Privacy-Preserving Design of Data Processing Systems in the Public Transport Context

Franco Callegati
DEI - Università di Bologna
Via Venezia, 52 - 47521 Cesena, Italy
franco.callegati@unibo.it

Aldo Campi
DEI - Università di Bologna
Via Venezia, 52 - 47521 Cesena, Italy
aldo.campi@unibo.it

Andrea Melis
DEI - Università di Bologna
Viale del Risorgimento, 2 - 40136 Bologna, Italy
andrea.melis6@unibo.it

Marco Prandini
DISI - Università di Bologna
Viale del Risorgimento, 2 - 40136 Bologna, Italy
marco.prandini@unibo.it

Bendert Zevenbergen
OII - University of Oxford
1 St Giles' - Oxford OX1 3JS, UK
bendert.zevenbergen@oii.ox.ac.uk

Abstract

The public transport network of a region inhabited by more than 4 million people is run by a complex interplay of public and private actors. Large amounts of data are generated by travellers, buying and using various forms of tickets and passes. Analysing the data is of paramount importance for the governance and sustainability of the system. This manuscript reports the early results of the privacy analysis which is being undertaken as part of the analysis of the clearing process in the Emilia-Romagna region, in Italy, which will compute the compensations for tickets bought from one operator and used with another. In the manuscript it is shown by means of examples that the clearing data may be used to violate various privacy aspects regarding users, as well as (technically equivalent) trade secrets regarding operators. The ensuing discussion has a twofold goal. First, it shows that after researching possible existing solutions, both by reviewing the literature on general privacy-preserving techniques, and by analysing similar scenarios that are being discussed in various cities across the world, the former are found exhibiting structural effectiveness deficiencies, while the latter are found of limited applicability, typically involving less demanding requirements. Second, it traces a research path towards a more effective approach to privacy-preserving data management in the specific context of public transport, both by refinement of current sanitization techniques and by application of the privacy by design approach.

Keywords: Privacy, anonymization, public transport, data analysis

Introduction

The current trend in management of public transport systems is to outsource services to multiple private operators, requiring them to integrate their ticketing and fare system with one another. When this concept is introduced in regions that used to have single local entities managing every aspect of ticketing, or that conversely left operators free to adopt incompatible, separate ticketing systems, a new layer of coordination must be put in place.

There are many examples of this kind of approach around the world, such as the Oyster card system in London, the Octopus card system in Hong Kong, or the Istanbulkart in Istanbul just to name a few. As a case study, this paper considers the Emilia-Romagna region of Italy, where the Regional Government has been running for several years a project to integrate the control processes of the various transport companies operating in the region. These companies typically operate over disjoint territories, and they used to manage independent and localized ticketing systems. The trend with the new regional system is to go more and more towards integration of tariffs, routes and ticketing, so that the citizen may buy a ticket in city by a given operator and use it in another city with another operator. While providing an improved service to citizen this approach also brings some additional burden, since a clearing system is needed to share the revenue of tickets sales according to the actual service each operator has provided (and thus, supposedly, to the real costs it has incurred).

Data detailing every trip, collected by public transport operators, previously confined to internal use only, now must be shared and can potentially harm passengers. The system that manages it does not merely need to control data disclosure, but has to be designed to manage potential risks during the collection and processing of data. This is a challenging task, which must manage privacy risks appropriately on the one hand, and preserve data utility to a level that guarantee usefulness for clearing purposes on the other hand. This paper illustrates the work the authors are doing to design and test the clearing system in a way that safeguards the protection of personal information, not as a result of some policy superimposed to the existing functions, but rather taking into account this requisite from the start, by applying the principles of Privacy by Design.

The manuscript is organized as follows. In Section 2 the local context and the general ideas behind the clearing system are briefly presented and reviewed, and research questions are stated. Section 3 gives an overview of the general principles of data sanitization for the purpose of safe release of sensitive information. Section 4 illustrates the risks connected with the release of sensitive information, focusing on the desanitization attacks that exploit public data sources, and Section 5 gives two examples of how these attacks can affect the clearing datasets. Section 6 describes the general principles of Privacy by Design, and outlines the direction of current and future work to apply them to the scenario of the clearing system, before conclusions are draft in Section 7.

The Clearing System Scenario

The Local Context

The Emilia-Romagna Region has approximately a population of 4.5 Million with an area of 22500 square Kilometres. About half of the population leaves in the 13 main cities that are lined along the ancient Roman road called “Via Emilia”[1] which gives its name to the Region. Emilia-Romagna is highly industrialized with a number of

[1] The *Via Aemilia*, named after the Roman consul Marcus Aemilius Lepidus, was completed in 187 BC and runs from Piacenza, in the central part of the largest plain (Pianura Padana) in northern Italy, to Rimini on the Adriatic sea shore in an almost straight line for about 250 km.

https://aisel.aisnet.org/pajais/vol7/iss4/4
DOI: 10.17705/1pais.07403

companies typically spread along the Via Emilia around the main urban centres. For this reason the mobility infrastructures are a key part of the logistics supporting the economy of the region.

The estimate of daily trips made by Emilia-Romagna citizens for work or leisure is about 9 Millions of which 25% walk or bike and 8% by public transport means. The public transportation system is built around a rail backbone basically parallel to the Via Emilia which links the main urban centres. Local transportation systems in the cities mostly use buses. The local systems are run by four large operators and a few small operators on specific routes.

The policy maker is the regional Mobility and Transport Councillorship which is competent, among many subjects, for the planning of the infrastructural network, regional and local mobility systems. Over the last decade the Councillorship pursued service integration and multi-modality of public mobility systems, promoting, in particular, the deployment of regional integrated fares with an investment of about 20 M€ in supporting hardware (central control systems, ticketing machines, vehicle monitoring systems etc.).

The issue of the *MiMuovo (I move)* chip card was the flagship project of the fare integration process, supporting multi-modal tickets valid over a given path spanning several operators and transport means. For instance a user holding a MiMuovo card with an integrated travel contract is allowed to use the bus (run by operator A) in his home town to reach the railway station, the train (run by operator B) to his/her working town and the bus (run by operator C) to his/her working place. To date about 300,000 MiMuovo cards have been deployed and are used daily.

Today the Councillorship is also fostering fare integration for single trip tickets that can be bought in any town and used in any other within the Region. This requires the operator selling the ticket to share the revenue with other operators, if they are involved in its use. This is called *clearing* process, and has to be implemented in a way accepted fairly by the whole set of operators involved, to guarantee the integrated system sustainability.

The Clearing System

The clearing system is based on a distributed architecture in which each operator is responsible for the management and maintenance of its own data. The data needed for the computation of the clearing function is collected in a clearing database located in a central processing centre, operated by a regional in-house company, in order to guarantee neutrality and to avoid disturbing the production systems of the operators.

The creation of the clearing database requires the sharing of the operators dataset in a standard, machine readable format, thus creating a possible threat as a consequence of secondary uses. Moreover the regional Councillorship aims at using the data for in-depth analysis of the transport system performance. Eventually, part of the datasets could be released to the public as open data.

Operators and public bodies do not have any effective control over future uses of their dataset once it is publicly available. Unfortunately the data about sales and usage may reveal issues the operators consider part of their industrial secrets and/or sensitive information in terms of personal privacy.

This problem can be (partly) mitigated by applying full anonymization safeguards, which is very difficult when the utility of the database is to be maintained. Moreover, it is possible to adequately inform the involved subjects of the intention to disseminate the dataset in an open data format, alerting them to potential risks, but this action can limit the degree of user acceptance, especially if the policy intentionally leaves open what kind of secondary uses of their data will be done. Therefore a trivial solution to the issue does not exist.

Research Questions

From the point of view of the users, in the widest sense that encompasses passengers, operators, and regulators, the most pressing questions to be answered are: with the current data storage and utilization processes, is it possible to breach data privacy and re-identify the data subjects? Which kind of processing and links to people and business-related issues are possible, for example by matching the clearing database data with other external databases? From the point of view of the researchers, these questions can be answered by analysing the underlying scientific and technical tools: what features do the current data sanitization algorithms exhibit? Is it possible to measure their effectiveness in any given scenario? Symmetrically, can the experience gathered from similar projects in other cities/regions point our research in the right direction, or are our requisites too specific?

Once the background analysis is complete, if it highlights deficiencies either in the basic technologies or in their application to specific scenarios, the research activity will be directed towards the definition of a more effective framework for public transport data sanitization.

Sanitization: A Critical Overview

The architecture outlined in Section 2.2 introduces two possible security attack vectors. The first one is the intrusion of an unauthorized party, in which data are subtracted from the primary database; this is a classical issue of information security and access control, and this work does not deal with its direct form; yet, it takes into account the similar situation of purposely releasing data for public use, considering that it could be enriched though correlation with external data sources, to the extent of disclosing details that should not be made public. The second one is called an insider attack, also referred to as an insider threat; this type of attack arises due to a malicious threat from somehow authorized actors, from inside the organizations that are legitimately involved in data collection and processing; the next chapter illustrates ways to perform this kind of attacks and corresponding effects.

