Please cite this paper as:
Francesca Pratesi, Lorenzo Gabrielli, Paolo Cintia, Anna Monreale, Fosca Giannotti,
PRIMULE: Privacy risk mitigation for user profiles,
Data & Knowledge Engineering (2020) 101786, vol. 125,
ISSN: 0169-023X, https://doi.org/10.1016/j.datak.2019.101786

PRIMULE: Privacy Risk Mitigation for User Profiles

Francesca Pratesi[a,b], Lorenzo Gabrielli[a], Paolo Cintia[a,b], Anna Monreale[a,b], Fosca Giannotti[a]

[a] *ISTI - CNR*
[b] *University of Pisa*

Abstract

The availability of mobile phone data has encouraged the development of different data-driven tools, supporting social science studies and providing new data sources to the standard official statistics. However, this particular kind of data are subject to privacy concerns because they can enable the inference of personal and private information. In this paper, we address the privacy issues related to the sharing of user profiles, derived from mobile phone data, by proposing PRIMULE, a privacy risk mitigation strategy. Such a method relies on PRUDEnce [1], a privacy risk assessment framework that provides a methodology for systematically identifying risky-users in a set of data. An extensive experimentation on real-world data shows the effectiveness of PRIMULE strategy in terms of both quality of mobile user profiles and utility of these profiles for analytical services such as the *Sociometer* [2], a data mining tool for city users classification.

Keywords: Mobile Phone Data, Call Detail Record, Privacy, Anonymization

1. Introduction

Nowadays, mobile devices record digital traces of different human activities such as movements, purchase transactions, preferences, opinions, and so on. Thus, they are an important source of information that enables the study of environmental monitoring, transportation, social networks, innovative demographic indexes and human behavior. In particular, the availability of CDR (Call Detail Record) data produced by mobile phones stimulated the research for sophisticated data mining algorithms suitable for understanding people habits and mobility patterns [3]. This type of data has been used also for monitoring population movements and displacement after disasters, such as earthquakes [4], for helping decision making in public health, particularly when considering the dynamics and spread of infectious diseases and the consequences of a natural disaster [5].

The opportunity of exploiting big data has attracted also the interest of official statistics [6]. Indeed, currently, a hot topic in official statistics is the exploitation of big data in combination with traditional data sources, in order

to improve quality, timeliness and spatio-temporal granularity of statistical information. As an example, in [2], Furletti et al. presented the *Sociometer*, a data mining tool for classifying users by means of their calling habits, uses the calling activities to infer a presence indicator of different categories of people in a city. It takes advantage of a methodology able to construct an aggregate and compact user call profile.

The use of human data for both understanding social phenomena and the development of data-driven services is getting common, but at the same time, raises the concern on leakage of personal information or re-identification. In fact, numerous services have been temporarily put to halt or even out of service because of such issues[1,2]. In practice, nowadays, the knowledge discovery related to human behavior comes with unprecedented opportunities and risks. The paradoxical situation we are facing is that we are running the risks, without fully catching the opportunities of big data. Indeed, on the one hand, we feel that our private space is vanishing in the digital world, and our personal data can be used without feedback and control; on the other hand, the same data are seized in the databases of companies (Telcom companies, insurance companies, and so on), which use legal constraints on privacy as a reason for not sharing it with science and society at large, keeping this precious source of knowledge locked to data analysts or service developers.

In Europe, policy-makers have addressed this problem with the General Data Protection Regulation (GDPR) [7]. This regulation responds to privacy and data protection threats associated with new data practices by strengthening protections for individuals, and also by harmonizing the legal framework to enable data to flow better within Europe. The GDPR introduces the practice of a Data Protection Impact Assessment and the application of the Privacy-by-Design principle in the creation of information systems. Thus, it is necessary to keep under control the privacy risk of users in the data and to enable knowledge discovery from raw data while preventing privacy violations by-design.

In this paper, we address the problem of guaranteeing privacy protection while using individual profiles for the extraction of additional knowledge, hidden in the data, by sophisticated data mining processes. In particular, our main goal is to guarantee privacy protection during the application of the *Sociometer* [2], that is considered a valuable tool for official statistics [8]. To this end, we propose PRIMULE (Privacy RIsk Mitigation for User profiLEs), a privacy risk mitigation strategy for making private a set of user profiles. PRIMULE relies on PRUDEnce [1], a privacy risk assessment framework that provides a methodology for systematically identifying risky-users in a set of data. PRIMULE, on the basis of the privacy risk assessment of user profiles, acts making similar profiles indistinguishable to eliminate possible risky profiles.

We conduct a detailed analysis of our approach using a real data set. In particular, we used a CDR dataset that covers 139 municipalities of Tuscany with

[1] Yomiuri - https://goo.gl/Pxiuny
[2] Tom Tom - https://goo.gl/J8tcuc

85 million CDRs from about 3 million customers in the month of November 2016 (4 weeks). The deep experimentation shows the effectiveness of PRIMULE. Indeed, after the privacy risk mitigation, the quality of the profiles is high in terms of similarity with respect to the original ones. This fact is also confirmed by the utility of the private profiles for the *Sociometer*, which is measured both in terms of classification and quantification performance. Empirical results demonstrate a good classification and quantification especially for the city user category of residents. In each experiment, we perform a comparison of PRIMULE against a method based on differential privacy [9]. Again, experiments show that our proposal provides much better results in terms of data quality and service utility.

The rest of this paper is organized as follow. Firstly, in Section 2 we report some relevant literature about privacy in mobile phone data. Then, in Section 3, we describe the basis of our work: *i)* the individual user profile describing the calling activity, *ii)* the *Sociometer* framework, and *iii)* the PRUDEnce framework used for the assessment and the mitigation of the privacy risk. In Section 4, we introduce the problem definition while in Sections 5 & 6 we present the privacy attack model and our mitigation strategy PRIMULE. In Section 7, we show the results of our experiments on real data, bringing in evidence the effectiveness of our approach on both individual privacy and accuracy of the results. Finally, Section 8 concludes the paper.

2. Related work

Relatively little work has addressed privacy issues in the publication and analysis of GSM data. In the literature, many works treating mobile phone data state that there is no privacy issue or at least the privacy problems are mitigated by the high spatial coverage of the cell phone. However, Golle and Partridge [10] showed that a fraction of the US working population can be uniquely identified by their home and work locations even when those locations are not known at a fine scale or granularity. For this reasons, the risk in releasing locations traces of mobile phone users appears very high.

Privacy risks, even in the case of releasing of location information with not fine granularity, are studied in [11], where authors present a study on 30 billion CDRs from a nationwide cellular service provider in the United States. They observed several location information for about 25 million mobile phone users in a period of three months. This study highlights important factors that can have a relevant impact on the anonymity. Examples are the value of N in finding the top N locations, the granularity level of the released locations, the availability of additional social information about users, and geographical regions.

When the spatial granularity level of the cell data is combined with time information and a unique handset identifier, all this information can be used to track people movements. This requires that a good privacy-preserving technique has to be applied when we analyze such data. Unfortunately, many current proposals, such as those presented in [12, 13], do not consider this aspect. However, the work in [13] is very interesting because studies user re-identification risks in

GSM networks in the case user historical data is available to characterize the mobile users a priori.

In [14], a study on re-identification of CDRs is presented, applying spatial and temporal generalizations. In particular, this paper is based on the concepts of unicity [15] and other related ones, which represents the percentage of users in a dataset who are re-identified by using p randomly selected data points from each user's records. This is a concept very related to our privacy risk; the main difference is that we do not pick random data points but we systematically select temporal slots as background knowledge. Noriega et al. [14] used similar levels of aggregation w.r.t. ours (indeed, they also use the municipality level for space and temporal slots of 6, 12 and 24 hours), but they assessed utility based on a qualitative survey targeted to experts, while we tested the usefulness of our strategy comparing the performance obtained by a service using original and private data.

In [16], de Montjoye et al. present a summary of the main models to treat the privacy problem in mobile phone data (although they are quite general to be applied in a variety of different contexts), spacing from limiting the release of data (e.g., under legal contracts or using privacy-through-security approaches), transforming the data by adding technical difficulties to attempts at re-identifying individuals, using synthetic data, or relying on a question-and-answer model, where answers can be both at the level of individuals or, more often, groups of individuals.

Also Argaoui et al. [17] propose a review of existing anonymization techniques, suggesting, for CDR, to apply reversible anonymization (essentially, cryptographic techniques) for all the data contained in the CDR, except for the URL visited by the customers, because the telecommunication operator have not in any case the legal right to consult this latter information.

In [18], authors presents a statistical disclosure control methods applied to CDRs. They focus on data obtained in a stream fashion, so the primary goal of the implemented method must be the efficiency.

In [19], we can find a useful survey on mobile phone data analyses. It starts with the study of the social networks generated by the mobile call graphs, then it provides some examples of services that it is possible to construct adding geographical information, i.e., cell towers, and temporal information. Finally, it reports some studies about privacy. As typical solutions, Blondel et al. [19] suggest to operate small modification of datasets, or to change frequently pseudo-identifiers: every day (as in [11]) or even every 6 hours (like in [20]). Unfortunately, this can lead strong limitation on analyses and services that can be performed. Otherwise, Blondel et al. [19] report results of [21], where it is suggested to use synthetic data, which can reproduce many features of mobility of users of Los Angeles and New York, to model the movements of people.

