P(V(e))} \quad (9)$$

RELATIVE REDUCED ENTROPY: Recall that the mutual information quantifies what we call the absolute privacy loss. In fact, there is an inherent randomness in the price distribution among locations. It is important to capture to what extent the original uncertainty about the locations can be reduced when a dataset of purchase events is given. The relative reduced entropy therefore captures the *relative* privacy, as the complement of the fraction of the conditional entropy over the location entropy. Given $H(l) = I(l,V(e)) + H(l \mid V(e))$, we compute the relative reduced entropy as $1-\frac{H(l|V(e))}{H(l)}$ over all purchase events.

The proposed evaluation metrics are independent of a particular adversarial strategy. In return, the output of the privacy leakage quantification only depends upon the employed dataset of purchase events. In the next section we present the datasets utilized for our experimental evaluation.

3 Datasets

There are only a couple of datasets accurately accumulating the worldwide product price information. For individual products (e.g., a Big Mac [6] or Starbucks coffee [9]), the average price values per country are available. Because a product often appears multiple times with different price values in the same country or city, the average is not a good estimator for elaborate studies. In the following, we describe the three independent price datasets considered in our work.

Symbol	Description
$\mathbb{C}$	Set of considered product categories
$\mathbb{L}$	Set of considered locations
e	Purchase event, e.g., element on a bill
v	Price value, e.g., the price of a product
u	Consumer id
p	Product, e.g., the name
c	Product category
m	Merchant: restaurant, market, transportation
l	A location
t	A timestamp
D	Price value dataset
n	Number of trials per location
α	Additive smoothing parameter
$V_{\text{scenario}}(e)$	Filters the content of e, given to the adversary
$M(c)$	Infers the merchant for a product category c
S_U	Series of purchase events
$H(X)$	Shannon Entropy of X
$I(X,Y)$	Mutual Information of X and Y
$\mathscr{H}_G$	History log of global purchase events
$\mathscr{H}_U$	History log of S_U of the target consumer

Table 1: Important notations used throughout the paper.

The first dataset, Numbeo [10], is a crowd-sourced dataset containing worldwide price values per product category, city and country. It is the most complete dataset of worldwide harvested prices available to our knowledge. We restricted our analysis to 23 frequently-bought product categories, listed in Table 8 and split the Numbeo dataset into two separate datasets: (i) two years of data as the Numbeo dataset and (ii) five months of data as the Numbeo test dataset (cf. Table 2). Numbeo performs sanity checks on the crowdsourced inputs, and we additionally filtered extreme outlier [4][7] from the data to account for possible mistakes from crowdsourced data. We identified 112 countries, with a total of 328,720 price values. Note that the provided data mostly contains prices from the US (18%) and India (14%).

The second dataset, referred to as the Chicago dataset [13], covers 84 stores in the Chicago metropolitan area over a period of five years. The data is sourced on a weekly basis from Dominick's supermarket stores. We sample 85 weeks with the most data, each containing on average 283,181 prices, spanning 28 product categories for an average of 6304 different products.

[6] Defined as the complement of the fraction of conditional entropy over the location entropy.

[7] $price < 25^{th}$percentile $- 3 \cdot$ interquartile range, and $price > 75^{th}$percentile $+ 3 \cdot$ interquartile range

The third dataset originates from Kaggle [8], a Machine Learning competition platform. The dataset contains 350 million purchase events from 311,539 consumers across 134 store chains. The data is anonymized, but contains the individual product price, product category, date of purchase and purchase amount. Most purchase events cost less than 25 USD. The country of the dataset is not disclosed, but purchase prices are given in USD and purchase amounts are described in the imperial system.

In order to estimate the location probability, an adversary requires the knowledge of the population in each location. On the country granularity, we use the data available from the World Bank [14] for the year 2013, while for the US city granularity we used the data from the US Census Bureau [40].

As described in Section 2.2, we increase the knowledge of the adversary with the product basket. A product basket details which and how many products an average person purchases, both in terms of quantity and monetary amount. We leverage a national product basket [5] from 2010 containing over 300 product categories in order to infer the ratio in which different products are bought over the year.

Numbeo Dataset (2 years)

Number of countries	112
Number of prices	328,720
Number of cities in the US	23
Number of prices in the US cities	11,686
Number of distinct product categories	23

Numbeo Test Dataset (5 months)

Number of countries	47
Number of prices	40,968
Number of distinct product categories	23

Chicago Dataset (5 years)

Number of stores	84
Number of total prices in top 85 weeks	24,070,437
Average number of prices per week	283,181 ± 6790
Number of distinct product categories	28
Average number of products per week	6304 ± 461

Kaggle Dataset (1 year)

Number of store chains	134
Number of purchase events	349,655,789
Number of consumers	311,539
Number of distinct product categories	836

Table 2: Statistics about the three price datasets

4 Experimental Evaluation

In this section we evaluate the adversarial models designed in Section 2.2. We start by presenting the assumptions and choices made for the evaluation.

4.1 Experimental Considerations

With respect to the value probability $P(v \mid l, c)$, we assume that the frequency of price values in the Numbeo dataset reflects the frequency of real-world purchase events with the corresponding price values. This is a natural assumption and is further motivated by the fact that e.g., Numbeo contributors likely entered the most popular price values for the considered product categories. Because our datasets contain a limited amount of products and product categories, our analysis is naturally confined to the available products. Note that, if the adversary knows the product categories of the purchases, e.g. milk, other categories such as apples can be ignored, which allows precise predictions with knowledge about few products. In order to compute the product category probability, $P(c \mid l)$, we only consider one national product basket and apply it to every country. Note that we do not use the product basket as an indicator of how much money is spent on average by a person, but rather as an indicator in which ratio products are bought.

SAMPLING PRICE VALUES: Given a location l, we generate *synthetic* consumer purchase events by sampling price values from the respective dataset. For the three datasets we consider adversaries with complete knowledge of the price values. In addition we instantiate an adversary with incomplete knowledge with the Numbeo test dataset. Given the product basket of the location l we compute the probability of a product category being sampled (cf. Equation 4). Thus, we sample each product category with the product category probability $P(c \mid l)$. For each location we repeat the sampling of the price values $n = 1000$ times and average the result.

ADDITIVE SMOOTHING PARAMETER: In the case of an adversary with incomplete knowledge, we make use of additive smoothing to avoid zero probabilities when aggregating the probabilities of multiple purchase events for locations (see Section 2.2.2). We choose a smoothing parameter $\alpha = 0.01$ which provides us with good results on our data (cf. appendix Figure 13).