A data sanitization phase is commonly proposed in the literature as the necessary step to prevent these issues; this phase as defined by Crawford et al. (2007), is "the process of altering [a dataset] so that it remains usable for beneficial purposes, while minimizing its use for harmful purposes". To properly define this process, the key issue to deal with is to understand what "keeping the beneficial purposes" and "minimizing the harmful purposes" mean. Ideally, the process should be able to manipulate the data in a way that prevents privacy attacks but at the same time preserves the possibility of performing many kinds of economic computations. To progress towards this goal, the existing literature is analysed to find (a) whether convincing measures of utility and vulnerability of the dataset exist and (b) whether existing algorithms result in positive trade-offs when applied to our context.

The literature was analysed as follows. Starting from the basic requisite of having to anonymize the data, the generally-applicable techniques of data anonymization were reviewed, highlighting their limitations.

Next, the application of these techniques to the clearing scenario was attempted, taking into account the literature on the evaluation of these techniques, namely verifying the impact of their known limitations, and introducing metrics that allow to determine whether a satisfying level of privacy is attained.

Subsequently, the symmetrical path was followed, starting from similar cases for which documentation of the process of transport data privacy protection exists, such as those of Montreal and Amsterdam. However, the analysis of the various aspects highlighted that, even though there

https://aisel.aisnet.org/pajais/vol7/iss4/4
DOI: 10.17705/1pais.07403

are important points of contact, these experiences did not need to take into account some requirements that turn out to be of critical importance for the clearing scenario. As a consequence, the techniques devised for those systems would leave ours subject to numerous types of attack, both of kinds already known in the literature, and of other kinds described in this paper as proofs of concept.

For each section reviewing a specific subject, a table is provided at the end, summarizing the most relevant literature sources, their contribution, and the open issues (both in terms of intrinsic limitations of the methods and of gap between the methods and the requirements of the specific scenario of this paper).

General-Purpose Sanitization Approaches in the Literature

As a preliminary consideration, to understand the way algorithms manipulate datasets to achieve the aforementioned results, it is useful to note that every approach is based on the classification of data elements according to the potentially sensitive information in three categories (Ranjit and Acharya, 2008; Zevenbergen et al., 2013):

Identifier attributes (or identifiers) can individually distinguish the data subject more or less directly. Typical identifiers include: name, address, social security numbers, mobile phone number, IMEI number.

Quasi identifier (or key) attributes can be used to identify a data subject using auxiliary sources of information, by linking to databases that contain identifying information. They are indirect identifiers of a data subject, which make an individual more distinctive in a population. Typical key attributes include: age, race, gender, date of birth, and place of residence.

Sensitive (or secondary) attributes cannot individually identify a data subject directly and may require significant amounts of auxiliary data to be useful for re-identification purposes. A data subject may then be identified individually through more sophisticated methods such as fingerprinting, rather than mere linking of databases. Examples include settings in an application, battery level measured over time, or location patterns.

In summary, the literature describes four main techniques of data anonymization.

***k*-anonymity** (Sweeney, 2002; Ciriani et al., 2007) is the most well-known technique for *generalization*. The basic principle here is to replace exact values with ranges, wide enough to guarantee that every attribute in a database appears with identical values in a given number of other rows, forming a group of *k* rows indistinguishable from each other. This approach may take the form, for example, of grouping subjects' locations into sufficiently large areas such that no set of locations is unique to any individual. The enforcement of *k*-anonymity requires the preliminary identification of the quasi-identifier. The quasi-identifier, as previously defined, depends on the external information available to the recipient, as this determines her ability to make correlations (not all possible external data sources are available to every possible data recipient); different quasi-identifiers can potentially exist for a given table. Many variations and improvements exist, yet *k*-anonymity techniques cannot hide whether an individual is in the dataset, and it performs poorly in protecting sensitive attributes against attacks based on background knowledge or on the knowledge of the details of its application. (L. Sweeney, 2002)

***l*-diversity** (Machanavajjhala et al., 2006) is an improvement of *k*-anonymity that require the sensitive attribute associated with each quasi-identifier to appear at least with *l* different values. Other refinements have been proposed, but as described also in Pingshui et al. (2013) processing a large dataset to achieve *l*-diversity is time-consuming and vulnerable

to inference attacks in presence of a series of updated publications of the same dataset, if it is simply re-anonymized with the same approach every time. Finally the added requirements proved to be neither necessary nor sufficient to prevent sensitive attribute disclosure (Li et al., 2007).

An example of the effects of the application of *k*-anonymity and *l*-diversity techniques on the dataset that is the object of our study is shown in Figure 1.

***t*-closeness** (Machanavajjhala, 2007) To improve robustness of *k*-anonymity, the same authors of *l*-diversity also proposed a privacy notion called *t*-closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should not be greater than a threshold *t*). In other words, an equivalence class is said to have *t*-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold *t*. A table is said to have *t*-closeness if all equivalence classes have *t*-closeness.

Differential Privacy is a process derived from cryptography. As defined in Roth (2014), it "aims to provide means to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying its records." Unlike other methods, differential privacy operates off a solid mathematical foundation, making it possible to provide strong theoretical guarantees on the privacy and utility of released data. The most used technique is called ε-differential privacy and it is modelled via a randomized algorithm; a theoretical definition is given in (Neustar, 2014) and summarized as follows.

	User Zip	Bus Line	Contract
1	47677	29	Day
2	47602	22	Month
3	47678	21	Month
4	47905	41	Day
5	47909	45	Year
6	47906	89	Day
7	47605	12	Week
8	47673	14	Day

Table 1. Original Public Transport Dataset

	User Zip	Bus Line	Contract
1	476**	2*	Day
2	476**	2*	Month
3	476**	2*	Month
4	4790*	>=40	Day
5	4790*	>=40	Year
6	4790*	>=40	Day
7	476**	<= 20	Week
8	476**	<= 20	Day

Table 2. The dataset 3-Anonymization version

	User Zip	Bus Line	Bus Zone	Contract
1	47677	29	300	Day
2	47602	22	400	Day
3	47678	27	500	Month
4	47905	43	550	Week
5	47909	52	570	Month
6	47906	47	550	Year
7	47605	30	400	Day
8	47673	36	800	Month
9	47607	32	90	Year

Table 3. Original Public Transport Dataset

	User Zip	Bus Line	Bus Zone	Contract
1	476**	2*	300	Day
2	476**	2*	400	Day
3	476**	2*	500	Month
4	4790*	>= 40	550	Week
5	4790*	>= 40	570	Month
6	4790*	>= 40	550	Year
7	476**	3*	400	Day
8	476**	3*	800	Month
9	476**	3*	90	Year

Table 4. The Table 3-diverse version

Figure 1 - Effects of anonymization techniques on the transport dataset

https://aisel.aisnet.org/pajais/vol7/iss4/4
DOI: 10.17705/1pais.07403

"A randomized function K gives ε-differential privacy if for all data sets D and D′ differing on at most one row, and all S ⊆ Range(K),"

$Pr[K(D) \subseteq S] >= exp(\varepsilon) \times Pr[K(D') \subseteq S]$

This formula can be interpreted as stating that the risk to one's privacy should not substantially (as bounded by *ε*) increase as a result of participating in a statistical database. Namely, that an attacker should not be able to learn any information about any participant that they could not learn if the participant had opted out of the database. This goal is pursued by adding some noise to the result of a query on the dataset. There exist many different mathematic mechanisms to do that; the most commonly seen in this context is the Laplace mechanism, which adds noise derived from the Laplace distribution. It has only one parameter, defining the standard deviation, or *noisiness*. This parameter should have some dependence on the privacy parameter, *ε;* it should also depend on the nature of the query itself, and more specifically, the risk to the most different individual of having their private information teased out of the data.

Differential privacy comes in many different forms and variations which have not been covered in detail, but they all have several limitations, due in particular to the high computational complexity that the cryptographic techniques could introduce in a big dataset. The main advantage of this approach is that its mathematical foundation makes it possible to actually measure the strength or the weakness of the results. The concept of differential privacy holds much potential, and is still the topic of active research.