In order to enable a privacy-respectful management of data, Cavoukian conceived the Privacy-by-Design paradigm [22]. This model represents a profound innovation w.r.t. the traditional methods: the idea is to have a significant shift from a reactive model to proactive one, i.e., preventing privacy issues instead of remedying to them. In [23], Monreale et al. investigated the application of

Privacy-by-Design in data mining domains, providing evidences that this principle can ensure a quality leap in the conflict between data protection and data utility. In this paper, we can also find an application of the Privacy-by-Design model on mobile phone data; while this is the starting point of our work, in [23] no general framework is implemented and, more important, no mitigation strategy is suggested to limit the privacy risk.

Another model of anonymization is the Differential Privacy, a privacy notion introduced in [9] by Dwork. The key idea is that the privacy risks should not increase for a respondent as a result of occurring in a dataset. Differential privacy ensures, in fact, that the ability of an adversary to inflict harm should be essentially the same, independently of whether any individual opts in to, or opts out of, the dataset. This privacy model is called ϵ-differential privacy, due to the level of privacy guaranteed ϵ. It assures a record owner that any privacy breach will not be a result of participating in the database since nothing, or almost nothing, that can be discovered from the database with his/her record that could not have been discovered from the one without his/her data [24]. Moreover, in [9], it is formally proved that ϵ-differential privacy can provide a guarantee against adversaries with arbitrary background knowledge. This strong guarantee is achieved by comparison with and without the record owner's data in the published data. It is important to note that the parameter ϵ, which specifies the level of privacy guaranteed, is public [25].
Here, we do not report the formal definitions of this privacy model, but we only describe one of the fundamental concepts of this technique, which is the global sensitivity [9]. The global sensitivity of a query is a function that maps underlying datasets to (vectors of) reals. Intuitively, the global sensitivity represents how much the result of a query can change when it is performed on the dataset or on a dataset close to it.
There are two popular mechanisms to achieve differential privacy: *Laplace mechanism*, which supports queries whose outputs are numerical [26], and *exponential mechanism*, which works for any queries whose output spaces are discrete [27]. The basic idea of the Laplace mechanism is to add noise to aggregate queries (e.g., counts) or queries that can be reduced to simple aggregates. This mechanism is suitable for our aim, since we have (aggregation of) numerical values.

Regarding the Differential Privacy [9, 25] in CDR data, in [28] authors apply Geometrical mechanism to a predetermined partition of a territory, using also Voronoi tessellation to keep track of the presence of individuals. In this paper, clustering and sampling with Fourier-based perturbation are used. Another work is [29], which presents DP-WHERE, an extension of [21] that includes differential privacy. However, Mir et al. [29] apply differential privacy only to aggregate information, such as the probability distribution of the homes (or workplaces) over the grid cells, the numbers of calls per day made by the users or the hourly distribution of the calls.

In [30], Acts et al. point out that risks related to privacy can be present also in aggregated information, since attackers can reconstruct even entire individual trajectories from aggregate location data, if aggregates are periodically and sufficiently frequently published (e.g., in every half an hour). Indeed, uniqueness

(which is one of the fundamental properties of location trajectories, as reported in [15], along with predictability, and regularity) is considered to have devastating effect on the utility of anonymized datasets, due to the fact that location data is typically high-dimensional and sparse. In this paper, authors present a method based on perturbation of location data, where a maximum number of observations is fixed (i.e., only a certain number of data per user, randomly selected, is maintained), in order to obtain a bound on the sensitivity level.

While Differential Privacy model offers strong guarantees regardless of any background knowledge an adversary can have, we could apply *Privacy-by-Design* paradigm tailoring the anonymization strategy on a specific (but not necessarily weak) background knowledge and on the service we want to deploy with mobile phone data. A privacy model that works well with *Privacy-by-Design* is k-anonymity [31, 32], which aims to ensure that each individual in a dataset cannot be distinguished from at least $k-1$ individuals whose information are also in the dataset.

To conclude our overview on existing methods and techniques, the idea to exploit clustering techniques to achieve anonymity is not new, but it is already described in the literature, as in [33, 34, 35].
In [33], Le Freve et al. use a greedy partitioning algorithm, where the defined regions cover the domain space. Authors also introduce the relaxed partitioning, that allows a potential overlap in the generalization, i.e., the same Quasi-Identifier value can be generalized in different ways if it belongs to different records. A Quasi-Identifier (QI) is a piece of information that is not a unique identifiers, but is sufficiently well correlated with an entity, and it can be potentially combined with other quasi-identifiers to create a unique identifier. We rely on the possibility to generalize the same QI in different values, but we simplify the algorithm: in our work, we do not need to choose how to partition the space, since for us the QIs are always treated together, like a unique block.
In [34], Byun et al. provide a greedy algorithm that populates a cluster at the time, assigning the closest record among the free ones, and they change cluster when it reaches the desired dimension. Our algorithm, instead, assigns records to clusters only basing on global order, ensuring that each assignment is a global optimum and not a local one.
In [35], Lin and Wei create the various clusters at the same time, and then they adjust the clusters removing and reassigning records to clusters. On the contrary, our method is an iterative and a greedy one, so the solution is built incrementally, but each decision is definitive.
Moreover, the quality evaluation of the previous works is based on information loss or similar metrics. To the best of our knowledge, we are the first that apply a clustering algorithm to reach anonymity and evaluate their solution with a real service.

3. Background

GSM (Global System for Mobile Communications) Network is a mobile network that enables the communications between mobile devices. The GSM pro-

tocol is based on a cellular network architecture, where a geographical area is covered by a number of antennas emitting a signal to be received by mobile devices. Each antenna covers an area called *cell*. In this way, the covered area is partitioned into a number of, possibly overlapping, cells, uniquely identified by the antenna. Cell horizontal radius varies depending on antenna height, antenna gain, population density and propagation conditions from a couple of hundred meters to several tens of kilometers. A Call Detail Record (CDR) is a log data documenting each phone communication that the telecom operator stores for billing purposes. The format of the CDR used in this work is the following: $\langle Timestamp, Caller\_id, d, Cell\_1, Cell\_2 \rangle$, where *Caller_id* is the pseudo-identifier of the user that called, *Timestamp* is the starting time of the call, d is its duration, *Cell_1* and *Cell_2* are the identifiers of the cells where the call respectively started and ended.

3.1. Individual Call Profiles

The concept of Individual Call Profile (ICP) is introduced in [2]. The ICP is an aggregated spatio-temporal profiles of an user, computed by applying spatial and temporal breaks on user's CDRs (Figure 1). In our case, the spatial aggregation is at the municipality level, while the temporal aggregation is by week, where each day of a given week is grouped on weekdays and weekend. We define an ICP for each municipality. A further temporal partitioning is applied to the daily hours. A day is divided into three time-slot, representing interesting partitioning w.r.t. to user profiling. We represent the ICP as a matrix, where values in each cell represent the number of *days* when call events occurred (i.e., independently by the total number of calls), normalized with the number of days composing that slot. For example, if an individual performs one call on Monday, ten calls on Tuesday, and two calls on Friday in the whole week, the value in the weekdays of that week will be: $(1+1+1)/5 = 0.6$. Once each ICP is built, we can assume that a call event occurred in a specific municipality is a proxy of user presence in that territory and in that temporal slot.

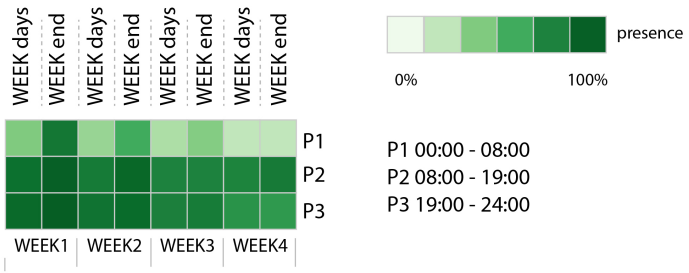


Figure 1: Individual call profile (ICP). Here, for the sake of simplicity, the intensity of presence is correlated with the saturation of the color instead of the percentage value representing the number of call activities.

3.2. Sociometer

ICPs are the input of *Sociometer* [36], a distributed data mining process for classifying users call behaviors. *Sociometer* relies on an analytical process consisting of several phases, showed in Figure 2. In particular, *phase 1* involves the *ICP Building* introduced in Section 3.1. Then, we group similar ICPs for *Prototypes Extraction*. The third part of the process (*Prototype Labeling*) assigns labels to the centroid of each prototype computed in the previous phase. The last phase is the *Label Propagation*: when each point of each prototype is labeled, it propagates the label of the prototypes to all the similar ICPs.

At the end of the process, for each ICP, we obtain a label that characterizes the individual in a specific class. In particular, five classes are considered:

- *Residents*: individuals who live and work in the same area; for this reason, their presence is significant across all days and all time slots for the specific municipality.
- *Dynamic Residents*: people who reside in some municipality A but work in a different one (B). The presence in A is expected to be significant always, except during working days and working hours (i.e., the time slot P2 of Figure 1).
- *Commuters*: people who reside (i.e., are Dynamic Resident) in some municipality B and whose work or study place is in A. The presence in A is expected to be almost exclusively concentrated during working days and working hours (i.e., the time slot P2 of Figure 1).
- *Visitors*: people that visit a municipality only a few times in the period.
- *Passing by individuals*: persons who are not actually living in a certain territory, but they merely traveled by the area covered by cells of the considered municipality.

