Passengers information in public transport and privacy: can anonymous tickets prevent tracking?

Gildas Avoine[•], Luca Calderoni[†][*], Jonathan Delvaux[•], Dario Maio[†], Paolo Palmieri[‡]

[•]*Information Security Group, Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium*

[†]*Department of Computer Science and Engineering, Università di Bologna, 47521 Cesena, Italy*

[‡]*Parallel and Distributed Systems Group, Delft University of Technology, 2628CD Delft, Netherlands*

Abstract

Modern public transportation companies often record large amounts of information. Privacy can be safeguarded by discarding nominal tickets, or introducing anonymization techniques. But is anonymity at all possible when everything is recorded? In this paper we discuss travel information management in the public transport scenario and we present a revealing case study (relative to the city of Cesena, Italy), showing that even anonymous 10-ride bus tickets may betray a user's privacy expectations. We also propose a number of recommendations for the design and management of public transport information systems, aimed at preserving the users' privacy, while retaining the useful analysis features enabled by the e-ticketing technology.

Keywords: Privacy, Public Transport, Sensitive Data Management, Privacy Preserving Technologies

1. Introduction

In 2013, East Japan Railway (JR East), the largest rail company in the country, announced the intention to sell to Hitachi corporation a large dataset of its passengers' travel histories (Geuss, 2013). This information has been gathered by JR East through its proprietary e-ticketing system, *Suica*. The company plans to anonymize these data by replacing sensitive information, such as names and addresses of card owners, with anonymous ID's. But is this enough to protect users' identities, and therefore their privacy? Historically, public releases of anonymized personal information have often proved to be dangerous for the privacy of the people that information concerned. In 2006, America On Line (AOL) released anonymized data regarding the search queries of millions of users of its web search engine. Even if the IP adresses of the users were replaced by anonymous identifiers, researchers and even journalists had little trouble finding the real names of people corresponding to the anonymous ID's, as proved by the famous case of user #4417749, presented in a New York Times article (Barbaro and Zeller, 2006). In this paper we show how the disclosure of travel histories can be equally dangerous, as travel data contain a great deal of information about the user, even when her real identity is concealed by means of anonymization.

The use of electronic tickets, usually smart cards, in public transportation networks has a number of potential benefits, both for the users and the provider of the transportation service. Often, their introduction also coincides with a more general modernization of the transportation infrastructure. Modern networks usually integrate a positioning system (GPS) for monitoring the movements of buses and trams, backed by a constant Internet connection to a central control infrastructure. Enabling location-awareness allows, for instance, to display real time information and waiting times for each line on the provider's website or on information screens at bus stops and represent a value-added *city-to-citizen* service in the smart urban ecosystem (Calderoni et al., 2012). Internet communication between vehicles and a central server can also be used to signal traffic congestion or unexpected issues efficiently, in both directions. These innovations help in making our cities smarter and greener, by improving the quality and reliability of the public transport service.

However, the technology enabling these features also

[*] Corresponding author. E-mail address: luca.calderoni@unibo.it

generates an unprecedented amount of information regarding user movements. And after such information is generated, the tendency among public transportation companies is to record it, rather than discard it when it exhausted its original goal. In this paper we discuss the privacy implications of such a large amount of data, and we analyze the potential consequences of its disclosure. As electronic tickets are generally characterized by a unique ID, and all trips are recorded, the information stored in the information system of a public transport company is nothing less than a detailed log of each user's movements and therefore should be treated as sensitive information. However, in this paper we show that even when personal information of the users are not stored in the system - or are anonymized - the threat to privacy remains. In fact, combining the data in the transportation company database with other publicly available source of information can ultimately be enough to identify a specific user, even in the case of anonymous tickets.

1.1. Contribution

In this paper we focus on anonymous, disposable 10-ride electronic tickets for public transportation. Such tickets can be generally bought anonymously through resellers or automated machines and, while they are identified through a unique ID, they do not carry any information on the owner's identity. Thus, these tickets are perceived as the most privacy-friendly by the users while, at the same time, retaining some of the advantages of personal travel passes, such as a lower cost per ride than single-ride tickets, and the ability to be used multiple times. The choice of anonymous tickets allows us to evaluate the potential effects of disclosure of travel histories to third parties, even when limited to a small number of rides and anonymized by removing personal information.