Table 1 - summary of general-purpose sanitization techniques

SOURCE	SUBJECT	CONTRIBUTION
Machanavajjhala et al.; 2006	k-anonimity	Practical application
Sweeney; 2002		General description
Machanavajjhala; 2007	l-diversity	General description of the techniques, and detailed explanation of the improvements over k-anonymity
Li et al.; 2007;		
Ciriani et al.; 2007;		
Machanavajjhala; 2007	t-closeness	General description of the techniques, and detailed explanation of the improvements over k-anonymity
Pingshui et al.; 2013;		
Roth; 2014	Differential Privacy	General description and guidelines for the application of the procedure
Neustar; 2014		

Sanitization in the Clearing Scenario

To ascertain the suitability of the illustrated techniques to the clearing scenario, the first step is to define the correct evaluation criteria, which could depend from:

1. The context of collection and usage of the data;
2. The structure of the data;
3. The objective of data processing.

As an example, the work of Brickell and Shmatikov (2008) measures the trade-off between privacy (i.e., how much the adversary can still learn from the sanitized records) and utility (i.e., the residual accuracy of data-mining algorithms executed on the sanitized records with respect to what could be found for legitimate purposes from the original data set). Their paper showed that *k*-anonymity provides no privacy improvement on the tested dataset; furthermore, *l*-diversity is no better than trivial anonymization. Another interesting work Cormode et al. (2013) tries to quantify the effectiveness of sanitization in terms of privacy impact of a data release. To this end,

the study introduces the idea of incorporating a metric over 'privacy breaches' based on a notion of empirical privacy, and evaluating the corresponding empirical utility of the released data. The measure of a privacy breach is defined as the increase in correct a-posteriori inferences obtained by an adversary about sensitive values in the data, using a Bayesian classifier with previous knowledge. The cited paper applies this metric to the main four techniques, and concludes that differential privacy often provides the best empirical privacy for a baseline utility level, but that for increasing utility levels it can be preferable to adopt methods like *t*-closeness or *l*-diversity. There are other works that pursue the same kind of investigation, that mainly derive as a conclusion the weakness of *k*-anonymity and *l*-diversity algorithms.

The main limitation of the reviewed literature is that only few papers interact with large amounts of data derived from public transport system. An exception is the paper by Ghasemzadeh, Fung, Chen, Awasthi (2013), which aims at preventing privacy attacks in a general sense, especially those damaging from a user's perspective. The proposed solution is an algorithm based on the LK-privacy model, using the approach of "identifying the LK-privacy requirement, and then eliminating the violating sequences by a sequences of suppressions with the goal of minimizing the impact on the structure of the user tracking." What the authors claim is that their anonymization algorithm thwarts identity record linkages, while effectively preserving the information quality in terms of its suitability for the generation of a probabilistic flow-graph. It is a very interesting result, yet insufficient in the scenario of a clearing system, where the user's privacy perspective is not the only one that must be protected; in fact, the insider threat is not taken into account.

Eventually, with the exception of differential privacy (which cannot be easily applied to huge amounts of data), it would seem that not a single sanitization solution is really effective. Actually, these studies demonstrate only how these techniques are not effective enough for the particular context taken into consideration. To properly evaluate their potential in our scenario, a more precise definition is needed for various characteristics, namely:

1. the privacy requirements;
2. the expected level of utility of datasets;
3. how to measure the effectiveness of algorithms at preserving these properties.

As regards privacy requirements and utility levels, it is possible to reason on the structure of sensitive values and quasi-identifier in our case study, represented in Figure 2. The sensitive values are the user's identity and location data; these values are the one to hide and protect. The means of transportation and the user's contract data, otherwise, can be classified as quasi-identifier values, since they could become sensitive if crossed with other information; at the same time, these are the data needed to calculate the clearing functions, so they cannot be depleted because of the precise information they carry. The goal of the anonymization process is to break the link between user identity and location, and to mask the QI values in a way that preserves the values needed for the clearing system.

As regards the metric used to quantify privacy, there is very little literature. Atzori et al (2007) created a metric called *δ*-presence to evaluate the risk of identifying an individual in a table based on generalization of publicly known data. This work shows that existing anonymization techniques are inappropriate for situations where *δ*-presence is a good metric (specifically, where knowing that an individual is in the database, as it very often happens to travellers of public transportation networks). So, despite its quality in general terms, this metric cannot be used in our context. Otherwise, the metric discussed in Ganta et al. (2008) is defined as the amount of "useful" data mining queries still existing

https://aisel.aisnet.org/pajais/vol7/iss4/4
DOI: 10.17705/1pais.07403

after the sanitization phase. As shown in the following section, an insider threat attack in our context is likely to take the form of a search pattern analysis; for this reason this measure of privacy seems to be more interesting.

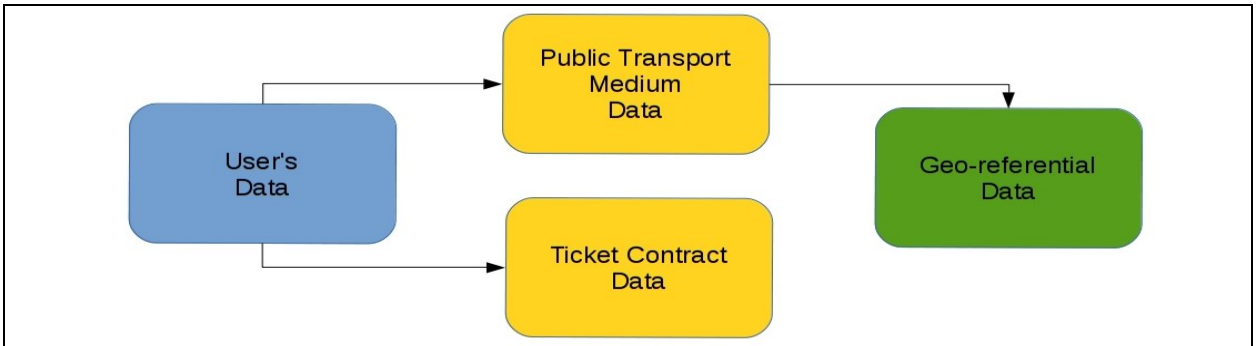

Figure 2 - Sensitive data elements and their logical connections

Table 2 - strengths and limitations of sanitization techniques applied to the public transport scenario

SOURCE	SUBJECT	CONTRIBUTION
Bishop et al.; 2010	k-anon., l-div., t-closeness limits and vulnerabilities	Outline the limits of the discussed techniques, and illustrate possible attacks (not tied to specific scenarios)
Barbaro and Zeller; 2006		
Atzori et al; 2007	Metrics to evaluate the known anonymization techniques in terms of amount of privacy and data utility	Introduces the delta-presence metric, which is useful to compare anonymization techniques in terms of effectiveness in hiding the presence of an individual in a dataset
Ganta et al.; 2008		Introduces a metric to measure the utility of the dataset after anonymization, in terms of feasible queries
Brickell and Shmatikov; 2008		Other works about the utility of the queries after anonymization using data mining techniques
Cormode et al.; 2013		

Threats

Known Attack Scenarios Against Anonymization

As illustrated in Section 3, the residual presence of one or more sensitive elements in a dataset is structural, both because their complete obliteration would remove any utility from the dataset and because most sanitization techniques have intrinsic limitations. There are various motivations driving attackers to exploit any possible data source to un-conceal information, as described for example by Narayanan and Shmatikov (2008) and summarized by Bishop et al. (2010). Government agencies are more and more involved in extensive surveillance and are eager to collect any kind of information, even remotely connected with individuals. Marketing campaigns often exploit behavioural targeting of advertisements, for which the construction of networks and the highlighting of patterns is essential. Investigators, stalkers and employers may want to target specific individuals, possibly starting from a vantage point in terms of background and context information.

In our scenario, the clearing datasets could be exploited by any of these categories wishing to infer information regarding the operators' business and/or the passengers of the public transportation system. Moreover, there is a specific insider threat: the participating companies may try to use data analysis to gain competitive advantages, both of the kind usually associated with market analysis (e.g. uncommonly profitable routes) and of the kind usually regarded as a trade secret (e.g. the optimization of the allocation of resources such as buses, trains, and personnel on board).

The main issue is that when pursuing their goals, adversaries are not limited to the analysis of the clearing data alone. Conversely, they can reap great benefits through correlation with many existing public databases. The first widely known case of identification through correlation of different public datasets dates back to 2006, when AOL released anonymous data about search queries and New York Times reporters were able to find the real name linked to the pseudonym 4417749 (Barbaro and Zeller, 2006). As noted by Bishop et al. (2010) this case also shows a peculiar effect of the failure of the privacy protection: since the user acted as a proxy for friends with no Internet access, her name was associated to many queries unrelated to her condition and habits. The same could happen with public transportation data. As an example, if zones of boarding and alighting are kept wide enough to conceal the exact location of a passenger, they could end up enclosing points of interest (hospitals, schools, recreational facilities, shopping districts, etc.) which could lead an attacker to draw wrong conclusions, possibly even more damaging to the victim than the correct ones.

While the AOL attack exploited various public records, a more recent episode targeted social networks (Narayanan and Shmatikov, 2008) exploiting correlated data from two networks attracting the same community of movie enthusiasts (the Netflix Prize and IMDB) to link user identities between the two datasets by correlation of their preferences. Social networks attract vast numbers of users, they collect every kind of personal information about them, and they can be conveniently searched by algorithms, thanks to the APIs provided to foster their growth by turning them into platforms for the development of social games and applications. Two recent studies regarding Twitter can be useful to understand the implication of leaving this kind of *digital footprints*. The *WhACKY!* application "harnesses the multi-source information from tweets to link Twitter profiles across other external services [. . .] to map Twitter profiles to their corresponding external service accounts using publicly available APIs." (Correa et al., 2012). Their study highlights how much Personally Identifiable Information (PII) can be programmatically gathered from social networks[2], as reported in Table 3, and how the correlations can fill the gaps to draw a complete picture of an individual. (Calvino, 2015 - in Italian - translates as "Stalking John Doe: surveillance, privacy and proximity in the age of Twitter") demonstrated how the correlation between the Twitter activity of an Italian user and the publicly available census data allows to reduce the uncertainty about the real-world identity and location of the victim to a mere 1-in-789 inhabitants of an area just 2500 square meters wide.