Through the dataset of labeled user profiles, we can also quantify the different classes of individuals present in the area and the flows of individuals among the different regions.

3.3. PRUDEnce

The *Sociometer* analyzes data gathered by a Data Provider (a telecom operator), and it could also be applied externally to the environment of the Data Provider (DP), for example because it does not have the necessary resources to run the service. For this reason, the DP could have the need to share data of its users with an external entity, i.e., a Service Developer (SD). As already discussed in Section 2, mobile phone data are personal data, thus it is necessary to ensure the right to privacy of the individuals described in the data.

As a consequence, the DP needs a mechanism to measure the privacy risk and to apply safeguards on risky data to get the desirable level of privacy protection. To this end, the DP can count on the framework PRUDEnce [1]. This

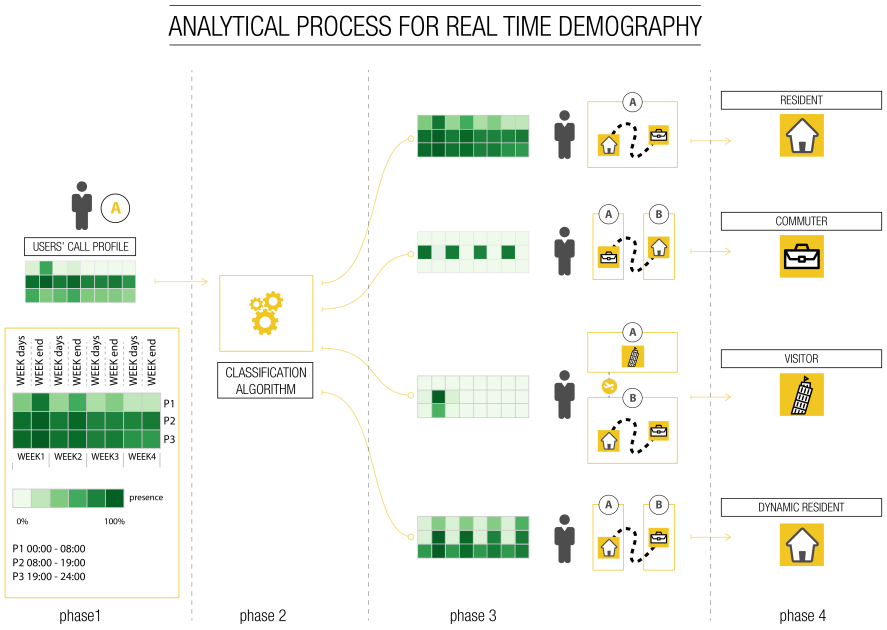


Figure 2: *Sociometer*. Starting from raw call data record, in phase 1, we first build for each user, for each zone an Individual Call Profile (ICP). Then, we apply a clustering algorithm to group users with similar behavior (phase 2). From each cluster, we extract a centroid (phase 3), and we label it w.r.t. the closest representative archetypes (phase 4).

framework, relying on the Privacy-by-Design paradigm [22, 37], offers a methodology that, by a privacy risk assessment module and a privacy risk mitigation module, first measures the empirical privacy risk of a specific set of data, and, then, reduces that risk if it is not compliant with the needs of the DP.

In order to verify the risk of privacy, through PRUDEnce, the DP queries its users data, producing a dataset suitable for the service to be developed (i.e., maintaining no more than the level of information really needed by the service[3]). Then, relying again on PRUDEnce, the DP: *(i)* identifies the background knowledge that an adversary might have about his/her target; *(ii)* simulates the attack based on that background knowledge, computing the privacy risk values for every individual; *(iii)* applies a privacy risk mitigation method (e.g., generalization [38], randomization [39], suppression [38]) on the dataset; and *(iv)* delivers the sanitized dataset to a third party, i.e., the SD who wants to apply the *Sociometer*.

PRUDEnce adopts the risk re-identification as privacy risk; the related attacks assume that an adversary gains access to a dataset and, using some background knowledge about an individual under attack, he/she tries to re-identify

[3]This is compliant with the *data minimization* principle described in the GDPR

that individual in the dataset. The background knowledge represents both the kind and quantity of information known by the adversary. We use b to indicate the specific background knowledge (e.g., the fact that a user performed a call in a particular location on a specific day). An individual is hence associated with several privacy risks, each for every background knowledge of an attack.

In the following, we provide the formalization of the above description, defining the measure that we use to quantify the privacy risk.
Let $\mathcal{D}$ be a database, D a dataset derived from $\mathcal{D}$ (e.g., an aggregated data format on time and/or space, such as the ICP introduced in Section 3.1), and D_u the set of records representing a user u in D, the probability of re-identification is defined as follow.

Definition 1 (**Probability of re-identification** [1])**.** *Given an attack, a function* matching(d, b) *indicating whether or not a record $d \in D$ matches the background knowledge b, and a function $M(D, b) = \{d \in D | matching(d, b) = True\}$, we define the* probability of re-identification *of an individual u in dataset D as: $PR_D(d = u|b) = \frac{1}{|M(D,b)|}$ that is the probability to associate record $d \in D$ to individual u, given the background knowledge b.*

Note that $PR_D(d{=}u|b) = 0$ if the user u is not in D.

4. Problem definition and proposed solution

As discussed in Section 2, mobile phone data are subject to privacy issues. Our aim is to enable the sharing of this kind of data achieving two important but conflicting goals: on the one hand, we surely want that adequate privacy guarantees are provided, in order to limit the privacy risk of the individuals described in the data; on the other hand, shared data should not be too distorted, in order to ensure that specific analyses, such as the *Sociometer* (Section 3.2), are still possible, maintaining a good quality level of service.

Thus, we propose to apply the framework PRUDEnce (Section 3.3) to mobile phone data, in particular to Individual Call Profiles (Section 3.1), in order to generate a privacy-preserving version of them which can be used, for example, to provide a privacy-aware census of population. First of all, the DP extracts the set of ICPs related to the required territory and the specified time window. Then, the DP evaluates the privacy risk associated with the set of ICPs, and, finally, if this risk is above a certain threshold, a mitigation strategy to unsafe ICPs is applied. At this point, only non-risky data will be shared with the SD, which may apply the *Sociometer* to the received safe ICPs, thus labeling each individual in the appropriate category of city users.

PRUDEnce strongly relies on the Privacy-by-Design paradigm [22], which, in order to design a privacy-preserving framework, requires some assumptions about: *(i)* the personal data that are the subject of the analysis; *(ii)* the attack model, i.e., the purpose of a malicious party that has an interest in discovering the sensitive data of certain individuals; *(iii)* the category of analytical queries that are to be answered with the data.

In our setting, the data are the ICPs, while the service to be provided is the *Sociometer*. Thus, we still need to define:

- a possible attack model to quantify the privacy risk, and
- a mitigation strategy for the privacy risk whether it is above a specified threshold.

Therefore, in Section 5, we describe the privacy risk assessment step, i.e., the attack model we want to simulate, while in Section 6, we focus on the task of privacy risk mitigation, presenting PRIMULE, our solution for achieving a trade-off between individual privacy and quality of the service.

5. Privacy Risk Assessment

The quantification of the probability of re-identification of each individual in the data requires to simulate a privacy attack. Our attack model is based on the *linking attack* [38], and it uses a specific and strong background knowledge. Indeed, in our setting, the attack is based on a perfect knowledge by the adversary of the call activities of his/her target in the observed area. In other words, for a specific time window and geographical area, the idea is to quantify the probability of re-identification of a target, in case the attacker would know *if* and *when* his/her target performed a call.

Exact Background Knowledge. We assume that the attacker knows exactly the call activities of a user U (i.e., the fact that he/she called someone and the time of these calls) during a visit at a certain location, for a certain time window (i.e., one week, two weeks, and so on). This means that with this knowledge, the adversary can build the corresponding ICP denoted by PB, where PB_{ij} represents a temporal slot (as shown in Figure 1). $PB_{ij} = -1$ if the attacker does not have any information about the call activity of the user in the period (i, j), while $PB_{ij} = v$ $(v \geq 0)$ if from the background knowledge he/she derives that the user was present in the area v times in the period (i, j). Note that for this last period PB_{ij} is exactly equal to the one owned by the telco operator. As an example, suppose that an adversary shadowed Mr. Smith for some period. With this information, for this period he/she can build an ICP as accurate as the one that can be found in the dataset.

Attack Model. The attacker, who gains access to the set of ICPs $\mathcal{P}$, uses the background knowledge PB on the user U to match all the profiles that include PB. The set of matched profiles is the set $C = \{P \in \mathcal{P} | \forall\ PB_{ij} \geq 0. PB_{ij} = P_{ij}\}$.