In the following, we evaluate up to three knowledge scenarios (cf. Section 2.3) for four location granularities: (i) across 112 countries worldwide; (ii) across 23 cities within the United States; (iii) across 84 stores within the Chicago metropolitan area; (iv) we distinguish among 134 store chains in a country.

4.2 Country Granularity

The adversary has to distinguish 112 candidate countries for each purchase event. We quantify the privacy given the three privacy metrics defined in Section 2.6. In particular, we performed our study in two settings. First, (i) we assumed that the adversary does not have complete knowledge. This means that the adversary receives purchase events from the Numbeo test dataset and estimate their location based on the Numbeo dataset. In the second case, (ii) the adversary assumes complete knowledge of price values, and therefore, the sampled prices are included in the price dataset which is adversarial knowledge.

Figure 3 shows the F_1-score for the first case based on the number of purchase events accessible to the adversary. Given one purchase event, the price, price_merchant and price_product-category knowledge scenario achieve an average of 0.38, 0.41 and 0.49 respectively. The high F_1-score after one purchase event shows, that even one event allows a decent prediction. We observe that the adversary is more likely to identify the correct location when it knows the product category of the purchase event. On the contrary, if the adversary has access to 10 purchase events, the respective F_1-scores are 0.80, 0.85 and 0.90. In other words, 10 purchase events significantly improve the ability of the adversary to identify the location of the purchase events. The reported values are averaged over $n = 1000$ iterations.

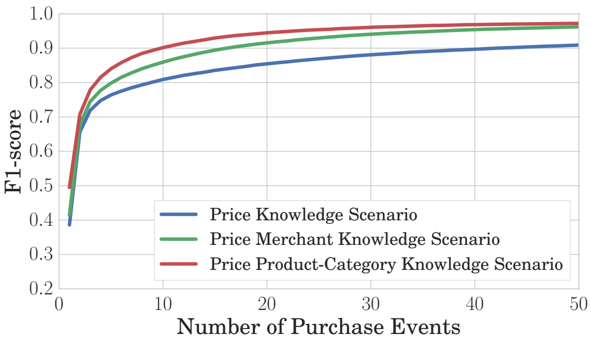

Figure 3: F_1-score for identifying the country given purchase events *sampled from the Numbeo test dataset*, corresponding to incomplete knowledge.

Figure 4 corresponds to the second case, where the adversary assumes complete knowledge of the price values. We observe that the adversary can distinguish more accurately between the possible locations.

The F_1-scores are averaged over all considered countries. Figure 14 in the appendix plots each considered country in the price knowledge scenario and shows that averaging does not hide poorly performing countries.

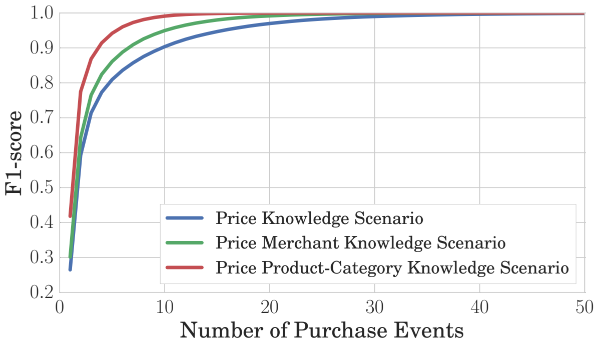

Figure 4: F_1-score for identifying the country given purchase events *sampled from the Numbeo dataset*, corresponding to complete knowledge. Averaging does not hide poorly performing countries (cf. appendix C).

Table 3 presents the results of the mutual information and the relative reduced entropy for each knowledge scenario. We observe that the price_product-category knowledge scenario reduces the entropy more significantly than the other knowledge scenarios. Naturally, this is because the price_product-category knowledge scenario provides the adversary with more information than the price knowledge scenario, thus effectively reducing uncertainty when identifying the location.

	Knowledge Scenarios		
	Price	**Price Merchant**	**Price Product-Category**
Mutual Information	0.539	0.841	1.703
Relative Reduced E.	0.114	0.178	0.360

Table 3: Mutual information and relative reduced entropy for the three knowledge scenarios when estimating the country of purchase events from 112 countries.

4.3 US City Granularity

In this section we analyze an adversary that aims to distinguish among the purchase events of 23 US cities. As before, we quantify the privacy based on the three privacy metrics defined in Section 2.6. We sample and test purchase events on the Numbeo dataset only, since our test dataset does not contain sufficiently many purchase events per considered US city.

Figure 5 illustrates the F_1-score depending on the number of purchase events. We observe, that after 10 purchase events, the F_1-score is greater than 0.7. Therefore, our methodology also provides accurate estimations

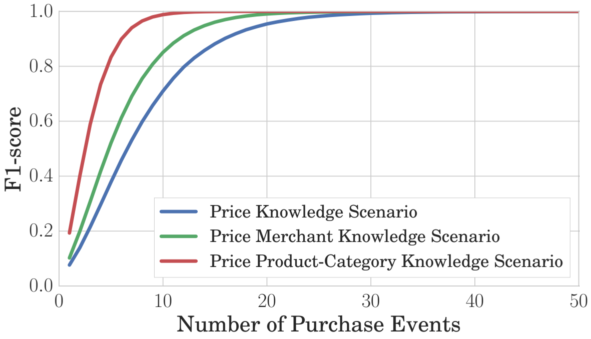

Figure 5: F_1-score for identifying the US city given purchase events for different knowledge scenarios. The purchase events are sampled from the Numbeo dataset.

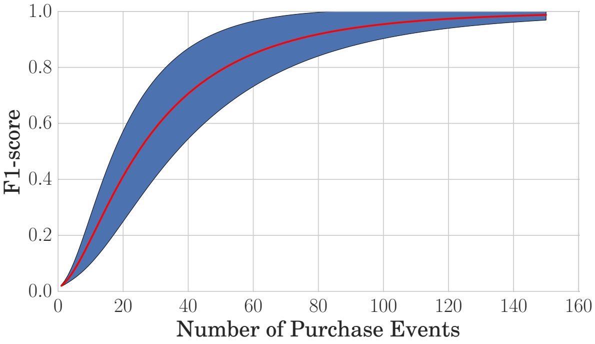

Figure 6: F_1-score and standard deviation over 85 weeks for identifying the store in the price knowledge scenario. Data sampled from the Chicago dataset among 84 stores.

on a city granularity. Table 4 reports the mutual information and relative reduced entropy when estimating the US city. We observe that the relative reduced entropies of country and city granularity match across the knowledge scenarios. This exemplifies the usefulness of the relative reduced entropy to highlight similarities across different price datasets.