It is worth noting that Golle and Partridge (2009) already studied the correlation of commutes with publicly available census data back in 2009, spurring speculation about how the increased use of location-capable devices would affect privacy (Narayanan, 2009). After four years, de Montjoye et al. (2013), working on location data from mobile telephone carriers, were able to claim that "four spatio-temporal points are enough to uniquely identify 95%

2 Many other sites yield less PII, yet can be used to link users with their location; just to give a few examples: Noisetube, FixMyStreet, OpenStreetMap, Panoramio

https://aisel.aisnet.org/pajais/vol7/iss4/4
DOI: 10.17705/1pais.07403

of the individuals" and that "even coarse datasets provide little anonymity". A final example of the risks associated with location data is carried by the recent study of possible privacy breaches as a consequence of the traces left when renting bikes in London. This is a very relevant case for this paper, since the kind of data used in the studies described hereinafter is strikingly similar to what could be found in our datasets. Siddle (2014) analysed a publicly available Transport For London dataset that contained records of bike journeys for London's bicycle hire scheme over a period of six months between 2012 and 2013, reaching the conclusion that "with a little effort, it's possible to find the actual people who have made the journeys". The study appeared in the news (Merriman, 2014), triggering TfL's remedial action. In the words of TfL's General Manager of Cycle Hire, Nick Aldworth "We're committed to improving transparency across all our services and publish a range of data for customers and stakeholders online. Due to an administrative error, anonymized user identification numbers were shown against individual trips made between 22 July 2012 and 2 February 2013. The data, which did not identify any individual customers online, was removed as soon as the matter was brought to our attention." This episode highlights that, on top of the privacy concerns that must be taken into account when designing the clearing system, leaks are possible.

Table 3 - PII accessible from social networks via APIs

	Flicker	Foursquare	YouTube	Last FM	Twitter	Facebook
Username	-	-	-	-	-	-
Name	V	V	V	V	V	V
Gender	-	V	V	V	-	V
Image	V	V	V	V	V	V
Relationship	-	-	V	-	-	-
Location	-	V	V	V	V	-
School	-	-	V	-	-	-
Company	-	-	V	-	-	-
Occupation	-	-	V	-	-	-
Hobbies	-	-	V	-	-	-
Music	-	-	V	-	-	-
Movies	-	-	V	-	-	-
Books	-	-	V		-	-
Contacts	V	V	V	V	V	-
Likes	V	V	V	V	V	-
Photos	V	-	-	-	-	-
Age	-	-	V	V	-	-
Videos	-	-	V	-	-	-
Description	-	-	V	-	V	-
Last access	-	-	V	-	-	-

Source: (Correa et al., 2012)

Table 4 - analysis of correlation attacks on transport-related databases

SOURCE	SUBJECT	CONTRIBUTION
Narayanan and Shmatikov 2008	Targeted attacks on anonymized datasets.	Actual demonstration of the general weakness illustrated in the previous section, and of even greater risks deriving from the availability of external data sources, impossible to control, which can be used to compute correlations through data mining or by filtering queries based on specific targets
Bishop et al.; 2010		
Correa et al.; 2012		
Calvino; 2015		
Golle and Partridge; 2009		
deMontjoye et al.; 2013		

Specific Attack Scenarios in Our Context and Countermeasures

The kind of correlation with external databases exemplified in the previous section is feasible in our context too. Not only it is possible to leverage online social networks in the same way, but also to browse many free-access databases of sensitive data, strongly related to the regional context. Just to give two examples, it is very easy to extract from the corresponding web sites all the professional data information (such as office address, office telephone number etc.) of the regional health organization staff, as well as of the university staff in all the cities that have an academic institution. These data are not sensitive if taken alone; in fact, the transparency laws of the public administrations mandate their availability; as shown in the next section, it is their combination with the public transport dataset that could allow privacy breaches.

If this were not enough, as explained previously the same sanitization algorithm are not free from attacks. In particular when there is the need to keep a good level of utility, algorithms as *k*-anonymity have been proven to be weak against attack where the adversary have a previous ("a-priori") knowledge. The work of Machanavajjhala et al. (2006) and the *l*-diversity algorithm have been created to overcome these de-anonymization issues of *k*-anonymity. As well explained in Ghasemzade et al (2013), anonymizing public transport data structured over a space with a high number of dimensions has been studied widely, but in general none of the proposed solutions takes into account the clearing scenario with its peculiar requisites about utility. In Ghasemzadeh et al. (2013) the differences between the different methods are clearly detailed.

Ghinita et al. (2008) propose a permutation method that groups transactions with close proximity and then associates each group to a set of mixed sensitive values. Terrovitis et al. (2008) propose an algorithm to *k*-anonymize transactions by generalization based on some given taxonomy trees. He and Naughton (2009) extend this method by introducing local generalization, which awards better quality. Xu et al. (2008) extend the *k*-anonymity model by assuming that an adversary knows at most a certain number of transaction items of a target victim, which is similar to our assumption of limited background knowledge of an adversary.

This is a very interesting model because it is definitely related to our scenario. It deals with attempts to gain a basic knowledge of some transaction rows, which is equivalent to get a certain number of possible travel records of a user: a valuable outcome for an attacker in our context. Yet, although their method addresses the high-dimensionality concern, it considers a transaction as a set of items rather than a sequence; this makes it useful to prevent attacks against the privacy of single users, but not to prevent attempts at general pattern discovery, which is typical of insider threat attacks. Therefore, it is not fully applicable to our problem, which needs to take into consideration the sequential ordering of travel data. Furthermore, Xu et al. (2008) achieve their privacy model by merely global suppression, which significantly decrease information quality on transport data.

The last reviewed model was developed by Chen et al. (2011). It studies the releasing of transport dataset while satisfying differential privacy techniques. Although they claim that their approach maintains high quality and scalability in the context of set-valued data and is applicable to the relational data, their method is limited to preserving information for supporting count queries and frequent item-sets, as opposed to Xu et al. (2008), and not passenger tracking. The combination of these two pieces of research is a very promising research direction towards a complete solution for our scenario.

https://aisel.aisnet.org/pajais/vol7/iss4/4
DOI: 10.17705/1pais.07403

Table 5 - specialized approaches to sanitization that may fit well the clearing scenario		
SOURCE	**SUBJECT**	**CONTRIBUTION**
Ghasemzadeh et al.; 2013;	LK-diversity approach	Specifically tested on transport data, this approach redesigns known techniques to overcome of their limitations, dealing especially with the preservation of useful information.
Ghinita, G. and Tao, Y. and Kalnis, P. (2008)	Anonymous publication of transactional data	Introduces tools such as flow-graphs and transactional probabilities, which are very effective to analyze the loss of useful information.

Case Studies

This section presents some simple case studies built along the lines of the illustrated threats, showing how such concepts may easily be applied to the case under consideration. Two different threats are considered:

1. an attacker who tracks the movement of a specific person (one of the authors) on the public transport network, exemplifying a threat to the privacy of individuals;
2. an attacker who is interested in understanding what are the more profitable areas in terms of regional tickets sold, to challenge the business of an operator, exemplifying a threat to trade secrets of operators .

A summary of the data items collected by operators for each ticket validation is described in Table 6. The definition of the minimal subset needed for clearing, and the anonymization of the selected fields, are the goal of the work in progress described in section 6. However, it is immediately possible to notice that the most sensitive attribute and the only direct identifier (in case of personal passes), i.e. the serial number of the ticket, cannot be omitted. Following its usage through the dataset (possibly over a period of time that cannot be known *a priori*) is the main function of the clearing system, which has to compute the share of revenue (generated when the ticket was bought) to distribute to each carrier which provided service to the ticket holder. In a broader sense, it is possible to define the required utility level of the dataset as being very similar to the goal of a potential attacker: that is, allowing to reconstruct a traveller's itinerary. It is worth detailing how this reconstruction happens, to understand also how an attacker could try the same process and follow a traveller.