Once defined both the possible background knowledge and the attack model, we can simulate this attack, in order to quantify the probability of re-identification of the user U, which is $\frac{1}{|C|}$. Clearly, a greater number of candidates corresponds to a better privacy protection.
In this paper, we propose a mitigation strategy that directly integrates the privacy risk assessment, which requires the simulation of this privacy attack.

6. PRIMULE: Privacy RIsk Mitigation for User profiLEs

We figure out a method that is based on the knowledge of the *Sociometer* process and has the goal to get a set of *safe* ICPs. A profile is considered *safe* if it is indistinguishable from at least others k-1. In other words, considering our attack model, a profile is *safe* if its probability of re-identification is at most $\frac{1}{k}$, where k is a parameter of PRIMULE (Algorithm 1).

Thus, our basic idea is to create groups of indistinguishable profiles by rendering equal those profiles which are already quite similar and thus assigning the unsafe ICPs to the closest group. We conceive a mitigation strategy using the k-anonymity privacy model depending on: *i)* the ICPs' properties, *ii)* the background knowledge and *iii)* the service to be developed. We recall that in our setting the service goal is the classification of typologies of city users (residents, commuters, dynamic residents, visitors and passing by individuals), as discussed in Section 3. In the following, for the sake of simplicity, we consider the probability of re-identification assuming the knowledge of the first 1/2/3/4 week(s). However, other portions of the ICPs can be considered and managed by PRIMULE (see Section 6.2).

Our approach is based on two major principles. The first one is that two ICPs are indistinguishable if they exactly match w.r.t. the portion of adversary background knowledge, because this is the actual data used to perform each attack. For example, if we are considering a background knowledge of 2 weeks, two individual call profiles are indistinguishable only if they are equal in the portion of them corresponding to the first 2 weeks. The second idea at the base of our reasoning is that indistinguishable profiles represent an equivalence class w.r.t. the background knowledge, thus, each element in the class has the same probability and we can manage the whole group as a single entity.

In a nutshell, the mitigation strategy creates groups of indistinguishable ICPs, then, it tries to aggregate ICPs groups which are as much as possible similar each other, two at a time. The aggregation strategy is a weighted average between the two groups of profiles. However, different policies could be defined. If the profiles are still not safe, the process is iterated.

6.1. Example of the working principle of the proposed strategy

In Figure 3, one can find a simplified example of the idea behind our approach. On the $x-$axis we indicate some single groups of indistinguishable ICPs (from A to H) with the correspondent cardinality. On the $y-$axis is reported the number of the current iteration step of the algorithm. We fix the privacy threshold to $k = 5$, i.e., a probability of re-identification of $\frac{1}{5}$. In the picture the unsafe groups B, C, D, F and G (i.e., groups with cardinality lesser than k) are represented in red, and the safe groups A and H (i.e., groups with cardinality higher or equal to k) are represented in black. In the first iteration, we merge the groups B and C, both with cardinality equal to 2, creating a new group (i.e., BC), consisting of 4 ICPs. The 4 elements of the BC groups are not the original ICPs coming from B and C, but are computed ICPs derived from the weighted average of the original elements of B and C. In general, more than two groups

can be aggregated at the same time, as we can see at Iteration 2. The process is iterated until at least an unsafe group exists.
It is important to point out that the merging operations are performed if at least one of the groups is unsafe. In the example, we can see how in Iteration 1 ($B \cup C$) both starting groups (i.e., B and C) are unsafe, while in the Iteration 4 ($A \cup BC$) the A group is already safe. However, it represents the best option (i.e., the most similar group) for the unsafe group BC; thus, we decide to merge these two groups in order to minimize the transformation of ICPs.
The last case is represented by the H group, which is safe from the starting point and it is not merged with any other group in the whole execution; this because it is too far from all the unsafe groups.

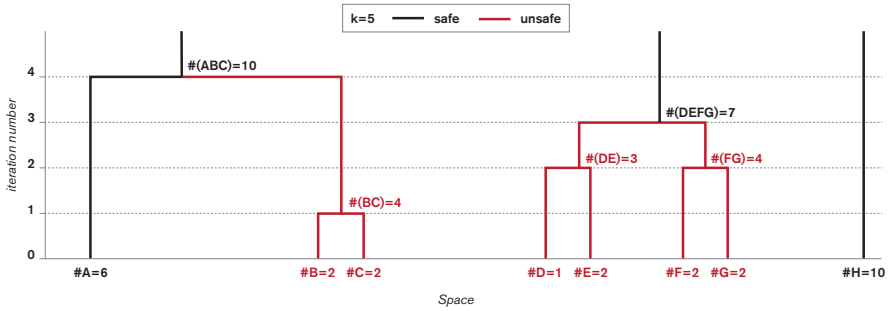


Figure 3: Example of the PRIMULE strategy.

Algorithm 1 PRIMULE($\mathcal{P}$, h, k, $dist$)

Inputs: The set of SCT profiles $\mathcal{P}$, the list of indexes representing the portion of profile known by the adversary h, the required group size k, the chosen distance $dist$.
Output: The privacy-preserving SCT profiles $\tilde{\mathcal{P}}$ (if it is possible) and a flag that indicates if the result is anonymous or not.
if $|\mathcal{P}| < k$ **then**
    return $\mathcal{P}, false$
//create two empty sets for groups ($\mathcal{G}$) and unsafe groups ($\mathcal{G}_{unsafe}$)
$\mathcal{G} = \emptyset$; $\mathcal{G}_{unsafe} = \emptyset$
//cluster profiles based on the background knowledge dimension
$\mathcal{G} = CreateGroups(\mathcal{P}, h)$
for all group $g \in \mathcal{G}$ **do**
    $count = ComputeCardinality(g)$
    if $count < k$ **then**
        $\mathcal{G}_{unsafe} = \mathcal{G}_{unsafe} \cup g$
$\tilde{\mathcal{P}}$ = ProfileMitigation($\mathcal{G}$, $\mathcal{G}_{unsafe}$, h, k, $dist$)
return $\tilde{\mathcal{P}}, true$

6.2. PRIMULE algorithm

Our strategy is shown in Algorithms 1 and 2. Algorithm 1 is composed of the following procedures: i) risk assessment phase and ii) mitigation phase.

Algorithm 2 PROFILEMITIGATION($\mathcal{G}$, $\mathcal{G}_{unsafe}$, h, k, $dist$)

1: Inputs: The set of group $\mathcal{G}$, the set of unsafe group $\mathcal{G}_{unsafe}$, the list of indexes representing the portion of profile known by the adversary h, the required group size k, the chosen distance $dist$.
2: Output: The privacy-preserving ICPs $\tilde{\mathcal{G}}$.
3: $\tilde{\mathcal{G}}$=$\mathcal{G}$
4: **while** $\mathcal{G}_{unsafe} \neq \emptyset$ **do**
5: **//create a list of couple of groups and relative distances** ($D = \langle g_1, g_2, dist(g_1, g_2)\rangle$)
6: $D = \emptyset$
7: **for all** unsafe group $g \in \mathcal{G}_{unsafe}$ **do**
8: **//search for the nearest group**
9: $d_{min} = ComputeMinimumDistance(g, \tilde{\mathcal{G}}, dist, h)$
10: $g_{min} = ComputeMinimumDistanceGroup(g, \tilde{\mathcal{G}}, dist, h)$
11: $D = D \cup \{g, g_{min}, d_{min}\}$
12: $D = OrderByDistanceAsc(D)$
13: **//create empty list of updated group** ($\mathcal{G}_{modified}$)
14: $\mathcal{G}_{modified} = \emptyset$
15: **for all** element $d \in D$ **do**
16: $g_2 = TakeSecondGroup(d)$
17: **if** $g_2 \notin \mathcal{G}_{modified}$ **then**
18: **//there is no conflict: merge the two groups**
19: $g_1 = TakeFirstGroup(d)$
20: $\tilde{\mathcal{G}} = \tilde{\mathcal{G}} \setminus g_1$
21: $\tilde{\mathcal{G}} = \tilde{\mathcal{G}} \setminus g_2$
22: $\mathcal{G}_{unsafe} = \mathcal{G}_{unsafe} \setminus g_1$
23: $\mathcal{G}_{unsafe} = \mathcal{G}_{unsafe} \setminus g_2$
24: $new\_P^h = WeightedAverage(g_1, g_2, h)$
25: $new\_cardinality = |g_1| + |g_2|$
26: $UpdateInformation(g_2, new\_P^h, new\_cardinality)$
27: **if** $new\_cardinality \leq k$ **then**
28: $\mathcal{G}_{unsafe} = \mathcal{G}_{unsafe} \cup g_2$
29: $\tilde{\mathcal{G}} = \tilde{\mathcal{G}} \cup g_2$
30: $\mathcal{G}_{modified} = G_{modified} \cup g_2$
31: **return** $\tilde{\mathcal{G}}$

Risk assessment phase. The risk assessment phase (Algorithm 1, lines 3-12) takes in input the set of original profiles $\mathcal{P}$ and the background knowledge h, i.e., the list of cells indexes representing the portion of profile known by a potential adversary. In the following, given h and a profile P, we denote by P^h the portion of the profile identified by the cells indexes in h.
Firstly (line 3), it is checked if the total number of profiles is big enough to guarantee a certain privacy threshold k. If no, the original profiles are returned, along with the information that it was not possible to anonymize them (line 4). This can happen if the privacy threshold is too high and the municipality is sparsely inhabited. Then, the algorithm groups together ICPs, having the same values in the portion P^h, by the function *CreateGroups* (line 8) and, then (line 12), selects the unsafe groups, i.e., groups with cardinality lower than k.