In this paper we present the case study of a real, city-wide public transport network in Italy. By analyzing and decoding the tickets issued by the company, we infer the information collected during their use. We use this knowledge to show that even anonymized and numerically limited travel histories are indeed enough to profile users with a great depth of detail. We also show that careful elaboration of these data, and comparison with other publicly available sources of information ultimately allows to find matching patterns and to statistically identify the user as belonging to a small, well-defined group. Empirical evidence produced by analyzing this case study proves that simple anonymization of the travel histories of public transportation users is not sufficient to protect their privacy, and therefore suggests caution in the disclosure or trade of such data without the informed consent of the users themselves. In order to address this issue, we propose a set of recommendations for the design and management of the information systems of transportation companies. Our solutions are both privacy-preserving and cost-effective, as they reduce the overhead in communication and storage of travel data to the information system, while avoiding costly renovations of current infrastructures.

In this paper we focus on a specific case study (the Italian city of Cesena and its public transport system) and we analyze the potential information disclosure for a specific set of users (university students). However, the problem we bring to light is indeed common to other cities and countries. If in this work we use public information on students' classes and housing, the same result could be achieved, for instance, using phone directories. Overall, the aim of this paper is not to prove a flaw in the design of a specific e-ticketing systems, but rather to show how the disclosure or sale of location-aware information, such as travel histories, even when anonymous (or anonymized) could become dangerous to the privacy of the concerned individuals: in fact, when data are combined with other sources of information, the presumed anonymity disappears.

1.2. Related Works

As discussed by Diaz and Gürses (2012), it is often difficult for individuals to know how their personal data are used by companies that hold them. While Diaz and Gurses mostly focus on sensitive data as defined by European Union regulations, their reasoning is equally valid when applied to companies collecting location data such as travel histories, as in the case of public transportation companies. In fact, users are often unaware of the risks of malicious surveillance, profiling or manipulation they are exposed to (Avoine et al., 2010). The security of this information is therefore best assured adopting the *Privacy by Design* paradigm, i.e. providing data anonymity by designing the appropriate protocols and procedures as hard-coded in the system itself (Diaz and Gürses, 2012). The privacy-friendliness of the infrastructure, if correctly implemented, does not necessarily hinder the business model (Liu et al., 2011). In the case of public transportation, electronic tickets raise privacy concerns for their ability to track users. Recent studies on the subject discuss this issue from the point of view of security against external attackers (Asadpour and Dashti, 2011; Avoine et al., 2010; Heydt-Benjamin et al., 2006; Sadeghi et al., 2008). A typical attacker is therefore some unauthorized person trying to monitor the movements of a victim, for instance

	Information	Time requirements	Privacy implications
Pricing	**P.1** A sorted list of each stamp performed by each single identified user, in order to compute his fare (for travels with connections, composed of multiple stamps). **P.2** Discounts according to well-defined categories (students, elderly, people with disabilities . . .). **P.3** Zone-based or stop-based discounts (special tickets, such as airport shuttles, special events/destinations passes etc.).	Pricing information should be stored for billing purposes only and thus they should be deleted after each invoice issuance.	For billing purposes personal information of the user might be needed. However, the short-lived time required to perform these operations reduces the privacy implications. Moreover the user is generally well-aware of the company disposal of his personal information required for billing (address, credit card data, . . .) as he provided them himself.
Statistics	**S.1** Total amount of passengers for each ride of a line, in order to monitor the line workload. **S.2** A sorted list of each stamp performed by users during a ride, in order to understand anonymous patterns in user' habits. **S.3** A sorted list of each stamp performed by each single user during a single day in order to understand one-way and return patterns in users' habits. **S.4** Total amount of journeys related to a single ticket, in order to monitor the workload of a long-term (monthly, yearly) pass.	This information needs to be stored for a long time as they are used to perform statistical analysis, even in a long-term year-over-year comparisons.	Trip information collected by transport companies are less sensitive than billing data, but more invasive for mainly three reasons. First, the user might not be aware of the recording and storing of this information, contrary to the case of billing data. Second, the information needs to be stored for significantly longer periods of time in order to be useful. Finally, travel histories contain location information, which imply the users habits and the places he regularly visits, along with time and frequencies of the movements. These are potentially more invasive to the privacy of the users than mere financial records.