Table 6 - Database table storing trip information	
Field name	Content
CONTRACT SUPPORT	Type of physical token
CONTRACT TYPE	Type of contract (single trip, pass, etc.)
VALIDATION TSP	Timestamp of ticket usage
VALIDATOR LOCATION	Placement of the validating equipment
CONTRACT RESELLER	Company which sold the ticket
VALIDATION LINE	Bus/tram/train line number
VALIDATION NR	Number of parallel validations of the same ticket (e.g. many passengers on a single pay-as-you-go ticket)
VALIDATOR SERIAL	Serial number of validating equipment
CONTRACT SN	Serial number of the ticket
VALIDATOR MODEL	Model of the validating equipment
VALIDATION ZONE	Fare zone where the ticket was validated
CONTRACT VALIDITY	Geographical extension of the contract (regional, urban, etc.)
CONTRACT ZONES	Number of fare zones the contract allows to traverse

The itinerary is "blurred" within the dataset, because in the studied system passengers validate tickets only when boarding a bus/train, but not when leaving it. Consequently, raw data does not show a sequential structure; each row represents a leg of a trip for a contract, but there is no direct connection with the next legs of the same contract. Each leg can be represented by a structure like (A)T1 → (B..Z),T2 meaning that on time-stamp T1 a user (Contract_SN) goes from A in a known direction (inferred from Validation_Line) leading to a set of possible stops, one of which is reached at T2. This introduces uncertainty in the computation of the number of traversed zones, which is needed by the clearing system when a vehicle of a different operator is used to continue the trip: in this case, the end of the first leg must be inferred from the validation that happens at the start of the second one.

To this end a *probabilistic flow-graph* can be exploited. According to Ghasemzadeh et al. (2013) a probabilistic flow-graph is a tree where each node represents a point in space-time, the edges corresponds to transitions between two places, leaving the origin at a given time to reach the destination at a different time, and each transition has an associated probability of being actually followed. For every node, there may also be a non-zero *termination probability*, which is the percentage of passengers who exit the transportation system at the node. By looking for validations of the same contract that are consecutive within a given time-frame, a possible itinerary can be identified. For example, if a validation (A)T1 could take a passenger to (B..Z), and there is a validation (D)T2, with the value of (T2-T1) falling within a given threshold, a non-zero probability can be associated to the edge (A)T1→(D)T2. The analysis can proceed seeking for destinations that can be reached from (D).

Table 7 and Figure 3 depict an example of the probabilistic flow-graph derived from one of our datasets for a few contracts. With enough samples, probabilities can be estimated with acceptable precision, and the graph becomes a faithful enough representation of the distribution of passenger over the network. At the same time, each set of itineraries for a given contract represents the habits of a passenger, enabling correlations with other data sources (places around the nodes, events close to the timestamps), and potentially leading to the association between the contract and a personal identity.

Table 7 - Table travel data in the sequential version

Serial	Sequential travel positions
70001112	(1, 245)T1->(2,249)T2->(3,248)T3->(1,245)T4
50004058	(1, 245)T3->(1, 245)T4-(1, 245)T6
50004077	(2,249)T2->(1, 245)T5
50004070	(4,260)T1->(2,249)T2->(1, 245)T5
70001386	(1, 245)T1->(2,249)T2
70001389	(1, 245)T3->(2,249)T4
75001498	(1, 245)T1->(2,249)T2->(1, 245)T4

https://aisel.aisnet.org/pajais/vol7/iss4/4
DOI: 10.17705/1pais.07403

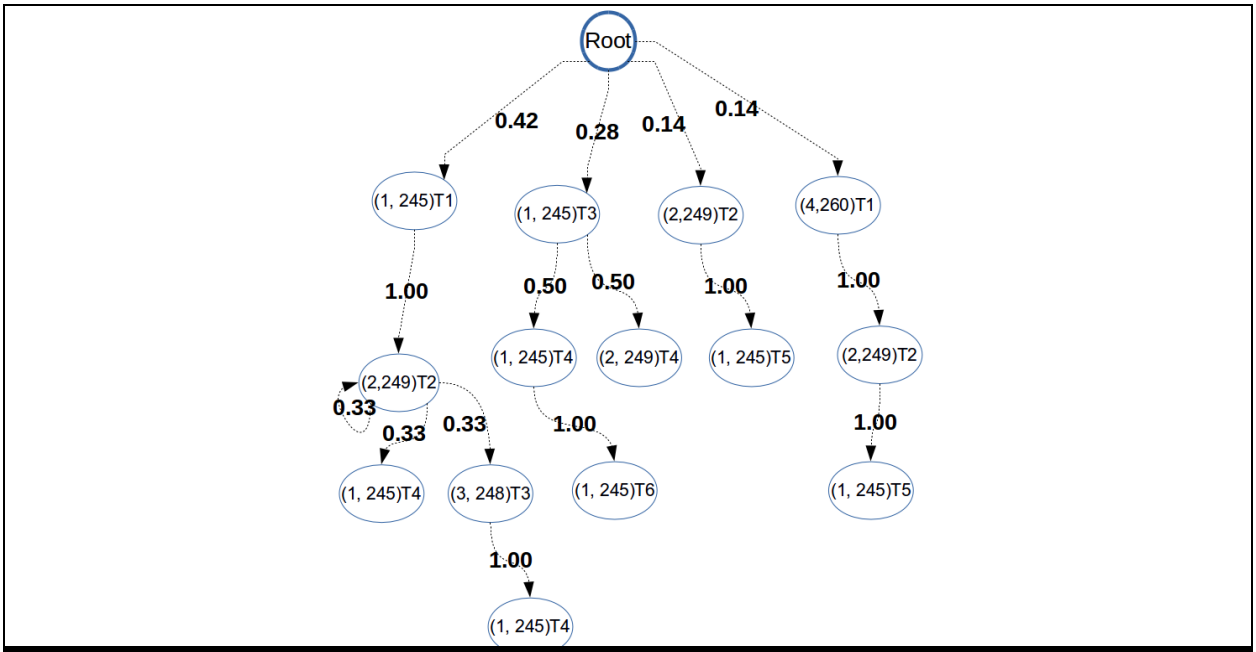

Figure 3 - Travel data represented as a probabilistic flow-graph

Stalking Franco Callegati

As a proof of concept, let us present the case of an attacker who wants to target one of the authors, to track his movement on the public transport network. Franco Callegati is a professor at the University of Bologna. He has a public web page detailing the address of his office which is located in the off-site campus of Cesena, about 60 Km East of Bologna, and showing that he works at the Department of Electrical, Electronic and Information Engineering, which has its central offices in Bologna. The phone directory lists his home address in Imola, a smaller town about 30 Km east of Bologna. Clearly from this basic data it can be inferred that Callegati will mostly travel to work from Imola to Cesena where he teaches and tutor students, but he will likely travel to Bologna too, for those sort of activities requiring physical presence related to the Department or to the University's central offices. Sometimes he will also travel from Cesena to Bologna (or the other way) when he has some commitment in both sites in the same day.

With this background, an attacker who has got a copy of the clearing dataset can associate the victim's identity with the serial number of his MiMuovo pass. Callegati is admittedly a very easy target, yet he serves us well for the purpose of giving a concrete and real example of usage of clearing datasets. Given the almost non-existent effort that allows a potential attacker to reach his goal, there is little doubt that "harder" targets can be exploited with some more, but still reasonable effort; as already illustrated at the end of section 4, even a coarse localization of commuting start and end points, when correlated with some background information about the victim, can yield significant results.

The dataset shows about 2,000 passengers boarding trains that leave Imola in the morning (all figures are computed as averages on working days). Since the validation occurs only at the start of the trip, their destination is not explicit, but of course it can be inferred with a good approximation by coupling the onward trip of a given ticket with the return trip. This further analysis yields slightly more than 1,000 passengers getting back from Bologna in the afternoon

(Fig. 4, pink arrow) and slightly less than 200 passengers getting back from Cesena (Fig. 4, grey arrow). The number of candidates drops dramatically when only passengers who travel alternatively to both Bologna and Cesena in different days of the week are considered. Only 14 MiMuovo users show a commuting pattern of this kind (Fig. 4, red arrow).

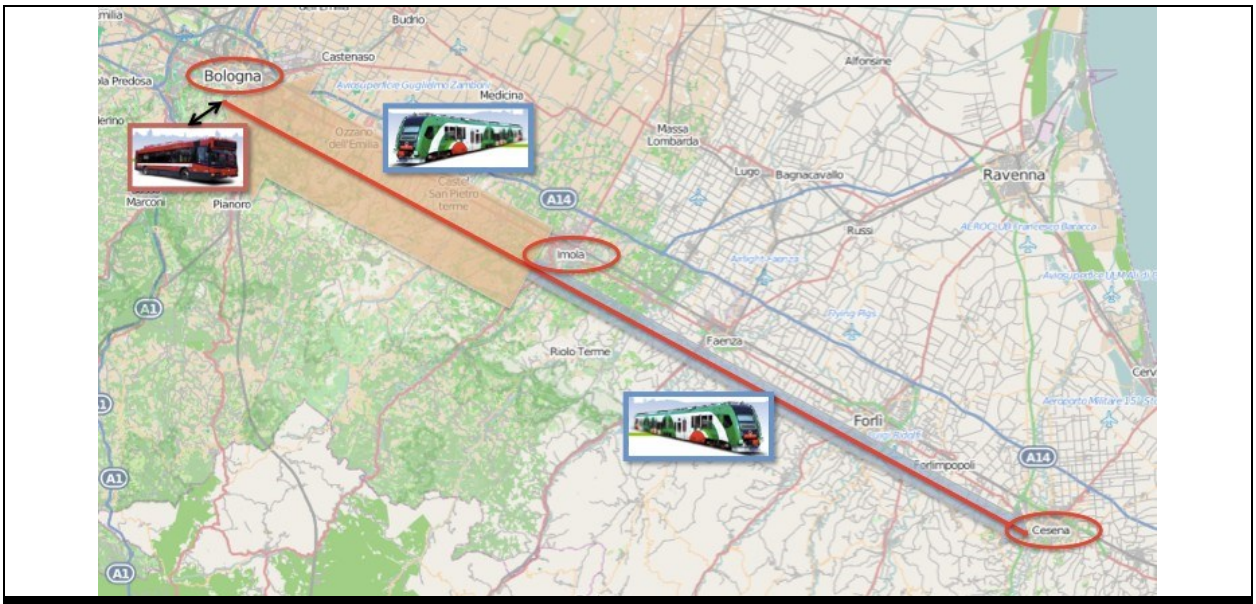

Figure 4 - Commuting flows from Imola to Bologna and Cesena

Note: Background cartography: http://www.openstreetmap.org/copyright - OpenStreetMap contribs