	Knowledge Scenarios		
	Price	**Price Merchant**	**Price Product-Category**
Mutual Information	0.368	0.572	1.164
Relative Reduced E.	0.101	0.157	0.319

Table 4: Mutual information and relative reduced entropy for the three knowledge scenarios when estimating the city of purchase events among 23 US cities.

4.4 Chicago Metropolitan Granularity

In this section, we analyze an adversary that aims to distinguish among the purchase events of 84 Dominick's stores within the Chicago metropolitan area. We sample the price values from the Chicago dataset, and assume an adversary with complete knowledge; we therefore do not apply additive smoothing. We consider the location prior probability $P(l)$ to be uniform, because we do not have reliable store popularity information for the Chicago area.

In Figure 6 we can observe that the adversary can identify a local store given 100 purchase events with high confidence. We expected a weaker result, since all stores are operated by the same chain, implying relatively similar price structures. We ran our attack on each of the 85 weeks with most data, averaged the results and report the standard deviation as shown in the blue area of Figure 6.

Table 5 shows that the Chicago price dataset reveals less information about the considered locations than the Numbeo dataset. This observation holds for both knowledge scenarios, and is consistent with the result that more price points are required to localize purchase events within the Chicago area.

	Knowledge Scenarios	
	Price	**Price Product-Category**
Mutual Information	0.280	0.569
Relative Reduced E.	0.044	0.089

Table 5: Mutual information and relative reduced entropy when estimating the store of purchase events for 84 stores in the Chicago metropolitan area.

4.5 Store chain granularity

The large-scale Kaggle dataset does not provide precise location information of purchase events, but allows the adversary to distinguish among 134 store chains. Knowing the store chain of purchase events effectively reduces the possible locations of the purchases. Note, that the prices of Kaggle are distributed over a year and the adversary therefore does not know the precise time of the purchase events.

We uniformly sample purchase events of different consumers and perform our attack on the Kaggle dataset. Figure 7 reveals that given approximately 250 price values we achieve an F_1-score of over 0.95 for the origin of the purchase events. Note, that the price_product-category knowledge scenario is particularly strong due

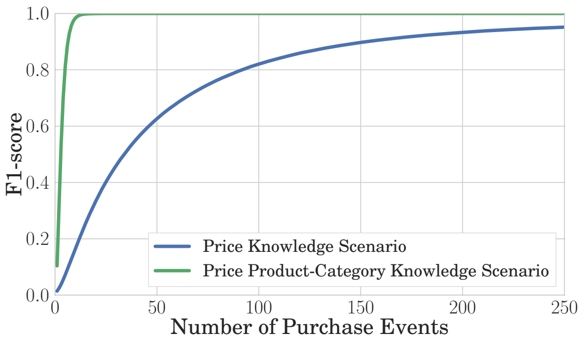

Figure 7: F_1-score for identifying the store chain. The purchase events are sampled from the Kaggle dataset.

	Knowledge Scenarios	
	Price	Price Product-Category
Mutual Information	0.456	2.256
Relative Reduced E.	0.068	0.337

Table 6: Mutual information and relative reduced entropy when estimating the store chain of purchase events for 134 chains.

to many product categories. This is reflected by the particularly high Mutual Information (cf. Table 6).

Given these results, we conclude, that our framework and methodology apply to a wide variety of different price datasets and allow us to quantitatively compare their respective privacy leakage. In the following, we extract further insights from our data to strengthen the attack.

4.6 Most Revealing Product Category

In this section we investigate which of the 23 considered product categories from the Numbeo dataset leak more information. This is a useful insight since an adversary would pick purchase events of this product category in order to increase the probability of correctly identifying their location. Therefore, with the mutual information we measure the extent to which the location entropy is reduced, given the purchase events of a particular product category. Contrary to the previous analysis, we evaluate the mutual information *per product category* based on the price_product-category knowledge scenario defined in Section 2.3. More specifically, we compute the mutual information using only purchase events of a particular product category.

The results of the evaluation can be found in Figure 8. According to this metric, the most revealing product categories are milk, a one-way ticket for local transportation, and a loaf of white bread. On the contrary, the product categories that disclose less information about a location are oranges, chicken breasts and rice.

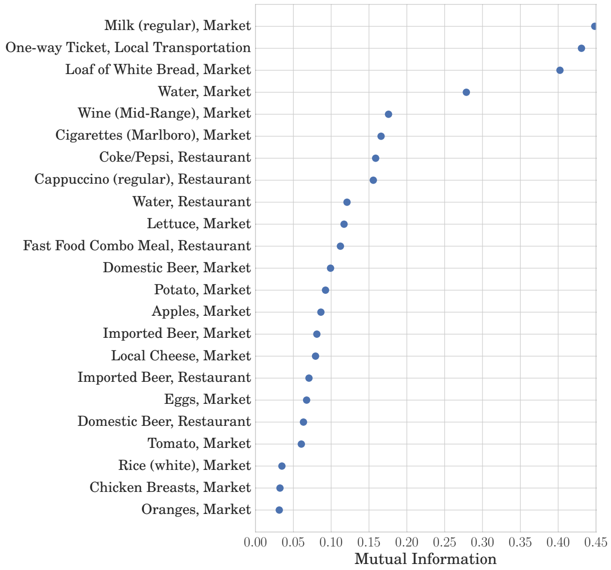

Figure 8: Mutual information of the respective product categories, the higher the mutual information, the more revealing is the product category.

4.7 Required Time Precision

Previously, we assumed that knowledge of the exact currency conversion rates is required to compare non-localized purchase events. Exact currency conversion rates, however, require a precise knowledge of the purchase event times. In this section, we show that our attack does not require the exact currency conversion rates, but also works if the adversary knows only the date or even week of the purchase, i.e. it has an uncertainty of 24 hours or 7 days in relation to the conversion rates. We therefore relax the requirements on the time precision.

Due to the conversion rate differences, the adversarial estimation of $P(v \mid l, c)$ is inaccurate. To compensate for the conversion rate differences, the adversary can use a price tolerance. We study two options for the tolerance: a static tolerance and a dynamic tolerance. For the static tolerance, the adversary estimates $P(v \mid l, c)$ in the presence of uncertainty by considering price values in the interval $[v - tol_s, v + tol_s]$ where the static tolerance tol_s is a small amount in global currency (e.g., 0.02 USD). The dynamic tolerance value tol_d is a percentage-wise estimate of uncertainty (e.g., 2%). To estimate $P(v \mid l, c)$ the adversary considers price values from the interval $[v \cdot (1 - tol_d), v \cdot (1 + tol_d)]$.