Table 1: Information typically handled by public transport companies.

by accessing the records of those movements stored on the ticket itself (usually a smartcard). For this reason, the studies usually conclude that no sensible information should be kept within the smartcard for longer than it is actually required for the correct functioning of the system. This is the case of Avoine et al. (2010), where the authors discovered, through an analysis of the Mobib smartcard (the public transport pass used in the city of Brussels, Belgium) the presence of unneeded information that could expose users to privacy threats. In this paper, instead, we are interested in privacy with respect to the company providing the transportation service. In a typical scenario of an RFID-based ticketing system, the smart card ticket is read on the vehicle in order to learn its unique identifier, which is then sent from the reader to the central server encrypted (Asadpour and Dashti, 2011), usually by applying a collision resistant hash function to the identifier. Unfortunately, this allows different stamps to be associated with the same user and therefore permits tracking (Sadeghi et al., 2008). In Kerschbaum et al. (2013) the Authors focus on electronic cash payments and bill processing in the e-ticketing scenario and discuss how to achieve a privacy preserving billing system based on asymmetric key encryption while in Peng and Bao (2010) a simple billing mechanism designed to avoid privacy leaks is proposed. Basically, it enables the public transport company not to collect the starting place and the ending one in order to compute the journey cost. Security and privacy issues related to *Near Field Communication*, a very common technology used for mobile-payments on public transports, are discussed by Salonen (2011). In Tseytin et al. (2006), the use of anonymous databases for collecting user movements is discussed. The authors show, from a theoretical point of view, that anonymity alone is not enough to protect users' privacy. In this paper we provide a real-world case study confirming their intuition, and we propose a set of plug-in privacy enhancements for existing information systems.

1.3. Laws pertaining public information

In this paper we show ways of de-anonymizing travel histories by comparing them to other sources of information. In order to show the viability of this approach, we use for this purpose only publicly available information. The case study we present focuses on Cesena's university students: we use therefore information from the local university dormitories and housing directories. In the following, we provide legal references showing how it is, in general, mandatory to maintain this personal information publicly accessible: this is due to transparency policies for applicants in merit rankings.

According to the Italian law D.P.R. 09.05.1994 n. 487 (published on the official gazette n. 185 on August 9, 1994), each ranking list related to a public competition must be published and accessible to the public. This is the case for instance of subsidized housing for students or public housing for disadvantaged people. More generally, it is commonly stated in law that the protection of personal data of an individual shall not apply when the exposure of such information is required by law itself for reasons of transparency and public access to information held by authorities. In practice, some public lists of various kinds will always be exposed, due to the need of balancing state responsibilities in terms of transparency and human rights in terms of privacy protection.

1.4. Outline of the Paper

After an introduction concerning information management in the public transport field, provided in Section 2, we discuss how even anonymous tickets can provide enough information to track the users of a public transportation network in Section 3. In particular, we present the case study of the service provided by Start Romagna in the city of Cesena, while in Section 3.1 we disclose the data recorded in a MiMuovo ticket after its use. We show how this information is sufficient to endanger the privacy of the users in Section 3.2. We propose solutions that can help mitigate the privacy issues we presented in Section 4.

2. Passengers Information Management in Public Transport

Modern transportation systems integrate monitoring technology, both location based (GPS) and users based (smartcard tickets), that generates real time information on the current status of the transportation network. This constant flow of information is generally elaborated in real time, but is also stored and kept, usually indefinitely, by the relevant actors.

Passengers data collected by transportation companies are used for a number of purposes, with the most obvious one being ensuring that passengers pay the bus fare. In general, this can be achieved by verifying the authenticity of the ticket and its validity for the current ride, and this kind of checks is only performed during the ride itself. In modern ticketing systems, however, information on bus rides is often transmitted to a central information system, where it is stored for an indefinite period of time. This allows the company to gather statistics and useful contextualized data that can be used later to monitor usage, propose modifications to the bus network and adjust frequency, fares, etc (de Grange et al., 2013). But, as the JR East/Hitachi case shows, these data are also of increasing interest to external actors, and therefore have an inherent monetary value. if internal policies or existing legislation do not prevent this, it is realistic to predict the emergence of a number of cases in which this information will be old to external (and possibly foreign) entities.

In Table 1, we discuss the main information units that are commonly kept in information systems of public transport companies (Sampaio et al., 2008). We divide them in two main subgroups, namely statistics and pricing.

This large amount of recorded data poses an inherent threat to the privacy of the citizens. This threat is not necessarily linked to the increased integration of electronic identification mechanisms. It is in fact the sheer amount of data that enables tracking and tracing of the citizens: when everything a person does is registered, identifying that person among a group is a trivial task. In the following, we discuss an example of how this is possible even when the available data are apparently limited, such as in the case of anonymous 10-ride tickets for the public transport network of a small city.

3. Case Study: the Cesena Bus Network

The Italian city of Cesena is a small-sized town (less than 100.000 inhabitants) hosting a university campus. The campus, part of the Università di Bologna, offers five different majors (computer science, electrical and bio-medical engineering, architecture, agricultural sciences and psychology) to around 4.000 students (Table 2). The city is served by a bus network, provided by the regional public transport company Start Romagna, counting 6 city lines and 13 connections to nearby cities. The local population of university students is one of the main customers of the transport network, and benefits from specific fare discounts.