The possible inferences do not stop here. It is easy to check that the Callegati's office in Cesena is near enough to the train station (15 minutes on foot), and that it is not conveniently served by public transport (direct bus only once an hour)[3]. An attacker could make an educated guess that Callegati will not take a bus when he leaves Cesena's train station, and thus eliminate candidates who do it. Conversely, the site of Callegati's department in Bologna is twice as far from Bologna's train station, compared to the Cesena situation, and much better connected to it by bus (6 to 8 connections per hour). In this case an educated guess would lead an attacker to consider the exclusion of candidates who *do not* board a bus in Bologna after reaching the train station. Note also that the clearing function can be computed without taking into account the line number, but in case the full database is leaked, or in case the line number is kept on record for secondary uses, it would be possible to further restrict the set of candidates to those boarding one of the two bus lines connecting the train station with the department, out of the 19 serving the train station.

In our tests, this is enough to pinpoint the victim. This result was reached without even taking into account another very valuable source of information, the timetable of the lectures in Cesena, which would allow establishing a precise spatio-temporal constraint to the victim's movements towards one of his usual destinations. In conclusion, by following these patterns, an attacker can identify Callegati's MiMuovo card ID and then follow his movements also

3 The whole public transport network of the Emilia-Romagna Region is on Google Transit

https://aisel.aisnet.org/pajais/vol7/iss4/4
DOI: 10.17705/1pais.07403

outside his most common habits by accessing the clearing database.

Unfair Competition

Ticket validation datasets contain potentially useful information for an operator wishing to uncover the business practices of its competitors or challenge their business practice. This kind of attack comes from the inside, and it is very difficult to deal with. Access control rules cannot be very strict against insiders, who enjoy not only the possibility of easier read-only access to datasets, but also the opportunity to inject carefully crafted data to stimulate the production of particularly useful outputs, like a cryptanalyst that is able to perform a chosen plaintext attack.

One of the most valuable pieces of information would be the planning strategy in the usage of vehicles, which is a crucial issue for a transport provider and that can be inferred at some extent by exploiting the information in VALIDATION ZONE, VALIDATOR LOCATION, and VALIDATION LINE.

Here a simple and realistic inference is shown, built by looking at the correlation between the type of ticket and the zone of its usage. It is a piece of information that can give a very small margin of profit by pushing sales of multi-trip tickets where they are most appreciated, making profits on the rate of unclaimed trips for lost tickets (what is not claimed for clearing remains in the pockets of the seller). This should clearly be a small percentage of the whole ticket volume. Nonetheless in today's competitive markets every source of income may be vital; moreover the examples show that this sort of analysis may pave the road to similar analyses in "business areas" which are not considered today, because they are impossible to accurately explore in absence of large datasets.

Over 30 data mining tests over the ticket validation datasets were performed (Melis, 2014) using the Weka software (Hall et al., 2009). The correlation of interest was best highlighted by means of cluster analysis, i.e. a set of exploratory techniques that aim to group the unity of a population in statistics on the basis of their similarity in terms of values taken by the observed variables. As an example, Figure 5 shows the result of cluster analysis according to the *Simple K-means* [4] classifier. It is clear that the attributes of the sold tickets form well-separated clusters, whose significance can be useful from a business perspective. Once this hypothesis is verified, a Bayesian [5] classifier allows to infer more details over some attributes. The structure of the clearing system allows an attacker to feed the Bayesian classifier a large amount of past knowledge from the snapshots. The result is that the algorithm is able to correctly predict the belonging to a given cluster of over 90% of new instances.

This result needs to be interpreted in a specific context in order to show the power of this kind of attacks. The Bayesian test shows that theoretically, by knowing only the contract type and the validation zone, it is possible to infer the correct serial number of the contract support, which would reveal the pseudo-identity of the contract holder. Beneath the privacy risks for the contract holder, this discovery would allow a company to determine the history of a pseudo-identity. This history would be revealing the type of contract, along with its movements, leading to a kind of profiling and possible definition of targeted offers that is usually regarded as unfair competition in our context.

A more general attack is also possible, again by using the results from cluster tests. Cluster analysis usually aims to group the elements of a population on the basis of their similarity, in terms of values found in the observed variables. However, if the

[4] http://weka.sourceforge.net/doc.dev/weka/clusterers/SimpleKMeans.html

[5] http://weka.sourceforge.net/doc.dev/weka/classifiers/bayes/NaiveBayes.html

focus is put on a particular attribute, for example the contract type, it becomes possible to trace the trend in terms of other variables, for example (see again Figure 5) how the contract is used in a group of specific zones. This could easily lead an operator to discover the contract distribution of a competitor. So by intercepting this market trend, once again, the opportunity may arise to engage strategies deemed as unfair competition.

Figure 5 - Results of a clustering analysis of the dataset with Weka

Privacy by Design

Having assessed the high potential risks associated with data sharing, it is necessary to investigate what is the best approach to build security into the clearing system from the early design phase, in which the authors are involved.

Definition

Privacy by Design (PbD) is the principle by which data protection and information privacy is built into information systems from the design stage. The idea builds on existing notions of value-sensitive design, code as law, and Privacy Enhancing Technologies (PETs) (Koops and Leenes, 2014). Considerations about how to protect people's data and personal information must enter the system development life-cycle from an early stage where architectural decisions to protect privacy can still be made (Cavoukian, 2009; Schaar, 2010; Spiekermann, 2012). Such early-stage design decisions are likely to be more effective for the protection of privacy in a new information system, as there are many more options available to designers than to the engineer who needs to patch the system following a privacy incident (Brown, 2013; Schaar, 2010).

Privacy is designed into an information system when data protection and information privacy principles are incorporated into the overall design of the system (Schaar, 2010), thereby ensuring that privacy becomes integral to the organisational priorities, project objectives, design processes and planning operations (Cavoukian, 2009). A design that protects data subjects' privacy and maximises data utility requires a multi-dimensional and sophisticated consideration of the risks, and how all the parts of the design operate together (Zevenbergen et al., 2013). Privacy by Design in European Law Recital 46 of Directive 95/46 of the European Union contains the requirement that "requires that appropriate technical and organizational measures be taken, both at the time of the design of the processing system and at the time of the processing itself, particularly in

https://aisel.aisnet.org/pajais/vol7/iss4/4
DOI: 10.17705/1pais.07403

order to maintain security". Article 17 of the directive adds the requirement that "the controller must implement appropriate technical and organizational measures to protect personal data against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure or access [...]".

Privacy by design goes further than mere information security, as recognised by the data protection by design and by default requirement in the proposed General Data Protection Regulation 2012/0011/COD. Article 23 of the proposed Regulation contains the requirement that controllers implement measures to ensure "only those personal data are processed which are necessary for each specific purpose of the processing and are especially not collected or retained beyond the minimum necessary for those purposes, both in terms of the amount of the data and the time of their storage." Although this requirement is a step in the direction of PbD, the proposed methodologies in literature go further than the European legislator has proposed.

Methodology

A synthesis from the literature shows the following factors are essential to an effective Privacy by Design strategy. This list does not go into detail about specific de-identification techniques, software patterns or other privacy enhancing technologies. Rather, it is a list of more abstract factors that contribute to a successful Privacy by Design approach:

Define privacy risk assessment, goals & strategy - An information system design should start with an assessment of the risks of the data that will be collected (see for example Rotter (2008). The project instigator should define clearly the goals he or she wants to achieve in terms of privacy protection, while consideration of how to reach these will follow in further steps (Spiekermann, 2012).

Holistic approach - The risk assessment and goals will inform the privacy design strategy, which is comprised of an iterative approach of the steps below (Hoepman, 2012). It is vital the information system is considered as a whole in a holistic approach. An effective privacy by design strategy will have privacy settings set by default, since software settings are unlikely to be changed by users (Cavoukian, 2009; Gross and Acquisti, 2005; Mackay, 1991).