We evaluated the attack to infer the country of purchase events with imprecise purchase times and compensated the time error with different tolerance values. To simulate imprecise purchase times, we converted the adversarial knowledge using conversion rates of 30 different days from the year 2014 and then converted the non-localized purchase events S_U using the previous days' conversion rates. As before, we computed the F_1-score to evaluate the quality of the estimated $P(l \mid S_U)$.

For static and dynamic tolerance values, we found that the attack is still accurate, i.e. reaches an F_1-score above 95% with less than 50 purchase events. A higher tolerance value has two opposing effects: (i) it compensates for differences in currency conversion rates and increases the number of correctly considered price values; (ii) a higher tolerance, however, also increases the number of incorrectly considered price values which fall into larger intervals. Therefore, the tolerance value presents a trade off between the true-positive and true-negative rate. Our experimental results reflect this trade off both for static and dynamic tolerance values (cf. appendix C). Based on our experimental results we propose a dynamic tolerance of 2% for a 24-hour time imprecision.

We also evaluated the uncertainty of one week on the currency conversion rates. We used real-world currency conversion rates that were seven days apart from each other. Figure 9 shows the result of this experiment for the different knowledge scenarios and a dynamic tolerance value of 2% on the Numbeo dataset. We conclude that our attack does not require precise purchase event times.

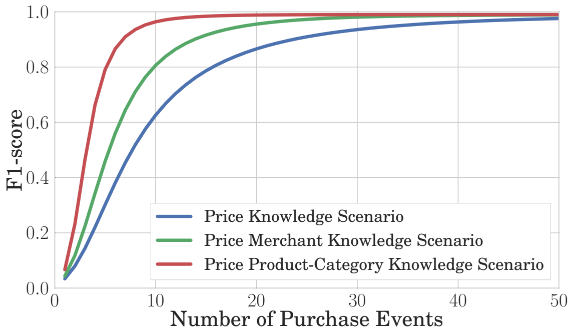

Figure 9: Dynamic tolerance of 2% with one week time uncertainty on the Numbeo dataset while estimating the country. Precise time allows an F_1-score of 0.95 after 10 purchase events whereas a one week time uncertainty achieves an F_1-score of 0.63.

Apart from using a tolerance value to compensate for purchase time uncertainty, we could also perform an exhaustive search over the time interval, apply the subsequently presented currency identification and pick the time at which the currency was most likely converted.

5 Currency Identification

We now consider the problem of identifying the currency of a purchase based on the price value. This is relevant since knowing the currency efficiently reduces the number of possible consumer locations. In the Bitcoin setting, one for example is able to obtain all transaction values. Thus, inferring the currency based on that value would leak location information.

Ideally, one would crawl Bitcoin transactions to create a dataset containing the true currency and the transaction amount and train a statistical model on this data. Ground truth, however, is unavailable (notwithstanding the fact that Bitcoin is currently mostly used for gambling [3, 12]). We therefore simulate block-chain consumer purchases by using the global prices from the Numbeo dataset. We then evaluate the quality of our prediction model. The trained model can then be used for a real attack on Bitcoin transactions.

We acknowledge that an adversary would have additional difficulties operating on the blockchain, e.g., it is unclear how to reliably filter irrelevant Bitcoin transactions. Nevertheless, Bitcoin change addresses [15] and tagged Bitcoin addresses [34] (e.g., from `blockchain.info`) can be identified. Moreover, when a Bitcoin transaction transfers e.g., full integer Bitcoins, it is likely a pure Bitcoin transaction, and less likely a transfer in a local currency expressed in Bitcoin.

The key insight to solve this problem is that individual product prices are not distributed randomly, but are usually rounded-off at values such as .25, .95 or .99. This is clearly visible in Figure 10, data is taken from Numbeo. Research in psychology related to price-setting [27] also support our assumptions.

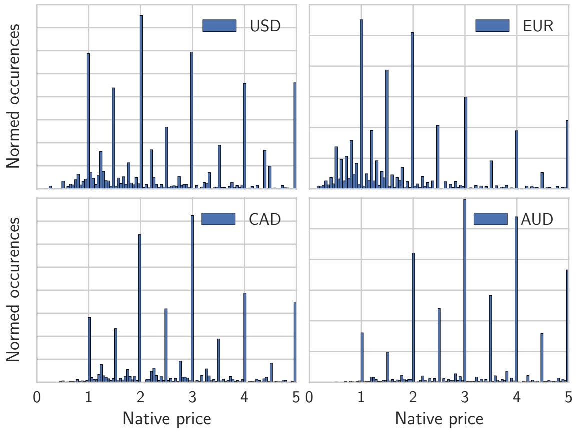

Figure 10: Distribution of prices in native currencies.

In order to infer the currency of a price value, we suggest a simple yet effective approach in which the price is converted to all known currencies and classified according to the following model:

$$
\begin{aligned}
p &= m \times v \times e, \\
c &= \arg\min_{e \in E} \{ |p - \lfloor p \rfloor|, |p - \lceil p \rceil| \},
\end{aligned}
$$

where v is the price in the global currency, $e \in E$ is the target currency conversion rate, p is the predicted price in the target currency. The effective role of m is to parametrize the assumption on the price distribution.

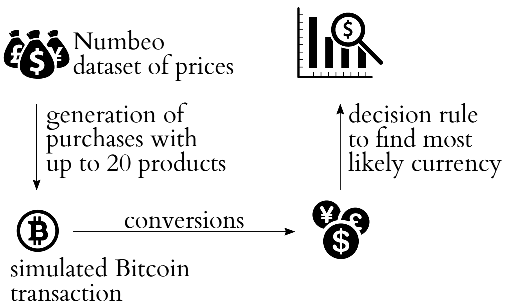
Numbeo dataset of prices
generation of purchases with up to 20 products
decision rule to find most likely currency
conversions
simulated Bitcoin transaction

Figure 11: Experimental setup for identifying the currency of purchases. We create 1,000,000 purchases by sampling real-world prices from the Numbeo dataset.