Mitigation phase. The mitigation phase (line 13 of Algorithm 1) acts on unsafe groups and is detailed in Algorithm 2, where:

1. For each unsafe group, the algorithm selects the nearest neighbor among all the groups, i.e., both safe and unsafe groups (Algorithm 2, lines 7-11). For identifying the nearest neighbor group the procedure uses a distance function *dist* only on the portion of the profiles P^h. We recall that h

identifies the cells indexes representing the background knowledge, i.e., the portion of profiles known by the adversary. Each couple of groups represents a potential assignment of the unsafe group to the other one, but, at this stage, no merging is performed yet.

2. The possible assignments found in the previous step are ordered from the most promising one (Algorithm 2, line 12), i.e., considering increasing distances.
3. Starting from the first couple, we perform the assignment procedure: we check whether the selected group was previously merged with other groups (Algorithm 2, line 17); if not, we merge the two groups (Algorithm 2, lines 19-26) and we check if the new group has a size of at least k (Algorithm 2, line 27). The merging operation also requires to make all the ICPs in the new group equal with respect to the portion of the profile identified by h. Therefore, the elements in the new group are transformed in such a way that the portion P^h of each profile is updated by computing the weighted average of all profiles in the group. In other words, for each group g the algorithm computes new_P^h as follows: $\forall (i,j) \in h.\ new\_P^h_{ij} = \frac{\sum_{P \in g} P_{ij}}{|g|}$.

Steps 1) and 2) are not deterministic because it is possible to have two equidistant groups from the same unsafe group, and two different couples of groups can have the same distance. This choice is reasonable for providing an efficient solution for the mitigation process. Obviously, we could add some clauses, like favoring bigger/smaller groups, in order to render the process deterministic.

The control described in Step 3) and reported in Algorithm 2, line 17 is necessary because the simultaneous crossed assignments can lead to the following situation. A and B are the nearest neighbors, thus: i) A is selected and merged with B, creating the group $A \cup B$, and ii) the group B is selected for being merged with A, but the two single groups no longer exist.

Computation Costs. We conclude with a final consideration regarding the time computational cost: the matching phase and the computation of distances can be performed in parallel, relying on distributed technologies like Map-Reduce. In order to assign each unsafe group to the nearest one, we adopt a k-nearest-neighbor strategy based on an implementation of the k-d tree [40]. Building a static k-d tree from n points costs $n \times log(n)$, while finding one nearest neighbor in a balanced k-d tree with randomly distributed points takes $log(n)$ time.

6.3. Correctness of PRIMULE

Now, we provide the proof of the algorithm termination (Theorem 6.1) and of the achievement of privacy requirements (Theorem 6.2).

Theorem 6.1. *Algorithm 2 terminates.*

Proof. If the privacy threshold k is greater than the number of ICPs, then the mitigation approach cannot succeed, and the Algorithm 1 immediately terminates. Otherwise, the algorithm terminates when there are no element left in $\mathcal{G}_{unsafe}$. In each iteration, we compute the nearest neighbor for every unsafe

group (as far as it can be, there is always a nearest neighbor), and we join two groups if there is not any conflict. A conflict occurs when a group is already treated in the current iteration, so at least the first attempt (the one that joins the two nearest groups) will succeed. Therefore, in every iteration, the number of (unsafe) group decreases by at least 1. Hence, the algorithm terminates. □

Theorem 6.2. *Algorithm 2 satisfies the privacy risk requirement.*

Proof. Established that there is an adequate number of ICPs (see Algorithm 1), the correctness of Theorem 6.2 is straightforward due to the presence of the loop. The program only terminates when there are no existing groups smaller than k, i.e., it terminates only when all the profiles have a probability of re-identification at most $\frac{1}{k}$ respect to the input background knowledge. □

6.4. Information loss

We can measure utility through well-known metrics, such as the information loss based on the Mean Square Error (MSE), presented in the Domingo-Ferrer and Torra work [41]. Indeed, our resulting clusters have an MSE which is: $\frac{1}{n}\sum_{i=1}^{n}(x_i - x'_i)^2$, where $1 \dots n$ are the indexes of original ICPs, and x_i and x'_i are respectively the original ICPs and the corresponding private version, limited to the portion of the profile known by a potential adversary, i.e., P^h, as explained in Section 6.2. This measure enables the identification of the maximum information loss that could be obtained with the PRIMULE transformation. Indeed, *in the worst case* we would obtain a single cluster, where the centroid is equals to the weighted average of the ICPs' values of all the original clusters. So, in this case, the information loss is given by $\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x})^2$, where $\hat{x}$ is the weighted average, cell by cell, of all the P^h. This represents a superior bound of the actual information loss.

7. Experiments

In this section, we compare the results of the evaluation of our approach, showing the outcome of the application of the *Sociometer* on three different sets of data. We present the *Sociometer* applied to ICPs treated with PRIMULE, our ad hoc mitigation strategy (Section 6.2), comparing its outcomes with the *Sociometer* applied to profiles without any kind of sanitization (i.e., this represents our baseline, since the ICPs are the original ones) and to profiles perturbed by an approach based on the differential privacy paradigm. For our experiments, we start from a CDR dataset that regards the territory of a significant part of Tuscany (139 municipalities out of 279). The dataset was provided by one of the major Italian mobile operators, and it consists of about 85 million CDRs from about 3 million customers. The covered period is the month of November 2016 (4 weeks). This data are already pseudonymized, and the pseudo-ids change every 4 weeks. Thus, we cannot link individuals among different supplies of data, and, for our purposes, 4 weeks represent the maximum possible background knowledge.

In our experiments, we managed data, and we implemented our algorithms using Spark on a Hadoop cluster composed by 4-nodes, each one with 6-cores Intel XEON@2.93Ghz, 24GB Ram, and 2 TB storage capacity.

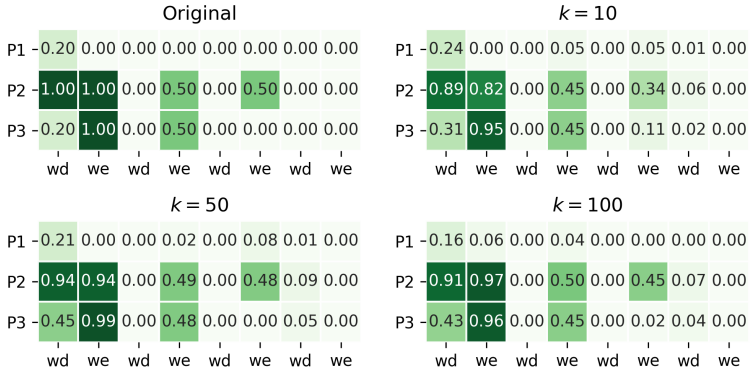
In the following, we show the evaluation of our approach by varying the parameter k (the threshold of the minimum number of indistinguishable profiles) from 10 to 100. Moreover, we use the Euclidean distance function to compute the profile similarity, and we set the background knowledge to 4 weeks.

Differentially Private Approach. As reported in Section 2, the Differential Privacy model provides strong privacy guarantees once the sensitivity of the query is established. Moreover, the Laplace mechanism is useful when data are numeric. Thus, we implemented an approach based on Differential Privacy as a competitor. The comparison has the goal to evaluate the performance of the two approaches in terms of user profiles utility. We are aware that differential privacy based approaches do not take into consideration any background knowledge and thus provide a different privacy guarantee. However, this analysis is useful to deeply understand the data utility implications that sometimes could lead to make data useless.

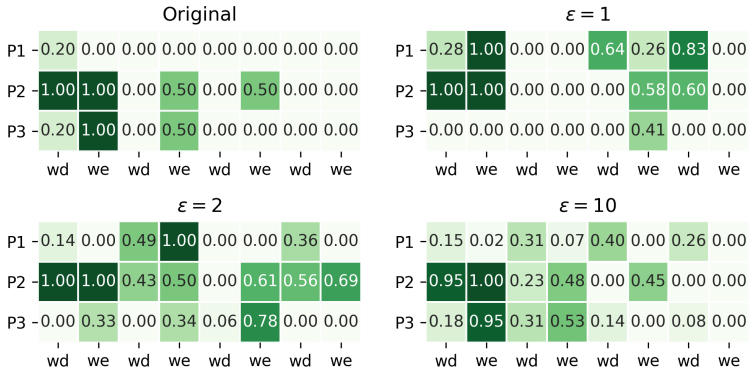
Basically, we add appropriately chosen random noise to the true query answer (i.e., the real value of each ICP's position), and we return the noisy answer. We apply the Differential Privacy fixing the sensitivity to 1 (i.e., the maximum value that each cell of the ICPs can have), and we extract an appropriate noise for each cell of each ICP; finally, we add these noise values to the ICPs, obtaining a differentially private version of it. Since the new values can be lower than 0 or greater than 1 (i.e., they can represent not informative values for the *Sociometer*), we apply a post-processing that forces the values inside the admissible range (i.e., from 0 to 1). Since it is a post-processing procedure this step does not affect the privacy guarantees [42]. Note that decreasing ϵ, a publicly known parameter, leads to greater privacy protection. This is due to the fact that the magnitude of the noise drawn from a Laplace distribution, which depends on both the global sensitivity of the query and the desired privacy level ϵ, becomes higher. In the following, we provide an example of a profile obtained by applying this approach, varying the privacy parameter ϵ.