Data minimisation & Purpose limitation - Information systems should be designed in a way that they use the minimal necessary personal data, without necessarily compromising on the functionality of the system. Therefore, the purpose of the system must be clearly defined and the combinations of collectable personal data analysed that are necessary for the full functionality (Brown, 2013; Hourcade et al., 2014; Schaar, 2010) . This makes data minimisation a necessary and foundational element of Privacy by Design (Gurses et al., 2007).

De-identification and Aggregation - When the necessary personal data have been identified, the project instigator should analyse to what extent the information system can be operated without the identifier fields of the data subjects in her database. The database should be de-identified or aggregated to the furthest extent possible. Although useful to increase the privacy of data subjects, de-identification and aggregation are never fully robust against re-identification practices and should not be considered as a means to circumvent the application of data protection law to the project (Danezis and Troncoso, 2013; de Montjoye et al., 2013; Ohm, 2009; Sharad and Danezis, 2013).

Secondary uses & Dissemination type - Before deciding how best to de-identify any collected data, the researcher must decide how the research data will be disseminated and which further uses the data will be suitable for. For example, it may be required that the data be shared in an open data format, whereby the risk of privacy harms will be significant, thus requiring the

adoption of a strong de-identification method. The project instigator can choose to share the data only upon request, and decide the method of dissemination on a case by case basis, along with suitable legal agreements, which should also be enforced if breached (Zevenbergen et al., 2013).

Transparency/openness - The OECD established that data controllers must be transparent about the information processes where personal data is processed (OECD, 1980). This is achieved by informing the potential data subject about the information processing before gathering informed consent in a lawful manner, while also complying with the data subject's rights such as maintaining the accuracy of the data and allowing access to correct the data (Brown, 2013; Schaar, 2010; Spiekermann, 2012).

Accountability - Privacy by Design is not merely a technical process, but must be complemented by information security, functionality, operational efficiency, organizational control and business processes that enable a trustworthy information environment (Koops and Leenes, 2014). A careful Security by Design plan must also be set up, whereby topics such as encryption and access limitations are ensured (Brown, 2013; Spiekermann, 2012). A plan must be established for when unforeseen risks materialise, and legal agreements on information sharing must be enforced effectively.

Application

In order to take the described factors into account when designing the clearing system, we propose to undertake an iterative approach described by the following steps:

1. **Data minimization** - take into account the needs of the legitimate analysers to define the smallest set of attributes that allow performing the intended computations;
2. **Sanitization techniques** - choose the perturbation and generalization algorithms that provide the most effective concealment of sensitive data without jeopardizing its utility for legitimate uses;
3. **Verification** - evaluate the results to formally verify that the privacy policies are respected and that the resulting dataset actually preserves its intended utility. If some basic constraint is found violated, or margins for further improvement are visible, the cycle is reiterated to achieve the foreseen refinement, otherwise the process stops. Note that this step could highlight an intrinsic contradiction between some of the privacy policies and some of the analysis requirements, leading either to the decision to relax some requirement or to the conclusion that the desired scenario is unachievable.

The last step of the iterative process calls for a formal metric to evaluate correctness and effectiveness. The literature already provides useful methodologies for this purpose.

Ganta and Acharya (2008) studied the general problem of *fusion resilient anonymization*. They stated the problem as a search for the optimal value of an objective which is the weighted sum of protection and utility. Protection is measured in terms of how hard is for an adversary to gain information by correlating the anonymized dataset with external sources. More formally, if P is the original sensitive dataset, and P^{I} is the sanitized release of P, an adversary can exploit information fusion techniques to derive a de-sanitized version $\hat{P}$ from P^{I}. The effectiveness of the applied sanitization process is measured as the *dissimilarity* between $\hat{P}$ and P.

Brickell and Shmatikov (2008) performed a similar analysis, again studying the trade-off between utility and protection, but on a more formal level. Their claim was quite

https://aisel.aisnet.org/pajais/vol7/iss4/4
DOI: 10.17705/1pais.07403

thought-provoking: sophisticated anonymization techniques offer no real advantages over trivial ones, i.e. datasets sanitized with complex application of generalization and perturbation algorithms, depending on the algorithm parameters, either provide no additional utility vs. trivially-sanitized datasets, or leak much more information to adversaries than what is gained in terms of legitimate analysis. Besides posing interesting questions that researchers in this field could find useful to orientate their efforts, their paper also provides formal definitions for various metrics related to privacy of data tables and utility of sanitized databases. In particular, they study privacy both under a syntactic perspective (pure statistical correlation) and under a semantic perspective, measuring the gain in adversarial knowledge afforded by the sanitized table.

Bishop et al. (2010) explore the topic by focusing on relationships that can be used to desanitize sensitive data. They model the problem of data sanitization as a double set of assertions, made of the constraints defining the privacy properties that must hold against adversarial attacks, and of the targets defining the information that the legitimate analysts want to extract from the sanitized dataset. They highlight the importance of defining a precise threat model as a requisite for drawing complete and concrete privacy policies, and a precise analysis policy that, in opposition to the privacy policy, puts limits to the sanitization process to avoid excessive loss of utility. They exploit *ontologies* to automate reasoning over these opposing requisites and solve the constraint satisfaction problem they represent. A question they leave open is: what is the most appropriate language to express these requirements?

To define our specific privacy by design process, we plan to investigate and possibly apply the common concepts and techniques presented in all of these works, in addition to a specific and noteworthy suggestion that comes from the last one. In the words of the authors: “Perhaps the most constructive approach is to provide two sets of relationships. The first lists those relationships that are known to hold in the raw data, and must not hold if desanitization is to be prevented. The second is a set of relationships that, *if they held*, would enable desanitization. The sanitizer can deal with the first set as appropriate. The second enables the sanitizer to perform a simple risk analysis, centred on two questions: (1) What is the probability that the relationships in this set hold; (2) What is the probability that the adversary will be able to determine that the relationships hold, and use that to desanitize the data?”. This approach seems to be especially useful because it allows both designing sanitization by evaluating the effectiveness of the planned techniques, and to measure and understand the risks deriving from future evolutions of the intended use of the datasets.

Conclusions

This manuscript highlights the privacy threats that can emerge from sharing or publishing the data related to usage of public mobility tickets. In the context of an integrated mobility system run by a set of operators, sharing data about tickets usage become mandatory for revenue clearing purposes. Unfortunately this may also pave the road to privacy attacks to individuals or institutions, such as, but not limited to those exemplified in this paper.

An ample discussion is presented of how the sanitization approaches could work in this scenario, and what their limitations are; it lays the ground for future research aimed both at improving the effectiveness of sanitization techniques in the specific scenario, and to derive generally-applicable principles.

As an alternative solution, the applicability of the “privacy by design” approach is examined. In the context of the design of the data sharing system for clearing purposes, its aim is to minimize the

likelihood of the emergence of privacy threats. The general-purpose definition of the approach is integrated with an implementation plan that takes into account a variety of literature sources to verify the effectiveness of the applied methodology, in order to iteratively converge towards the solution that strikes the most appropriate balance between data utility and privacy, possibly quantifying the effects of a breach.

The first set of research questions, regarding the suitability of existing techniques to protect privacy in the public transport scenario, has thus being answered in a substantially negative way. The second wave of research questions, on the possible definition of a more effective framework to devise privacy-enhanced data management processes, has been partially addressed. The present study meets its main limitations here: the negative findings from the review phase, and the realization that similar initiatives actually deal with less demanding requisites, were useful in highlighting the deficiencies of current approaches and lead to devise suggestions on how the framework could be structured, yet its concrete development will be the subject of future work.

References

Barbaro, M. and Zeller, T. (2006).”A face is exposed for aol searcher no.4417749,” New York Times http://www.nytimes.com/2006/08/09/technology/09aol.html.

Bishop, M., Cummins, J., Peisert, S., Singh, A., Bhumiratana, B., Agarwal, D., Frincke, D. and Hogarth, M.(2010). “Relationships and data sanitization: A study in scarlet,” In *Proceedings of the 2010 Workshop on New Security Paradigms*, NSPW ’10, pp.151–164, New York, NY, USA.

Brickell, J and Shmatikov, V. (2008) “The cost of privacy: Destruction of data-mining utility in anonymized data publishing”. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’08, pp. 70–78, New York, NY, USA.

Brown. I. (2013). “Britain’s smart meter programme: A case study in privacy by design.” *Social Science Research Network Working Paper Series*.

Calvino, C. (2015). “Stalking pincopallino: sorveglianza, privacy e prossimità al tempo di Twitter.” *Rivista Geografica Italiana*, 122(1), pp.67-94.

Cavoukian, A. (2009). “Privacy by design the 7 foundational principles implementation and mapping of fair information practices.”

Ciriani, V., De Capitani di Vimercati SSF., Samarati, P., Yu, T. and Jajodia, S. (2007). *Secure Data Management in Decentralized Systems*. Springer-Verlag.

Chen, R., Mohammed, N., Fung, B.C.M., Desai, B. C. and Xiong, L. (2011). *“*Publishing set-valued data via differential privacy” *Proc. VLDB Endowm., pp.1087–1098.*

Cormode, G., Procopiuc, C. M., Shen, E., Srivastava, D. and Yu T. (2013). "Empirical privacy and empirical utility of anonymized data", *IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, pp. 77-82.