5.1 Experimental Results

To simulate purchase events containing more than one product, we create n synthetic datasets based on the original Numbeo dataset: We sample 1,000,000 purchases, each purchase containing randomly up to n products, $n \in [1, \ldots, 20]$. We then apply our algorithm on each dataset separately and report the scores as shown in Figure 11. We evaluated our hypothesis under two assumptions (i) the true conversion rates[8] used to convert the price value to the global currency are known (i.e. the Bitcoin to local currency exchange rate is accurate) and (ii) we do not aim to predict minor currencies which correspond to a multiple of major currencies (e.g., 1 Bahamas Dollar equals to 1 USD). Note that we weaken our adversary by considering each product combination as equally likely while generating the purchase events.

In order to determine the best performing parameter m on the given price value dataset, we choose the one ($m = 10$) which maximizes the F_1-score. The Numbeo dataset features 155 currencies which our algorithm then distinguishes upon. In Figure 12 we plot the overall weighted F_1-score depending on the number of products contained in a purchase. We fit the observed data to an exponential decay function, and conclude that even a purchase with 50 prices has an F_1-score above 0.4. For comparison, a random baseline achieves an F_1-score of nearly zero (approximately 0.008).

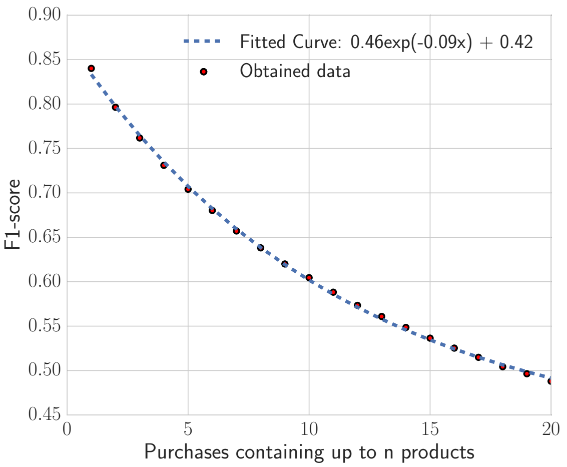

Figure 12: F_1-score for identifying among 155 currencies for purchases containing up to 20 products. Purchases with few products are more likely to reveal the currency.

In the case that the date of the transaction is not known, the currency conversion rate is not known. Because there are considerable fluctuations of exchange rates on even a daily bases, we hypothesize that not knowing the exact date of the transaction has an extremely negative effect on the classification algorithm. We support this hypothesis by using the exchange rates from a randomly selected day in 2014, instead of the correct ones. While for some stable currencies the effect is small (e.g., F_1-score for USD changes by 0.04), for majority of other currencies the F_1-score deteriorates to below 0.1. In order to assess our algorithm's performance among different currencies we reported the individual F_1-score in Table 7 and focus on major currencies. It shows, for example, that we can clearly distinguish whether the price is in Chinese Yuan (CNY) or not, therefore achieving a very high F_1-score.

Currency	F_1-score	Currency	F_1-score
RUB	0.98	TRY	0.93
CNY	0.96	EUR	0.89
INR	0.95	GBP	0.86
BRL	0.95	CAD	0.84
AUD	0.94	USD	0.58

Table 7: Top 10 major currencies and their respective difficulty of identification. CNY can be reliably guessed.

[8] The conversion rates are sampled from openexchangerates.org (30.06.2014)

The quality of the suggested model is established by very high values of the F_1-score [38], which demonstrate that this side channel information can be used in order to narrow down the currency and thus the possible locations of purchase events.

6 Related Work

Location Privacy. Blumberg [18] *et al.* provide a non-technical discussion of location privacy, its issues and implications. Gruteser and Grunwald [25] initiate major research in the area of the anonymization approaches to location privacy. Further, Narayanan *et al.* [32] investigate location privacy from a theoretical standpoint and present a variety of cryptographic protocols motivated by and optimized for practical constraints while focusing on proximity testing. Shokri *et al.* [37] propose a formal framework for quantifying location privacy in the case where users expose their location sporadically. They model various location-privacy-preserving mechanisms, such as location obfuscation and fake location injections. This work is orthogonal to ours, since in our setting the consumers are not willingly revealing their locations. Voulodimos *et al.* [41] address the issue of privacy protection in context-aware services through the use of entropy as a means of measuring the capability of locating a user's whereabouts and identifying personal selections. Narayanan [31] and Shmatikov propose statistical de-anonymization attacks against high-dimensional micro-data. We do not rely on their methods, since we are not aiming to de-anonymize the consumers. De Montjoye *et al.* [42] show that consumers can be uniquely identified within credit card records with only a few spatiotemporal triples containing location, time and price value. Contrary to their work, we focus on the price values and we localize instead of identify consumers.

Payment systems. The privacy implications of public transaction prices have been widely ignored. One prominent example is Bitcoin [36] [19], where transactions are exchanged between peers by means of pseudonyms. The actual transaction prices are archived and publicly available. The literature features many different methods for analyzing the privacy implications of Bitcoin, e.g., by means of appropriate heuristics [15], tainting [24], or other techniques [35] [23]. Reid and Harrigan [34] analyze the flow of Bitcoin transactions in a small part of the Bitcoin log, and show that external information like publicly-announced addresses, can be used to link identities and organizations to some transactions. In [30] the authors propose Zerocoin, a cryptographic extension to Bitcoin that augments the protocol to allow for fully anonymous currency transactions using a distributed ECash scheme. To the best of our knowledge only two contributions [16] [17] have aimed to hide the transaction prices in Bitcoin.

Price rigidity. Herrmann and Moeser [26] perform a quantitative analysis on price variability and conclude that prices are often rigid for several weeks. Pricing strategies for identical brands, however, vary significantly among retailers. Their observations match the studies of the Big Mac index [6] (the Economist), the Starbucks coffee index [9] (the Wall Street Journal) and the Ikea Billy Bookshelf index [2] (Bloomberg). The former studies show that prices of identical products from a single brand vary across locations. Dutta *et al.* [22] find that retail prices respond promptly to direct cost changes as well as upstream manufacturers' costs. Hosken and Reiffen [28] find that each product has a price mode—a price that the product stays at most of the time. Note that Hosken's non-public dataset contains nearly as many price observations as our Numbeo dataset.