7.1. Effect of the mitigation strategies on ICPs

In order to show the effect of the two privacy transformations on the single ICPs, we report in Figure 4 some private versions of one of the original ICPs, and we highlight the differences. The different versions of ICPs depicted in the figure correspond to different levels of privacy protection. Figure 4 (a) reports the ICPs obtained applying PRIMULE, our ad hoc mitigation strategy, with $k = 10, 50, 100$, while Figure 4 (b) shows the ICPs obtained applying the differentially private approach with $\epsilon = 1, 2, 10$. In both cases, we provide a comparison with the original ICP. As we can see, there are some fluctuations in the single values of the ICPs obtained by our strategy, but the general behavior of the individual is preserved since we try to merge similar profiles. In the differentially private ICPs, we decided to vary the privacy threshold given by



(a) PRIMULE

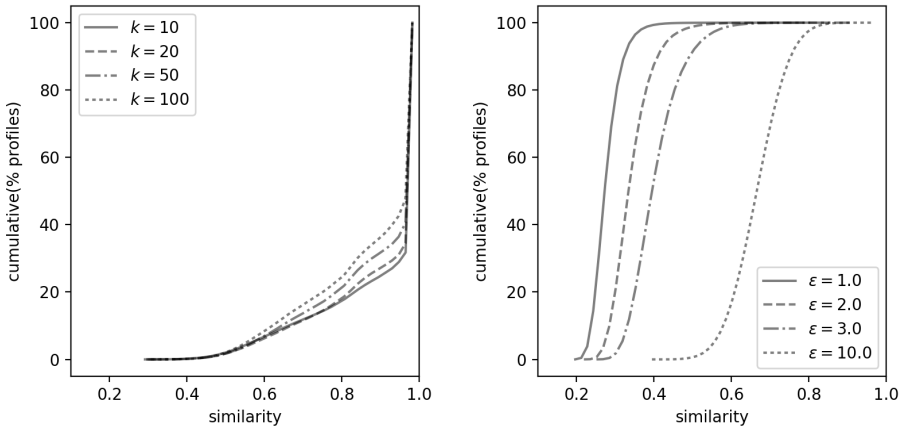


(b) Differential Privacy

Figure 4: Example of one ICP transformed using the two mitigation strategies, and varying the correspondent privacy protection level.

the ϵ parameter, which affects the shape of Laplace distribution and, thus, the magnitude of the noise added to the original values (we recall that greater ϵ leads to lower privacy protection). In particular, we chose $\epsilon = 1$ that is often considered the maximum valid value [43], but also other occasionally used values as $\epsilon = 2$ [44, 45] and $\epsilon = 10$ [46]. Here, we observe that with $\epsilon = 1$, we have important variations w.r.t. the original ICP. As an example, we can observe the second column, where the 0 value (which represents an absence during the specific period) becomes 1 (which indicates that the individual performed at least a call in both the weekend days), and, vice-versa, the last 1 becomes 0. This behavior is due to the fact that this approach does not try to preserve the

general profile, but it only aims to hide the real values of ICPs.



(a) PRIMULE effects on similarity

(b) Differential privacy effects on similarity

Figure 5: Cumulative curves, for both approaches and varying the privacy parameters, representing the ICPs having a certain similarity with the original correspondent profile.

Now, we can analyze the global situation of the two transformations on ICPs. In particular, we show how much the proposed mitigation strategy and the one based on differential privacy affect the real values of ICPs. Hence, in Figure 5 we report the similarity of the original ICPs w.r.t. the sanitized ones, both with the PRIMULE approach (Figure 5 (a)) and the differential privacy strategy (Figure 5 (b)). In the picture, we illustrate the similarity with respect to the Euclidean distance since is one of the simplest distances that can be used. The two plots show the cumulative curves that represent the percentage of ICPs having a certain similarity with respect to their correspondent original ICPs. As already reported in the previous analysis, for both the mitigation approaches, we also show what happens to vary the correspondent privacy parameter, i.e., the indistinguishability threshold k and the privacy budget ϵ.

As one can see in Figure 5 (a), with $k = 100$ we have that around 50% of sanitized ICPs are almost identical to the original ones (similarity greater than 0.95). Decreasing the privacy threshold to $k = 10$, we get a massive increase of data quality since around 70% of ICPs have this high value of similarity.

On the contrary, the cumulative curves depicted in Figure 5 (b) have a totally different shape. Even using a quite weak privacy guarantee, i.e., $\epsilon = 10$, we have that less of 10% of sanitized ICPs have a similarity with the original ones of at least 0.8. Instead, about 80% of sanitized ICPs have a maximum similarity with their correspondent ones of around 0.7. This is drastically reduced if we increase the privacy budget to 3: a similarity of at least 0.7 is obtained only by 3% of ICPs, while 80% of ICPs have a maximum similarity of around 0.4. If we increase the privacy guarantees again, we have that the vast majority of

sanitized ICPs have a similarity with the original data between 0.25 and 0.4 when $\epsilon = 2$ and between 0.2 and 0.3 when we fix $\epsilon = 1$.

7.2. Sociometer application on both real and private data

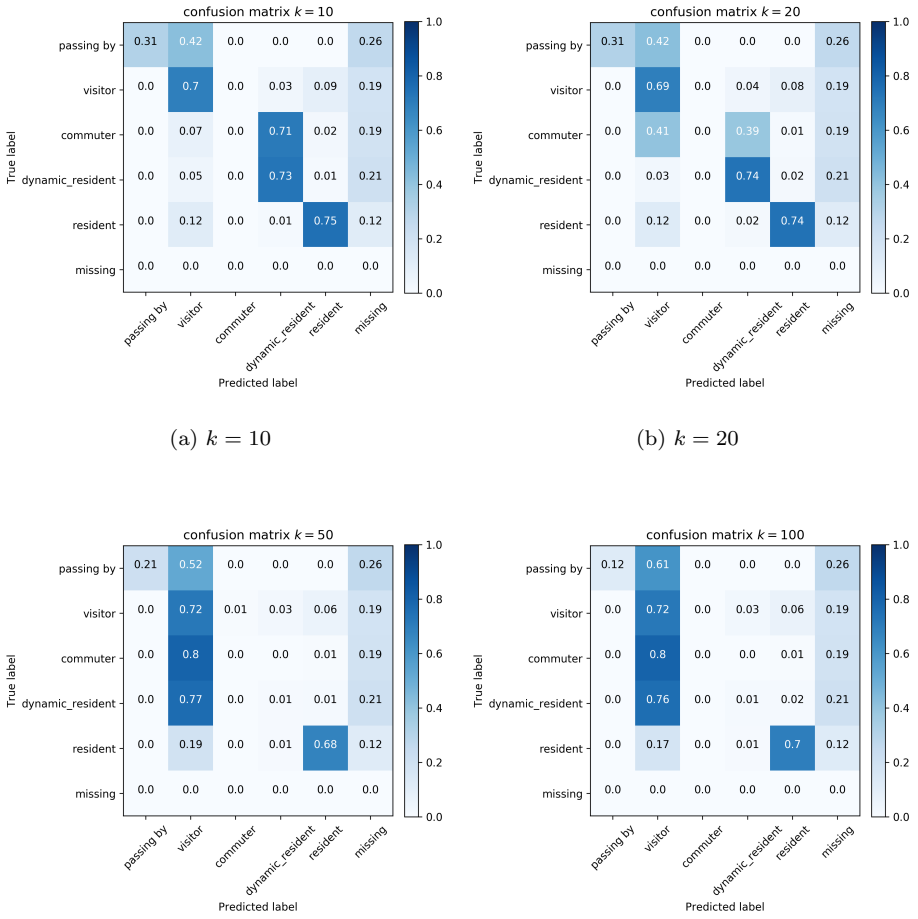
In this section, we evaluate the impact of our mitigation strategy on the classification and quantification of the five categories recognized by the *Sociometer*. To this end, we apply the *Sociometer* to the original data and to the private data and analyze the results. We compare the impact of both our mitigation strategy and that one based on differential privacy.

To evaluate the effects on the classification, we analyze the confusion matrices by considering the result of the *Sociometer* applied to the original ICPs as the actual class of city users, since we do not have any other official information about their real labels. We also report an analysis of the f-score. To assess the effects of the privacy transformations, we compare the results of the quantification task on the original data with those obtained applying the quantification on private data.

Classification Evaluation. Figure 6 & Figure 7 show the confusion matrices obtained by applying the *Sociometer* on ICPs anonymized by our strategy and the differential private approach, respectively.