After a recent renovation of the bus network and the introduction of electronic tickets (called MiMuovo), an

Major	1st	2nd	3rd	Tot.
Agriculture	221	151	207	579
Architecture	133	127	167	427
Psychology	506	606	441	1553
Computer Science	229	140	190	559
Engineering	329	299	407	1035
Total	**1453**	**1350**	**1458**	**4261**

Table 2: Students enrolled in Bachelor's degrees offered at the Cesena Campus, by year (survey for the academic year 2012-13).

intelligent information system has been put in place by the transportation company: buses collect information about the passengers when reading the electronic tickets and send this information, along with their current position (learned from a GPS receiver) to the central system. The central system stores these data and updates both the company website and digital information screens at bus stops with real-time information on waiting times for the different bus lines. This technological system offers various advantages to the users: RFID tickets are faster to stamp, and real-time information on buses, also available through a mobile application, is precious for avoiding long waits at a bus stop. However, the amount of data collected by the system exposes the users to tracking. Personal tickets, such as monthly and yearly passes, are directly linked to a user, and both a government issued ID and the university card have to be showed upon purchase. Since each use is registered by the system, they allow the creation of a complete profile of the user's movements over the years. Many users may not find this problematic, but others are more concerned about their privacy. A privacy-concerned user can however opt for a different kind of ticket offered by the transportation company: an anonymous 10-ride pass. This ticket retains some of the advantages of monthly and yearly passes, such as being less expensive than a single-ride ticket (especially with a student discount), but can be bought anonymously at newsstands and convenience stores. The user might therefore reasonably expect tracking to be impossible, and consequently a better privacy protection.

3.1. Analyzing an Anonymous Bus Ticket

A 10-ride disposable MiMuovo ticket contains a Mifare Ultralight contactless integrated circuit (IC). Such an IC belongs to the cheapest memory-based contactless technologies commercially available. The ticket is nothing more than a 64-byte memory (EEPROM) readable and writable remotely, through a high-frequency radio channel. The IC does not contain any mechanism to protect the access to the memory or to ensure the confidentiality of the stored data. A security mechanism, though, allows anyone to lock memory areas such that write operations on these areas are no longer functional afterward. However, read-access to the memory cannot be prevented.

The interface of the IC relies on the widely used ISO-14443 standard, and the memory access is compliant with the well-known ISO-7816 standard. Reading a Mifare Ultralight IC is consequently quite easy using commercial readers and softwares. For example, an NFC-compliant smartphone with an appropriate application (e.g., *Tag Info*) is enough to read the memory of a MiMuovo ticket.

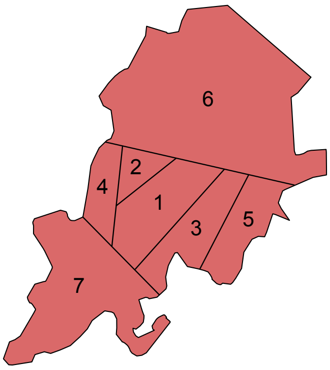

Figure 1: Cesena district partitioned into seven areas, according to which bus line each area is covered by. The bus and train station is in area 1.

Reading the memory actually means "obtaining a verbatim copy of the memory". The content of the memory is not encrypted but it must be decoded in order to retrieve the intelligible information. This can be easily done because the public transportation tickets usually contain common information, e.g., date, remaining trips, bus line, identity, location, and the encoding method is generally based on public standards, e.g., ISO-1545. Performing a differential analysis is usually enough to complete the full decoding of the ticket: in such an approach, the ticket is punched a few times, taking care that only one information (date, location, bus line,...) varies at a time. This allows identifying the fields encoded in the memory. Doing so, one can retrieve the information stored in a MiMuovo ticket. The memory actually contains 32 bytes of technical data whose write-access is partially locked, and 32 additional bytes that are freely modifiable. The latter bytes contain two 16-byte fields such that they are refreshed cyclically when the ticket is punched. Each field essentially contains the journey identifier, validation date and time, connection time, and also the geographical zone. The zone is a particularly sensitive information in terms of privacy.

3.2. Breaking Anonymity

We analyze the information collected by an anonymous ticket in its geographical context. In particular, we are interested in the topology of the bus network serving Cesena. As it is usually the case in small cities, Cesena is served by a number of bus lines, all of which have the central bus station as a starting point. From there, different lines branch out to reach different areas of the city.