Correa, D., Sureka, A. and Sethi, R. (2012). “Whacky.! - what anyone could know about you from twitter.” In *Privacy, Security and Trust (PST), Tenth Annual International Conference on*, pp. 43–50.

Crawford, R. (2007) “Sanitization models and their limitations” In *Proceedings of the 2006 Workshop on New Security Paradigms*, NSPW ’06, pp.41–56, New York, NY, USA.

DOI: 10.17705/1pais.07403

Cynthia, D. and Roth, A. (2014), "The Algorithmic Foundations of Differential Privacy", *Foundations and Trends® in Theoretical Computer Science*, *9(3–4), pp.211-407*

Danezis, G. and Troncoso, C. (2013) "You cannot hide for long: De-anonymization of real-world dynamic behaviour." In *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society*, WPES '13, pp. 49–60, New York, USA.

de Montjoye, Y. A., Hidalgo, C. A., Verleysen, M. and Blondel, V. D. (2013). "Unique in the crowd: The privacy bounds of human mobility." *Nature SR*.

Ganta, S. R. and Acharya, R. (2008). "On breaching enterprise data privacy through adversarial information fusion." In *Proc. of the 2008 IEEE 24th International Conference on Data Engineering Workshop*, ICDEW '08, pp. 246–249, Washington, DC, USA. IEEE Computer Society.

Ganta, S. R., Kasiviswanathan, S.P. and Smith, A. (2008) "Composition attacks and auxiliary information in data privacy." In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '08). ACM, New York, NY, USA, pp.265-273.

Ghasemzadeh, M., Fung, B.C.M., Chen, R. and Awasthi, A. (2014). *"*Anonymizing trajectory data for passenger flow analysis," *Transportation Research Part C: Emerging Technologies, 39, pp. 63-79, ISSN 0968-090X.*

Ghinita, G., Tao, Y. and Kalnis, P. (2008) "On the anonymization of sparse high-dimensional data." In: *Proceedings of the 24th IEEE International Conference on Data Engineering (ICDE),* pp. 715–724.

Golle, P. and Partridge, K. (2009) "On the anonymity of home/work location pairs." *In Proceedings of the 7th International Conference on Pervasive Computing*, Pervasive '09, pp. 390–397, Berlin, Heidelberg. Springer-Verlag.

Gross, R. and Acquisti, A. (2005) "Information revelation and privacy in online social networks." *In Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society*, WPES '05, pp. 71–80, New York, NY, USA.

Gurses S., Troncoso, C. and Diaz, C. (2007) Engineering privacy by design.

Hall, M. (November 2009) "The weka data mining software: An update." *SIGKDD Explor. Newsl.*, 11(1),pp.10–18.

He, Y. and Naughton, J. F. (2009). "Anonymization of set-valued data via top-down, local generalization" *Proc. VLDB Endowm., 2(1),pp. 934–945*

Hoepman, J. H. (2012) Privacy design strategies. *Computing Research Repository.*

Hourcade, J. P. (2014) "Electronic privacy and surveillance." In *Proc.of the Extended Abstracts of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI EA '14, pp. 1075–1080, New York, NY, USA.

Koops, B. J. and Leenes, R. (2014) "Privacy regulation cannot be hardcoded. a critical comment on the 'privacy by design' provision in data- protection law." *Int. Rev. Law Comput. Technol.*, 28(2), pp.159–171.

Li, N., Li, T. and Venkatasubramanian, S. (2007). "t-closeness: Privacy beyond k-anonymity and l-diversity." In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pp. 106–115.

Machanavajjhala, A., Gehrke, J. and Kifer, D. (2006). "*l*-diversity: Privacy beyond *κ*-anonymity." In *Proceedings of the*

22Nd International Conference on Data Engineering, IEEE Computer Society.

Mackay, W. E. (2001) "Triggers and barriers to customizing software." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '91, pp. 153–160, New York, NY, USA. ACM.

Melis, A. (2014) "Data sanitization in a clearing system for public transport operators," master thesis in computer science, defended Jul 17, 2014 at the University of Bologna.

Merriman, C. (2014) Boris bikes location data could be used to track you.

Narayanan, A. and Shmatikov, V. (2008) "Robust de-anonymization of large sparse datasets." In *Proc. of the 2008 IEEE Symposium on Security and Privacy*, SP '08, pp.111–125, Washington, DC, USA. IEEE Computer Society.

Narayanan, A. (2009). Your morning commute is unique: On the anonymity of home/work location pairs.

Li, N., Li, T. and Venkatasubramanian, S. (2007) "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on* , pp.106-115.

Nergiz, M. E., Atzori, M. and Clifton, C. (2007) "Hiding the presence of individuals from shared databases." In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data* (SIGMOD '07). ACM, New York, NY, USA, pp.665-676.

OECD. (1980) Oecd guidelines on the protection of privacy and transborder flows of personal data.

Ohm, P. (2009) "Broken promises of privacy: Responding to the surprising failure of anonymization." *UCLA Law Review*, 57(9-12), pp.1701 – 2010.

Rotter, P. (2008) "A framework for assessing rfid system security and privacy risks." *IEEE Pervasive Computing*, 7(2), pp.70–77.

Schaar, P. (2010) "Privacy by design." *Identity in the Information Society*, 3(2), pp.267–274.

Sharad, K. and Danezis, G. (2013) "De-anonymizing d4d datasets." In *Proc. of the 13th International Symposium Privacy Enhancing Technologies*, Lecture Notes in Computer Science, Berlin, Heidelberg. Springer- Verlag.

Siddle, J. (2014) I know where you were last summer: London's public bike data is telling everyone where you've been.

Spiekermann, S. (2012) "The challenges of privacy by design." *Commun. ACM*, 55 (7), pp.38–40.

Sweeney, L. (2002) "K-anonymity: A model for protecting privacy." *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5), pp.557–570.

Terrovitis, M., Mamoulis, N. and Kalnis, P. (2008). "Privacy-preserving anonymization of set-valued data" *Proc. VLDB Endowm., 1(1), pp.115–125.*

Pingshui, W. and Jiandong, W. (2013). "L-Diversity Algorithm for Incremental Data Release" *Appl. Math. 7(5), pp.2055-2060.*

Xu, Y., Fung, B.C.M., Wang, K., Fu, A.W.-C. and Pei, J. (2008) "Publishing sensitive transactions for itemset utility." In: *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM), pp. 1109–1114.*

Zevenbergen, B., Brown, I., Wright, J. and Erdos, D. (2013). "Ethical privacy guidelines for mobile connectivity measurements." *Social Science Research Network*.

DOI: 10.17705/1pais.07403

About the Authors

Franco Callegati received his Master and Ph.D. in Electrical Engineering in 1989 and 1992 from the University of Bologna, Italy. He joined the University of Bologna as assistant professor in 1995 and since 2001 he has been serving as associate professor. His research interests are in the field of teletraffic modelling and performance evaluation of telecommunication networks, and more recently he focused on automation issues of the network control plane. He is senior member of the IEEE. He is currently co-chairing the Education committee of the IEEE SDN initiative.

Aldo Campi received the degree in electronic engineering from the University of Bologna, Italy, in 2004 and obtained his PhD Telecommunication Engineering in the same university in 2009. His current research interests include Software Defined Networking and signalling protocols for Cloud computing over dynamic optical networks, with special emphasis on the deployment of the signalling protocol to bring application-aware functions to network nodes. Since 2009 Aldo Campi has been involved in teaching activities in the field of Communication Networks at the EI Department of the University of Bologna, School of Engineering, Cesena Campus, Italy.

Andrea Melis is a researcher fellow at the University of Bologna. He received his Master in Computer Science at University of Bologna with a thesis about possible privacy exploitation attacks on a public transport system; in 2013 he also obtained another Master at Polytech of Sophia-Antipolis. Actually his job focuses on the security and privacy aspects of the Emilia-Romagna Regional Government project called "Mi Muovo" (I move). More recently, his interests started including other topics related to the cyber security aspects of public services such as malware analysis, penetration testing, vulnerability discovering and privacy risk assessment.

Marco Prandini is a research associate at the University of Bologna, Italy, where he got his PhD in electronic and computer engineering in 2000. His research activities started in the field of public-key infrastructures and later moved to subjects related to computing systems security, high availability, and system administration. More recently, his interests started including the application of security to social fields where computing devices are becoming pervasive, like social networks and e-government.

Bendert Zevenbergen joined the Oxford Internet Institute to pursue a PhD on the intersection of privacy law, technology, social science, and the Internet. He runs a side project that aims to establish ethics guidelines for Internet research, as well as working in multidisciplinary teams such as the EU funded Network of Excellence in Internet Science. He has worked on legal, political and policy aspects of the information society for several years. Most recently he was a policy advisor to an MEP in the European Parliament, working on Europe's Digital Agenda. Previously Bendert worked as an ICT/IP lawyer and policy consultant in the Netherlands. Bendert holds a degree in law, specialising in Information Law.