Figure 6 highlights that, for lower privacy levels (i.e., $k = 10$ and $k = 20$), visitors, dynamic residents and residents are quite well preserved. On the contrary, passing by individuals are misclassified because they are confused with visitors (actually, these two classes are really similar in terms of call activities). Commuters tend to disappear because they are very few (only 3% of the total number of individuals), so for our algorithm is quite difficult to find similar groups that are also commuters. If we increase the privacy level to higher values (i.e., $k = 50$ and $k = 100$, which are quite high thresholds), we can see that all the classes apart from residents are mainly classified as visitors.
However, it is worth to notice that the increase in the privacy level does not determine a decay in the quality of the residents' classification. This result is due to the fact that these users are individuals with quite equally distributed calls in the period, so our mitigation strategy succeeds in maintaining this variety. This is a quite remarkable result because it can enable the development of other analytical tools able for example to discover fictitious residences: this is a notorious problem in Italy, where many people declare fictitious residences in order to avoid payment of taxes.
Visitors are quite correctly labeled, too. However, this set of users includes a quite large number of dynamic residents. This could be due to the fact that dynamic residents call prevalently during the weekends and evenings, so it is very likely that they have an ICP quite similar to a visitor, e.g., a tourist. For the same reason, dynamic residents are very often misclassified as visitors. Finally, commuters are often classified as visitors. Once again, it is quite straightforward to establish that these two categories can have quite similar profiles, at least in a time window of one month, like in our case.

It is worth noting that these results are referred to the general situation, i.e., analyzing ICPs without dividing them by their municipalities. We chose to report only the total situation, i.e., a summarization of all the municipalities because our experiments show that the population density of the municipalities does not affect the general structure of matrices too much. Indeed, all the confusion matrices related to the same privacy threshold but to different municipalities show the same behavior.

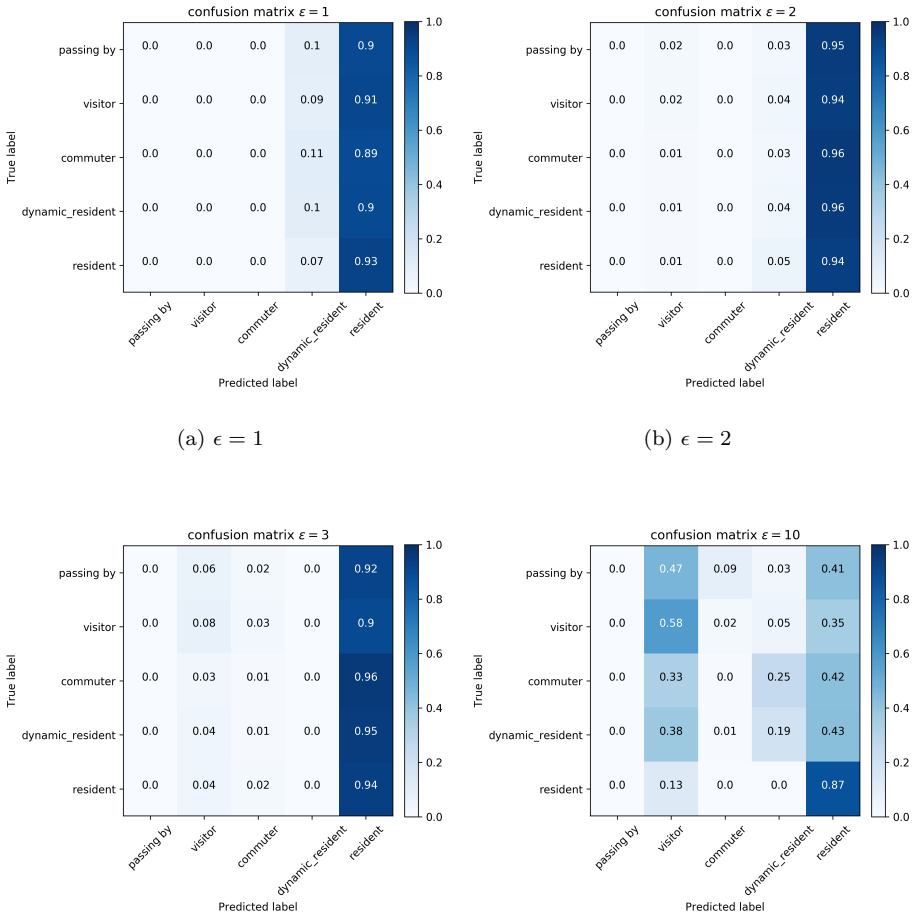


(a) $k = 10$ (b) $k = 20$

(c) $k = 50$ (d) $k = 100$

Figure 6: Confusion matrices obtained using PRIMULE, varying the privacy protection parameter from k=10 to k=100, respectively the lowest and highest privacy thresholds we tested.

In Figure 7, we report the confusion matrices of the *Sociometer* applied to the differentially private data. Here, for $\epsilon = 1, 2, 3$ (Figure 7 (a), (b) and (c),



(a) $\epsilon = 1$ (b) $\epsilon = 2$

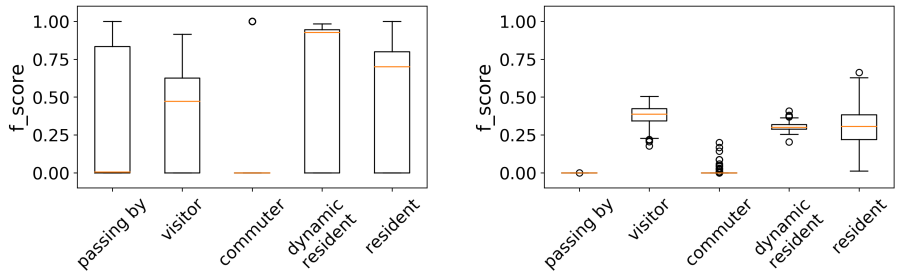
(c) $\epsilon = 3$ (d) $\epsilon = 10$

Figure 7: Confusion matrices obtained using the differential privacy approach, varying the privacy protection from $\epsilon = 1$ to $\epsilon = 10$, respectively the highest and the lowest privacy thresholds we tested.

respectively) we observe how the almost totality of individuals are labeled as residents. This is due to the fact that introducing noise for each cell, the 0s tend to disappear, and thus, all the ICPs become quite dense, with call activities distributed in the whole period, i.e., they correspond to residents. This is a direct consequence of the general solution offered by differential privacy approaches: the strong point is that they can work with low adjustment to several scenarios; the weakness is that they are not suited for a specific case, thus they cannot take into consideration any particular properties regarding the data (ICPs) or

the service (the *Sociometer*) during the data transformation.
If we increase ϵ to 10, we have a different situation: residents are well preserved, but the other actual classes are split among the different labels. However, even using a very big ϵ, we obtain values that are worse than the one obtained using PRIMULE, since visitors are not well predicted either.

In Figure 8, we report the f-score obtained with PRIMULE (Figure 8 (a)) and with the differential privacy approach (Figure 8 (b)), showing the distribution of values for all the municipalities under analysis. We observe that the f-score for both passing by individuals and commuters is very low in the median. However, for the remaining classes, using PRIMULE we obtain quite sparse results, but the f-score medians are around 0.5 for visitors, around 0.7 for residents and above 0.9 for dynamic residents. On the contrary, using the differential privacy strategy, values are more stable but always below 0.5 for all categories.



(a) PRIMULE f-score ($k = 20$)

(b) Differential privacy f-score ($\epsilon = 10$)

Figure 8: f-score comparison.

Quantification Evaluation. The *quantification task* [47] aims to accurately estimate the number of positive cases (or class distribution) in a test set, using a training set that may have a substantially different distribution. Figure 9 shows the presences in any user category measured after applying our method with $k = 10$ and $k = 100$, compared with those computed after applying the Differential Privacy approach with $\epsilon = 1$. Once again, our ground truth is defined as the stock of presences observed by the *Sociometer* technique applied to the original data. Here, we prefer to show the results associated to actual municipalities (instead of a summarization), because we want to show the real amount of individuals in each category, instead of the percentage, in order to provide a more accurate picture of the real situation. Moreover, the quantification results present a greater variability among the cities. For the sake of simplicity, the figure refers to four cities of different sizes, which are quite representative of our findings. In order to perform the comparison using only one plot per municipality, we decided to limit the shown results concerning the chosen parameters. Indeed, we report two instances of PRIMULE, in order to investigate its out-

come better, and only one case of the differentially private approach. In this latter case, in order to be as fair as possible, we choose to display the greater ϵ, which is worse for the privacy perspective but it is the better in terms of utility.

Our methodology provides a good estimation of the individual presences for the resident, dynamic resident and passing by categories.We obtain an overestimation of visitors and completely destroys the information regarding the commuter. This behavior is generally present in all four cities albeit with a different impact. In the specific cases of Pisa and Vecchiano, many passing bys are classified as visitors. This outcome is due to the fact that ICPs belonging to these two categories are very similar since their correspondent individuals are present only a few times. However, it is important to point out that if we would join visitors and passing by users into a single category, the quantification would give quite accurate results in all the municipalities.