In general, apart from the city center, no two lines cover the same area. This allows us to roughly divide the city in seven zones, according to which line they are served by. Zone number 1 is the city center, where the bus and train station are located, while zones numbered from 2 to 5 are neighbors covered by different bus lines. Zones 6 and 7 represent instead suburban locations served by provincial buses.

As most Italian city campuses, university sites and lecture halls are spread around the city, with each major having a different location. Our partition of the Cesena district also reflects this distribution: in particular, Psychology and Computer Science are located in two different buildings both in zone 1, Engineering is in zone 4, Architecture in 2, while Agriculture is outside of the city boundaries and therefore served by buses of zone 6. University buildings are usually best reached from one specific bus stop. Therefore, users of student-discounted tickets can be easily divided according to their major when a significant number of stamps are made at one of those bus stops. Moreover, the timestamps of those stamps can help an observer identify the year of study of a specific student/bus user: in fact, each class schedule is different depending on the year a student is enrolled in (in Figure 2, the bachelor in Computer Science).

If the student in question is lodged in university housing (which usually hosts students coming from outside the region, the most likely ones to use public transportation) profiling him based on his use of the bus pass can be even more successful. Places in the dormitories, located in zone 3, are assigned through a public selection whose results are available online. The concurrent analysis of this information, all of which is publicly available, with data collected through bus tickets by the transportation company can be enough to disclose the identity of the owner of a bus pass even when the pass is supposed to be anonymous, as in the case of 10-ride tickets.

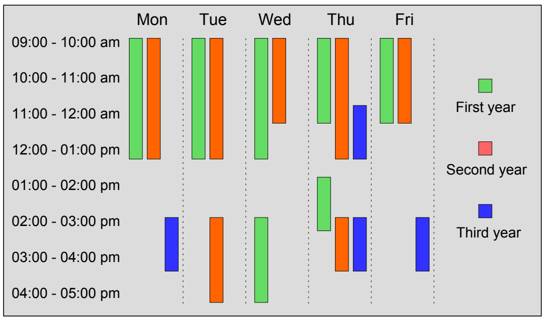

Figure 2: Time and location of classes make years distinguishable. Here, the schedule for the major in Computer Science.

In fact, according to the most recent report on public transportation published by the Italian national statistics institute, ISTAT, the number of university students regularly using public transportation in the city where their campus is located is 31% (Istituto Nazionale di Statistica (ISTAT), 2011). This figure is consistent with publicly available data on discounted tickets issued by Start Romagna for the city of Cesena and, combined with the numbers in Table 2, means that groups of users sharing the same characteristics (major, year of study, housing, ...) are composed of a handful of students only.

4. Plug-In Privacy Enhancements

As most of the privacy threats concerning pricing information can be overcome deleting related data after each invoice issuance, in the following we focus on raw data used for monitoring, statistics and similar purposes. The main aggregate information needed by public transportation providers in these cases are summarized in Table 3. Aggregates are basic information needed to compute meaningful statistics.

For example, if the bus company wants to analyze if the frequency of a bus line is adequate to demand, aggregate S1 is to be used. For this kind of aggregate it is enough to keep track of the bus line identifier and the specific ride identifier. Aggregates S2 and S3 are used instead in monitoring complex usage patterns, such as frequency and location of stopovers between bus lines during a single ride (S2) and behavior of the userbase on an outbound and then inbound journey (called coupled routes, S3). An abundance of the same entries on aggregate S2 may indicate that a direct link between one stop and another should be established. Aggregate S3 can instead detect patterns in the behaviors of the users. For instance, it lets the company understand the route a user usually takes during his working (or studying) day, allowing to study one-way and return patterns. In order to do that, the ticket identifier is also required. Finally, storing the ticket identifier alone is enough to know the usage load of a long-term subscription.

Recording this information, however, realizes the threat to privacy we discussed in the previous section. In Schwaig et al. (2013) the authors show that consumers usually consider corporations responsible for any inappropriate use of personal information. The organization management should therefore enforce adequate information privacy policies and promote an information systems emphasizing built-in privacy preserving features.

	Aggregate	Minimal field set	Privacy implications
S1	Passengers for each ride of a line	*Ride-id*, *Line-id*	Uncontextualized trip information. No threat.
S2	Routes (first stamp and changes)	*Ride-id*, *Line-id*, *Timestamp*, *Stop-id**	Exact trip information, but unlinked to the user. Reduced threat.
S3	Coupled-routes (returning ticket)	*Ride-id*, *Line-id*, *Timestamp*, *Stop-id**, *Ticket-id*	Aggregate requires all sensitive information. The ticket-id must therefore be removed or encrypted to prevent linking with other records.
S4	Journeys per ticket	*Ticket-id*	Number of rides counted. Implies the frequency a user travels.