7 Conclusion

Having a systematic methodology to reason quantitatively about the privacy leakage from datasets containing price relevant information is a necessary step to avoid privacy leakages. While further tests with more datasets will help to generally claim that price values alone can reveal the location of a purchase, our empirical results provide evidence that with relatively few purchase events it is possible to identify a consumer's location. In this paper, we have raised the following two questions: How much location information is leaked by consumer purchase datasets? How can it be quantified with the considered adversarial model and knowledge? In our proposed framework, we have modeled several adversaries and quantified the privacy leakage according to different dimensions. We make extensive use of Bayesian inference in our framework to model the different attack strategies. Our framework can be easily applied to any price dataset of consumer purchases and allows one to compare the privacy leakage of different datasets. We applied our methodology to three real-world datasets and achieve comparable results. In addition, we presented a novel methodology for identifying the currency of a purchase knowing only the total purchase price. The results presented in this paper strongly motivate the need for careful consideration when sharing price datasets and should be considered when designing public ledger cryptocurrencies.

References

[1] A Face Is Exposed for AOL Searcher No. 4417749, 2006. Available from: http://www.nytimes.com/2006/08/09/technology/09aol.html.

[2] Ikea Billy Bookshelf Index, Bloomberg, 2009. Available from: http://www.bloomberg.com/apps/news?pid=newsarchive&sid=a.K4T4ypP9ko.

[3] Lightspeed venture partners, 2013. Available from: http://lsvp.com/2013/08/23/at-least-half-of-all-bitcoin-transactions\-are-for-online-gambling/.

[4] NIST/SEMATECH e-Handbook of Statistical Methods, 2013. Available from: http://www.itl.nist.gov/div898/handbook/.

[5] Anonymized for review, 2015.

[6] Big Mac Index, The Economist, 2015. Available from: http://www.economist.com/content/big-mac-index.

[7] Consumer panel data and retail scanner data across the United States., 2015. Available from: http://research.chicagobooth.edu/nielsen/.

[8] Kaggle, Acquire Valued Shoppers Challenge, 2015. Available from: https://www.kaggle.com/c/acquire-valued-shoppers-challenge.

[9] More (or Less) Brew for your Buck, Starbucks coffee price, 2015. Available from: http://online.wsj.com/news/articles/SB10001424127887324048904578319783080709860.

[10] Numbeo, database of user contributed data about cities and countries worldwide., 2015. Available from: http://www.numbeo.com.

[11] Ripple, cryptocurrency, 2015. Available from: https://ripple.com/.

[12] Satoshidice, 2015. Available from: https://satoshidice.com.

[13] Store-level scanner data collected at Dominick's Finer Foods., 2015. Available from: http://research.chicagobooth.edu/kilts/marketing-databases/dominicks/dataset.

[14] World Population, The world bank, 2015. Available from: http://data.worldbank.org/indicator/SP.POP.TOTL?order=wbapi_data_value_2009+wbapi_data_value+wbapi_data_value-first&sort=asc.

[15] Elli Androulaki, Ghassan Karame, and Srdjan Capkun. Evaluating user privacy in bitcoin. In *Financial Cryptography and Data Security*, 2013. http://eprint.iacr.org/2012/596.pdf.

[16] Elli Androulaki and Ghassan O Karame. Hiding transaction amounts and balances in bitcoin. In *Trust and Trustworthy Computing*, pages 161–178. Springer, 2014.

[17] Eli Ben-Sasson, Alessandro Chiesa, Christina Garman, Matthew Green, Ian Miers, Eran Tromer, and Madars Virza. Zerocash: Decentralized anonymous payments from bitcoin. In *Security and Privacy (SP), 2014 IEEE Symposium on. IEEE*, 2014.

[18] Andrew J Blumberg and Peter Eckersley. On locational privacy, and how to avoid losing it forever. *EEF*, 2009.

[19] Joseph Bonneau, Andrew Miller, Jeremy Clark, Arvind Naryanan, Joshua A. Kroll, and Edward W. Felten. SoK: Bitcoin and second-generation cryptocurrencies. In *IEEE Security and Privacy 2015*, May 2015.

[20] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[21] Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013.

[22] Shantanu Dutta, Mark Bergen, and Daniel Levy. Price flexibility in channels of distribution: Evidence from scanner data. *Journal of Economic Dynamics and Control*, 26(11):1845 – 1900, 2002.

[23] Meiklejohn et al. A fistful of bitcoins: Characterizing payments among men with no names. In *Proceedings of the 2013 Conference on Internet Measurement Conference*, IMC '13, pages 127–140, New York, NY, USA, 2013. ACM.

[24] Arthur Gervais, Ghassan Karame, Srdjan Capkun, and Vedran Capkun. Is bitcoin a decentralized currency? *IEEE Security and Privacy Magazine, 2014.*

[25] Marco Gruteser and Dirk Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the 1st international conference on Mobile systems, applications and services*, pages 31–42. ACM, 2003.

[26] Roland Herrmann and Anke Möser. Price variability or rigidity in the food-retailing sector? theoretical analysis and evidence from german scanner data. Technical report, 2003.

[27] Judith Holdershaw, Philip Gendall, and Ron Garland. The widespread use of odd pricing in the retail sector. 8:53–58, 1997.

[28] Daniel Hosken and David Reiffen. Patterns of retail price variation. *RAND Journal of Economics*, pages 128–146, 2004.

[29] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[30] Ian Miers, Christina Garman, Matthew Green, and Aviel D Rubin. Zerocoin: Anonymous distributed e-cash from bitcoin. In *Security and Privacy (SP), 2013 IEEE Symposium on*, pages 397–411. IEEE, 2013.

[31] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, 2008.

[32] Arvind Narayanan, Narendran Thiagarajan, Mugdha Lakhani, Michael Hamburg, and Dan Boneh. Location privacy via private proximity testing.

[33] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems*, InfoScale '06, New York, NY, USA, 2006. ACM.

[34] Fergal Reid and Martin Harrigan. An analysis of anonymity in the bitcoin system.

[35] Dorit Ron and Adi Shamir. Quantitative analysis of the full bitcoin transaction graph. 2013. http://eprint.iacr.org/2012/584.pdf.

[36] S. Nakamoto. Bitcoin: A Peer-to-Peer Electronic Cash System, 2009.

[37] Reza Shokri, George Theodorakopoulos, George Danezis, Jean-Pierre Hubaux, and Jean-Yves Le Boudec. Quantifying location privacy: the case of sporadic location exposure. In *Privacy Enhancing Technologies*, pages 57–76. Springer, 2011.

[38] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

[39] Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671:1–34, 2000.

[40] U.S. Census Bureau, Population Division. Annual Estimates of the Resident Population for Incorporated Places of 50,000 or More, Ranked by July 1, 2013, 2014.