Table 3: Aggregates typically used by public transport companies. Each aggregate needs a set of atomic data in order to be computed. Note that as public transport companies usually know at any time the exact position of each controlled vehicle (thanks to GPS) the field *Stop-id* can be inferred from the field *Timestamp*.

In fact, a well designed system architecture can prevent privacy threats while allowing the same computations to be performed. In Figure 3 we show three different approaches to the information exchange between the local recording point (e.g. the bus) and the central storage system. In the first model, the validation machine sends to the central server the field set required to compute the discussed aggregates at the time of the ticket punch. Data are transmitted in a single atom with the identifier unencrypted. This basic and unfortunately commonly used model does nothing to prevent the potential privacy breaches discussed above, as the identifier for the ticket or pass is directly linked to timestamps, bus lines/stops and so on.

We propose a more advanced and privacy friendly approach in the second system model. Here the same data are transmitted to the same remote storage, but in four different atoms. Each atom contains only the minimal data set required to compute a single aggregate. Delivering these atoms individually and at different times and encrypting the *ticket-id* (for aggregate S3) actively breaks the link between the identifier and other information and therefore enhances privacy. We note that, as aggregate S3 is intended to monitor daily patterns, the encrypted ticket identifier has to stay the same only for the duration of the day. Therefore, we could apply a one-way hash function on the string `"ticket-id"+"yyyy/mm/dd"`, producing unique, one-day encrypted identifiers for each different ticket.

Following these design principles and assuming that the hardware of ticket punch machines can be trusted, the *privacy by design* paradigm is achieved, as the privacy preserving properties are hard-coded into the system itself and the public transportation company only receives data that are natively anonymized. At the same time, the company is still able to compute meaningful statistics. For example, receiving the minimal field set related to atom S1 (*Ride-id*, *Line-id*) separately from other field sets, does not prevent computing the aggregate *Passengers for each ride of a line*, but prevents performing an implicit time-based analysis in order to link this information to information of other atoms.

Finally, the third system model reaches a similar privacy enhancement by transmitting atoms to different remote servers, controlled by different business units within the company or by different companies. In fact, as described by Jiang and Clifton (2006), distributing data among autonomous and independent sites provides protection for individual data. The assumption here is that different controllers do not collude with each other by sharing information. In our case, performing a vertical partitioning of data concerning aggregates S1, S2 and S3 (and encrypting the *ticket-id* in S3 at the server side) on one side and aggregate S4 on the other, prevents profile reconstruction by a `JOIN` operation on the atoms timestamps or other sensitive fields which could disclose the ticket identifier relative to specific routes, timings and positions. To deploy such a system on an existing platform, it is sufficient to partition the previously collected data and to start recording new ones in separate business units. The most common choice would be to deploy a relational database via a hosting service. In order to do that it would be advisable to consider a database instance with a read replica designed for disaster recovery. The database should be able to support write-intensive policies and should be up and running 24 hours a day. Focusing on *Atom 4* and considering a 64 byte record required to store a single atom, a municipality reporting 1 million validations per day would produce approximately 2 GB data monthly. However, this model does not reach the *privacy by design* paradigm, as the level of privacy achieved is strictly related to the trust implied in the parties managing the information,

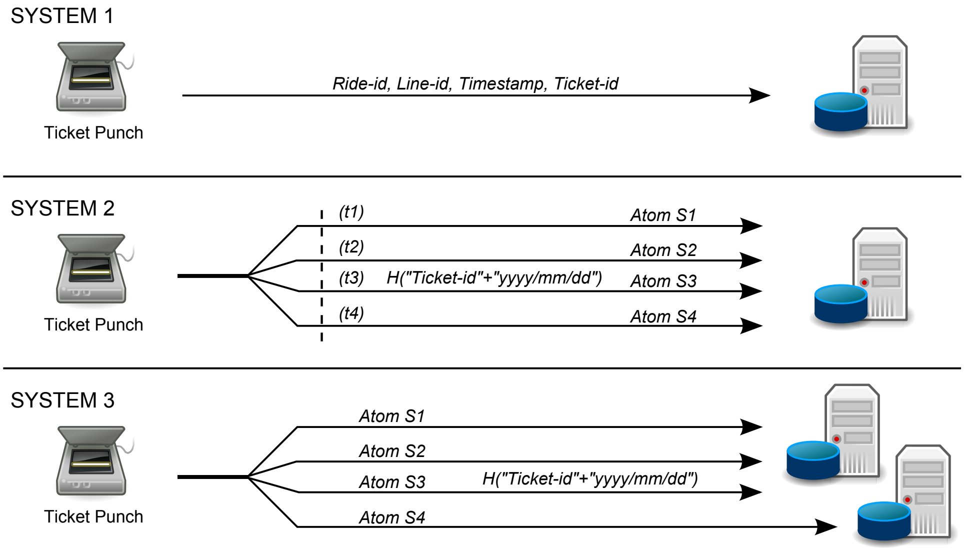

Figure 3: Different models for raw information management in the public transport scenario as discussed in Section 4.

for the whole lifetime of the system.

4.1. *Costs and potential disadvantages of the proposed solutions*

The proposed systems introduce several benefits concerning user's privacy, but also come with some potential disadvantages in terms of costs of deployment and data processing limitations. In order to deploy the second system proposed in Figure 3, it is mandatory to embed into the ticket punch machines a *jitter* that introduces randomness in atom delivery times and an algorithm for computing a secure hash function. These changes usually require a firmware upgrade. However, should the currently used processor not be capable of performing such tasks, the hardware upgrade costs are also to be considered. Depending on the complexity of the modifications required by the ticket punches and the data processing strategies applied, an implementation cost should also be added. For these reasons, the third system is preferable when privacy needs to be achieved in an already existing system, as it does not imply the costs associated with the upgrade of validating machines already installed on vehicles. In this case, the only measure involving ticket punches in an existing system would be a firmware update in order to send *Atom 4* to a different server. The most meaningful cost introduced in this model is data hosting, a service we assume is provided by a third company. In general, these costs depend on the current status of the system used by the company. For instance, buying new compliant ticket punches could be less cost effective than to upgrade existing ones. We sum up this considerations in Table 4, where we propose two different scenarios: in the first one, the transport company is about to deploy an entirely new system, while in the second one the company decides to upgrade the existing system instead.

	Build from scratch	System upgrade
System 2	System design	System design, hardware upgrade or purchase, firmware upgrade, server-side processing update
System 3	Data hosting	Slight firmware upgrade, data hosting, server-side processing update

Table 4: Costs introduced adopting the privacy preserving systems proposed in Section 4. In the *build from scratch* case we exclude the purchase of ticket punches as they should be buyed even if a traditional system would be adopted.

Concerning limitations on data processing, both systems preserve the ability to compute all the aggregates proposed in Table 3. Problems may arise if the public transport company attempts to compare routes belonging to the same anonymous ticket on a basis of more than one day, in order to study the user's movements and learn more complex patterns. This is however exactly the reason transport companies should introduce such systems: to guarantee users that their data are not used for malicious tracking and increase customers' trust in public transport systems, and, consequently, encourage more widespread usage of public transport itself.

5. Conclusion

The amount of intelligent technology that we use every day, knowingly or not, has seen in recent years an exponential increase. Sensors, smart cards and many other intelligent devices are now deployed all around us, and take an active part in shaping our lifestyles. The amount of information generated or collected by these devices is enormous, and the fall in prices and great increase in capability of storage devices, made possible what would have been unimaginable before: keeping this information indefinitely. While these technological developments greatly benefit our life, we are still only beginning to grasp their side effects. The most common and obvious concern is about privacy: what control do we have on our personal information, once it has entered the system? Can we really remain anonymous if we want to?

In this paper we analyze the risks of collecting and storing passengers information in the context of public transport. We not only show that real anonymity is almost impossible to achieve even when using anonymous passes, but we also show how the actual amount of data needed to identify (with some statistical error margin) and track a user is surprisingly low. In order to show the realism of this threat, we perform a detailed hardware analysis of the smart card tickets used by the transportation company Start Romagna (active in the city of Cesena, Italy), and we identify the information stored in the chip after each use of anonymous, disposable 10-ride tickets. This allows us to learn what information is collected at each validation and, since these data are then sent to the company databases and stored, to have a clear picture of what information is available to the public transportation company.

In light of this knowledge, we analyze the possible uses of this information: we discuss both the legitimate goals of monitoring traffic in order to improve the service and collect statistics by the company and the malicious tracking capabilities of an evil actor. We also analyze the current implementation of the information system that collects, stores and processes this information, and its privacy implications. Since the aim of this study is to improve the state of the art and build privacy-preserving public transport information systems, we propose management recommendations and system models that can achieve the same legitimate goals in a more privacy-friendly way, without any need of modifying the current infrastructure.