[41] Athanasios S Voulodimos and Charalampos Z Patrikakis. Quantifying privacy in terms of entropy for context aware services. *Identity in the Information Society*, 2(2):155–169, 2009.

[42] Vivek Kumar Singh Alex Sandy Pentland Yves-Alexandre de Montjoye, Laura Radaelli. Unique in the shopping mall: on the reidentifiability of credit card metadata. *Science, 2015*.

Appendix A: Probability calculations

In the following we clarify the individual steps for calculating the probabilities derived in Section 2.

$$\begin{aligned} P(v) &= \sum_{l\in\mathbb{L}} P(l,v) = \sum_{l\in\mathbb{L}}\sum_{c\in\mathbb{C}} P(l,c,v) \\ &= \sum_{l\in\mathbb{L}}\sum_{c\in\mathbb{C}} P(l)\cdot P(c \mid l)\cdot P(v \mid l,c) \end{aligned} \tag{10}$$

$$\begin{aligned} P(c,v) &= \sum_{l\in\mathbb{L}} P(l,c,v) \\ &= \sum_{l\in\mathbb{L}} P(l)\cdot P(c \mid l)\cdot P(v \mid l,c) \end{aligned} \tag{11}$$

$$\begin{aligned} P(v \mid l) &= \frac{P(l,v)}{P(l)} = \frac{\sum\limits_{c\in\mathbb{C}} P(l,c,v)}{P(l)} \\ &= \sum_{c\in\mathbb{C}} P(c \mid l)\cdot P(v \mid l,c) \end{aligned} \tag{12}$$

$$\begin{aligned} P(c,v \mid l) &= \frac{P(l,c,v)}{P(l)} = \frac{P(l)\cdot P(c \mid l)\cdot P(v \mid l,c)}{P(l)} \\ &= P(c \mid l)\cdot P(v \mid l,p) \end{aligned} \tag{13}$$

$$\begin{aligned} P(l \mid v) &= \frac{P(l)\cdot P(v \mid l)}{P(v)} \\ &= \frac{P(l)\cdot \sum\limits_{c\in\mathbb{C}} [P(c \mid l)\cdot P(v \mid l,c)]}{\sum\limits_{l'\in\mathbb{L}}\sum\limits_{c\in\mathbb{C}} [P(l')\cdot P(c \mid l')\cdot P(v \mid l',c)]} \\ &= \frac{P(l)\cdot \sum\limits_{c\in\mathbb{C}} [P(c \mid l)\cdot P(v \mid l,c)]}{\sum\limits_{l'\in\mathbb{L}} P(l')\cdot \sum\limits_{c\in\mathbb{C}} [P(c \mid l')\cdot P(v \mid l',c)]} \end{aligned} \tag{14}$$

$$= \frac{\frac{\text{Population}(l)}{\sum_{l'\in\mathbb{L}} \text{Population}(l')} \cdot \sum\limits_{c\in\mathbb{C}} [\frac{\text{Basket}(l,c)}{\sum_{c'\in\mathbb{C}} \text{Basket}(l,c')} \cdot \frac{D(l,c,v)}{D(l,c)}]}{\sum\limits_{l'\in\mathbb{L}} \frac{\text{Population}(l')}{\sum_{l''\in\mathbb{L}} \text{Population}(l'')} \cdot \sum\limits_{c\in\mathbb{C}} [\frac{\text{Basket}(l',c)}{\sum_{c'\in\mathbb{C}} \text{Basket}(l',c')} \cdot \frac{D(l',c,v)}{D(l',c)}]} \tag{15}$$

$$\begin{aligned} P(l \mid c,v) &= \frac{P(l)\cdot P(c,v \mid l)}{P(c,v)} \\ &= \frac{P(l)\cdot [P(c \mid l)\cdot P(v \mid l,p)]}{\sum\limits_{l'\in\mathbb{L}} [P(l')\cdot P(c \mid l')\cdot P(v \mid l',c)]} \end{aligned} \tag{16}$$

$$= \frac{\frac{\text{Population}(l)}{\sum_{l'\in\mathbb{L}} \text{Population}(l')} \cdot \frac{\text{Basket}(l,c)}{\sum_{c'\in\mathbb{C}} \text{Basket}(l,c')} \cdot \frac{D(l,c,v)}{D(l,c)}}{\sum\limits_{l'\in\mathbb{L}} [\frac{\text{Population}(l')}{\sum_{l''\in\mathbb{L}} \text{Population}(l'')} \cdot \frac{\text{Basket}(l',c)}{\sum_{c'\in\mathbb{C}} \text{Basket}(l',c')} \cdot \frac{D(l',c,v)}{D(l',c)}]} \tag{17}$$

$$\begin{aligned} P(l \mid S_U) &= P(l \mid V(e_1),V(e_2),\ldots,V(e_n)) \\ &= \frac{\prod\limits_{i=1..n} P(V(e_i))}{P(V(e_1),V(e_2),\ldots,V(e_n))} \\ &\frac{\prod\limits_{i=1..n} P(l \mid V(e_i))}{P(l)^{n-1}} \\ &= \frac{\prod\limits_{e\in S_U} P(l)P(V(e) \mid l)}{P(V(e_1),\ldots,V(e_n))\cdot P(l)^{n-1}} \\ &= \frac{P(l)\cdot \prod\limits_{e\in S_U} P(V(e) \mid l)}{P(V(e_1),\ldots,V(e_n))} \end{aligned} \tag{18}$$

Appendix A.1: Probability calculations

Based on its knowledge, the ideal adversary computes the following probabilities by computing the fractions of events.

$$P(l) = \frac{|\{e|e \in \mathscr{H}_G : e.l = l\}|}{|\mathscr{H}_G|} \tag{19}$$

$$P(v) = \frac{|\{e|e \in \mathscr{H}_G : e.v = v\}|}{|\mathscr{H}_G|} \tag{20}$$

$$P(c,v) = \frac{|\{e|e \in \mathscr{H}_G : e.c = c \wedge e.v = v\}|}{|\mathscr{H}_G|} \tag{21}$$

$$P(v \mid l) = \frac{|\{e|e \in \mathscr{H}_G : e.l = l \wedge e.v = v\}|}{|\{e|e \in \mathscr{H}_G : e.l = l\}|} \tag{22}$$

$$P(c,v \mid l) = \frac{|\{e|e \in \mathscr{H}_G : e.l = l \wedge e.c = c \wedge e.v = v\}|}{|\{e|e \in \mathscr{H}_G : e.l = l\}|} \tag{23}$$