While the threat we discuss in this work is specific to the infrastructure of public transport in Cesena, this experiment brings to light problems that are common to all e-ticketing systems that store travel histories. By showing how to de-anonymize travel histories of the passengers, we stress how important it is to get prior informed consent of the users before disclosing or selling even anonymized extracts of such data. Having been able to recognize users by comparing only to limited, publicly accessible information, we also pose the problem of what could be actually achieved in case the same analysis was to be performed by companies or other entities that have also access to data generated by social media (social networks, location-aware applications, etc.). We can only speculate how big a threat to users' privacy the unchecked sale of travel histories might become.

Acknowledgments

The authors would like to thank LAB ID s.r.l. and, in particular, Francesco Pirola and Daniele Zanoli for the useful information provided on the MiMuovo smart card.

References

Asadpour, M., Dashti, M. T., 2011. A privacy-friendly RFID protocol using reusable anonymous tickets. In: 10th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (IEEE TrustCom-11). IEEE Comp. Soc., Changsha, China, pp. 206–213.

Avoine, G., Martin, T., Szikora, J.-P., March–April 2010. Lire son passe navigo en un clin d'œil. Multi-System & Internet Security Cookbook – MISC 48.

Barbaro, M., Zeller, T. J., 2006. A face is exposed for aol searcher no. 4417749. *The New York Times*, 9 August 2006. Avbl.: `http://www.nytimes.com/2006/08/09/technology/09aol.html`.

Calderoni, L., Maio, D., Palmieri, P., 2012. Location-aware mobile services for a smart city: Design, implementation and deployment. JTAER 7 (3), 74–87.

de Grange, L., Gonzlez, F., Muoz, J. C., Troncoso, R., 2013. Aggregate estimation of the price elasticity of demand for public transport in integrated fare systems: The case of transantiago. Transport Policy 29 (0), 178 – 185.

Diaz, C., Gürses, S., 2012. Understanding the landscape of privacy technologies. In: Proceedings of the Information Security Summit '12. pp. 58–63.

Geuss, M., 2013. Japanese railway company plans to sell data from e-ticket records. *Ars Technica*, 7 July 2013. Avbl.: `http://arstechnica.com/business/2013/07/japanese-railway-company-plans-to-sell-data-from-e-ticket-records/`.

Heydt-Benjamin, T. S., Chae, H.-J., Defend, B., Fu, K., 2006. Privacy for public transportation. In: PET 2006. Vol. 4258 of Lecture Notes in Computer Science. Springer, pp. 1–19.

Istituto Nazionale di Statistica (ISTAT), 2011. Focus: Trasporti urbani. *Istat*, 5 April 2011. Avbl.: `http://www.ontit.it/opencms/export/sites/default/ont/it/documenti/files/ONT_2011-04-06_02603.pdf`.

Jiang, W., Clifton, C., 2006. A secure distributed framework for achieving *k*-anonymity. VLDB J. 15 (4), 316–333.

Kerschbaum, F., Lim, H. W., Gudymenko, I., 2013. Privacy-preserving billing for e-ticketing systems in public transportation. In: Workshop on Privacy in the Electronic Society.

Liu, Z., Bonazzi, R., Fritscher, B., Pigneur, Y., 2011. Privacy-friendly business models for location-based mobile services. J. of Theoretical and Applied Electronic Commerce Research 6 (2), 90–107.

Peng, K., Bao, F., 2010. A secure rfid ticket system for public transport. In: Foresti, S., Jajodia, S. (Eds.), DBSec. Vol. 6166 of Lecture Notes in Computer Science. Springer, pp. 350–357.

Sadeghi, A.-R., Visconti, I., Wachsmann, C., 2008. User privacy in transport systems based on rfid e-tickets. In: Bettini, C., Jajodia, S., Samarati, P., Wang, X. S. (Eds.), PiLBA. Vol. 397 of CEUR Workshop Proceedings. CEUR-WS.org.

Salonen, J., 2011. Evaluating the security and privacy of near field communication - case: Public transportation. In: Jung, S., Yung, M. (Eds.), WISA. Vol. 7115 of Lecture Notes in Computer Science. Springer, pp. 242–255.

Sampaio, B. R., Neto, O. L., Sampaio, Y., 2008. Efficiency analysis of public transport systems: Lessons for institutional planning. Transportation Research Part A: Policy and Practice 42 (3), 445 – 454.

Schwaig, K. S., Segars, A. H., Grover, V., Fiedler, K. D., 2013. A model of consumers' perceptions of the invasion of information privacy. Information & Management 50 (1), 1–12.

Tseytin, G., Hofmann, M., Lyons, D., O'Mahony, M., 2006. Tracing individual public transport customers from an anonymous transaction database. J. of Public Transportation 9 (4), 47–60.