Appendix B: Dataset information

Category and Merchant	Unit	Prices
Market Product Categories		
Apples	1 kg	11876
Chicken Breasts	1 kg	11893
Cigarettes (Marlboro)	1 pack	12712
Domestic Beer	one 0.5 liter bottle	10243
Eggs	12 units	14617
Imported Beer	one 0.33 liter bottle	9484
Lettuce	1 head	8966
Loaf of White Bread	0.5 kg	14633
Local Cheese	1 kg	10975
Milk (regular)	1 liter	17197
Oranges	1 kg	10289
Potato	1 kg	10891
Rice (white)	1 kg	10924
Tomato	1 kg	10539
Water	1.5 liter bottle	12762
Wine (Mid-Range)	1 bottle	11893
Restaurant Product Categories		
Cappuccino (regular)	1 unit	21539
Coke/Pepsi	one 0.33 liter bottle	21351
Fast Food Combo Meal	1 unit	21794
Domestic Beer	one 0.5 liter bottle	19128
Imported Beer	one 0.33 liter bottle	18048
Water	one 0.33 liter bottle	21691
Local Transportation Categories		
One-way Ticket	1 unit	15275

Table 8: Product categories of the Numbeo dataset.

Appendix C: Further Experimental Results

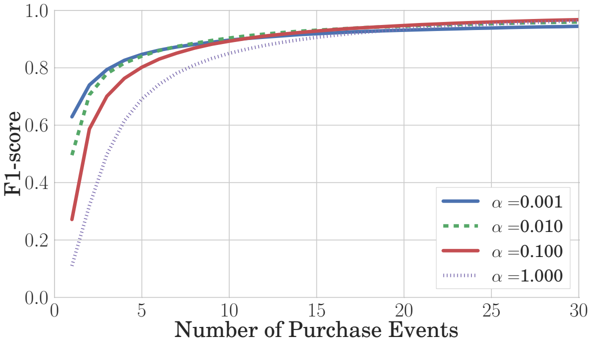

Figure 13: Comparison of different α-parameters for additive smoothing based on the price_product-category knowledge scenario.

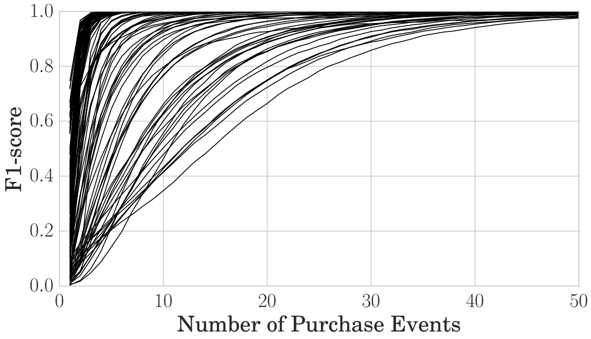

Figure 14: F_1-score of each individual country for the price knowledge scenario. The purchase events are sampled from Numbeo. We observe that no country performs poorly.

Appendix C.1: Required Time Precision

Figure 15 shows, that a larger tol_s will improve the overall F_1-score, but more purchase events are needed to filter out the false positives. Similarly, for the dynamic tolerance in Figure 16, a higher value for tol_d provides a better prediction for many purchase events, but a worse prediction for few purchase events. The figures show the experiments for the price_product-category knowledge scenario, however, we note that the results are analogous to the other scenarios. Based on these results we propose a dynamic tolerance of 2% in the case of a 24-hour time imprecision on the conversion rate.

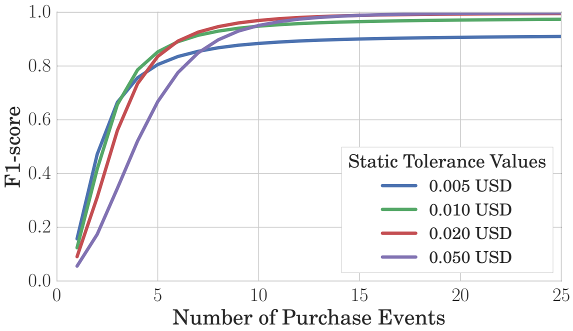

Figure 15: Using static tolerance values to compensate for imprecise time information (one day uncertainty) in the price_product-category knowledge scenario.

Appendix C.2: Motivating example

Since products appear in a multitude of price values, it is at first unclear how accurately price values can identify a location. To illustrate why purchases can be localized,

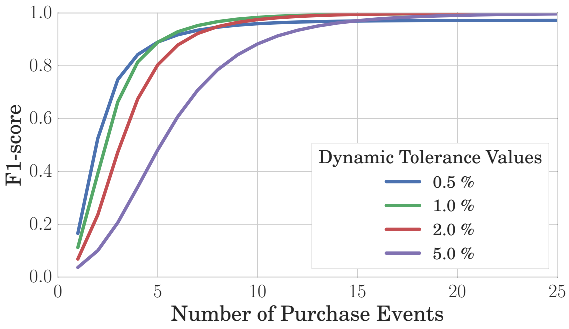

Figure 16: Using dynamic tolerance values to compensate for imprecise time information (one day uncertainty) in the price_product-category knowledge scenario.

we focus on an example of the product category *domestic beer (0.5L bottle)*, which can be bought in nearly every country. The price values are taken from the Numbeo dataset [10]. Figure 17 shows the distribution of price values of beer in USD for four countries. We observe that ranges of prices clearly differ for India and the other countries, while prices in Australia are more likely to be higher than in the US and Canada, where distributions of prices are similar. Given a beer price above 3 USD, in this case, it is highly likely that the purchase has not occurred in India.

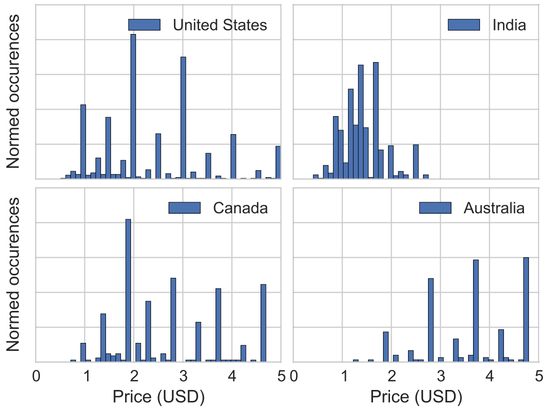

Figure 17: Distribution of domestic beer prices (0.5 Liter) in 4 countries. Numbeo prices, converted to US Dollar.