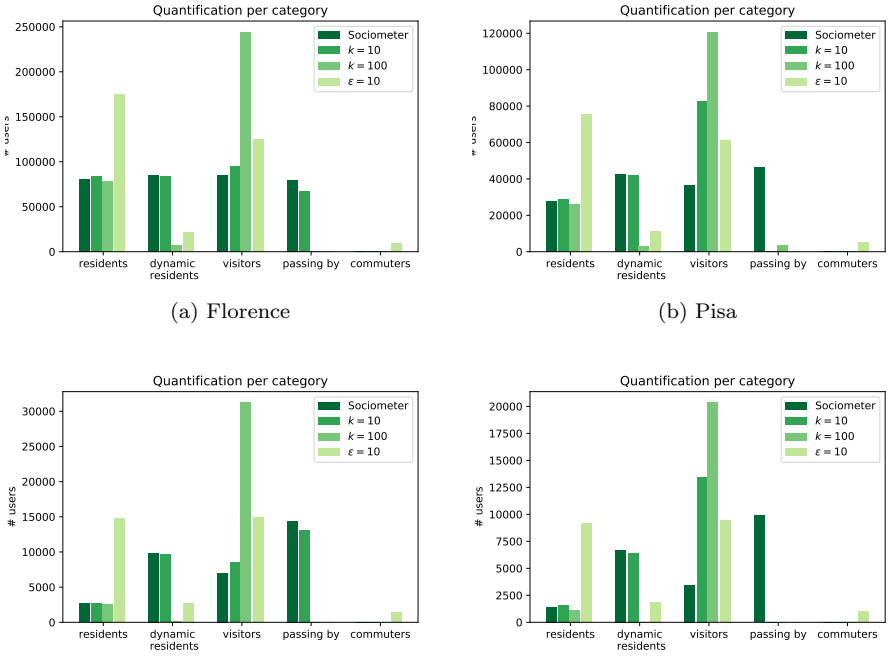
On the other hand, it is immediately evident that the differential privacy technique tends to turn a large majority of users into residents, even if we are reporting results using a high ϵ parameter. This result has already been observed in Figure 7 and in the description of the general confusion matrix. The same considerations about the reason of this overestimation hold.

8. Conclusion

In this paper, we have studied the problem of guaranteeing privacy protection for individual user profiles, which describe the call activity registered by mobile phones. Our mitigation strategy, called PRIMULE, provides privacy protection by making similar profiles indistinguishable to eliminate possible risky cases. The proposed approach relies on the privacy risk assessment framework PRUDEnce [1] for the identification of risky profiles. PRIMULE is particularly tailored to individual call profiles, which are the building blocks of the *Sociometer* [2]. In the definition of the mitigation strategy, we took into consideration how the *Sociometer* works for trying to maintain some properties in the profiles useful for the quality of this service. However, we observe that PRIMULE may be applied to mitigate the privacy risk of any set of profiles represented by numerical matrices.

After proving the theoretical privacy protection provided by PRIMULE, we have also demonstrated its effectiveness by a wide set of experiments on CDR data covering the territory of Tuscany. The results showed that the quality of the profiles is good in terms of both classification and quantification performance for some important categories of city users. In particular, the private version of profiles enables a good quantification especially of residents, while in terms of classification the *Sociometer* is able to get a good performance also for visitors. Experiments also showed a high similarity between original and private ICPs, suggesting that we could go a step further the *Sociometer* goal, trying to conceive new services that need a similar kind of data.

In each experiment, we also compared PRIMULE against a method based on differential privacy [9]. Our results showed that, in general, PRIMULE



(a) Florence

(b) Pisa

(c) Collesalvetti

(d) Vecchiano

Figure 9: Quantification, in four different municipalities, of the five categories labeled by the *Sociometer* applied to the original ICPs, to the ICPs returned by our mitigation approach (varying the anonymity threshold k) and to the differentially private ICPs. Florence (a) and Pisa (b) are quite large cities (their mobile users are around 350K and 180K, respectively), while Collesalvetti (c) and Vecchiano (d) are small towns (around 44K and 30K callers).

outperforms the differential privacy approach since the latter tends to transform any user profile into profiles representing residents.

As future work, we can try both to consider different attacks and to improve the ad hoc mitigation strategy. As an example, we could extend the background knowledge, and thus the attacks, by considering more than one municipality at the same time. Regarding the mitigation strategy, we could evaluate the nearest profiles with other distance functions, such as cosine similarity, gravity model or even a distance tailored to the service. Lastly, we could consider distance matrix approximation methods in order to reduce the impact on the computational time of the distance between groups.

9. Acknowledgements

This work has been supported by Project EU H2020-654024 SoBigData Infrastructure, under grant agreement No. 654024.

References

[1] F. Pratesi, A. Monreale, R. Trasarti, F. Giannotti, D. Pedreschi, T. Yanagihara, PRUDEnce: a System for Assessing Privacy Risk vs Utility in Data Sharing Ecosystems, Transactions on Data Privacy 11 (2018) 139 – 167.

[2] B. Furletti, L. Gabrielli, C. Renso, S. Rinzivillo, Analysis of gsm calls data for understanding user mobility behavior, in: IEEE Big Data, 2013.

[3] M. Nanni, R. Trasarti, B. Furletti, L. Gabrielli, P. V. D. Mede, J. D. Bruijn, E. de Romph, G. Bruil, MP4-A Project: Mobility planning for Africa, in: In D4D Challenge @ 3rd Conf. on the Analysis of Mobile Phone datasets (NetMob), 2013. URL: `http://perso.uclouvain.be/vincent.blondel/netmob/2013/D4D-book.pdf`.

[4] L. Bengtsson, X. Lu, A. Thorson, R. Garfield, J. von Schreeb, Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in haiti, PLOS Medicine 8 (2011) 1–9. URL: `https://doi.org/10.1371/journal.pmed.1001083`. doi:`10.1371/journal.pmed.1001083`.

[5] N. Oliver, A. Matic, E. Frias-Martinez, Mobile network data for public health: Opportunities and challenges, Frontiers in Public Health 3 (2015) 189.

[6] T. Tuoto, F. De Fausti, R. Radini, L. Valentino, M. Savarese, F. Fabbri, M. R. Spada, Challenges and opportunities with mobile phone data in official statistics, in: Conference of European Statistics Stakeholders, 2018.

[7] European Parliament & Council, General data protection regulation, in: L119, 4/5/2016.

[8] B. Furletti, L. Gabrielli, F. Giannotti, L. Milli, M. Nanni, D. Pedreschi, Use of mobile phone data to estimate mobility flows. measuring urban population and inter-city mobility using big data in an integrated approach, in: 47th SIS Scientific Meeting of the Italian Statistica Society, 2014.

[9] C. Dwork, Differential Privacy, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 1–12. doi:`10.1007/11787006_1`.

[10] P. Golle, K. Partridge, On the anonymity of home/work location pairs, in: International Conference on Pervasive Computing, Springer, 2009, pp. 390–397.

[11] H. Zang, J. Bolot, Anonymization of location data does not work: A large-scale measurement study, in: Proceedings of the 17th annual international conference on Mobile computing and networking, ACM, 2011, pp. 145–156.

[12] N. J. Croft, M. S. Olivier, Sequenced release of privacy accurate call data record information in a gsm forensic investigation., in: ISSA, 2006, pp. 1–14.

[13] Y. De Mulder, G. Danezis, L. Batina, B. Preneel, Identification via location-profiling in gsm networks, in: Proceedings of the 7th ACM workshop on Privacy in the electronic society, ACM, 2008, pp. 23–32.

[14] A. Noriega-Campero, A. Rutherford, O. Lederman, Y. A. de Montjoye, A. Pentland, Mapping the privacy-utility tradeoff in mobile phone data for development, arXiv preprint arXiv:1808.00160 (2018).

[15] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, V. D. Blondel, Unique in the crowd: The privacy bounds of human mobility, Scientific reports 3 (2013) 1376.

[16] Y.-A. de Montjoye, S. Gambs, V. Blondel, G. Canright, N. De Cordes, S. Deletaille, K. Engø-Monsen, M. Garcia-Herranz, J. Kendall, C. Kerry, et al., On the privacy-conscientious use of mobile phone data, Scientific data 5 (2018). doi:doi:10.1038/sdata.2018.286, published online 2018 Dec 11.

[17] S. Arfaoui, A. Belmekki, A. Mezrioui, Privacy increase on telecommunication processes, in: 2018 International Conference on Advanced Communication Technologies and Networking (CommNet), IEEE, 2018, pp. 1–10. doi:DOI:10.1109/COMMNET.2018.8360266.

[18] M. Nunez-del Prado, J. Nin, Revisiting online anonymization algorithms to ensure location privacy, Journal of Ambient Intelligence and Humanized Computing (2019) 1–12. doi:doi.org/10.1007/s12652-019-01371-6, accepted: 18 June 2019.

[19] V. D. Blondel, A. Decuyper, G. Krings, A survey of results on mobile phone datasets analysis, EPJ Data Science 4 (2015) 10.

[20] Y. Song, D. Dahlmeier, S. Bressan, Not so unique in the crowd: a simple and effective algorithm for anonymizing location data, in: Proceeding of the 1st International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security co-located with 37th Annual International ACM SIGIR conference, PIR@SIGIR 2014, Gold Coast, Australia, July 11, 2014., 2014, pp. 19–24. URL: http://ceur-ws.org/Vol-1225/pir2014_submission_11.pdf.

[21] S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, W. Willinger, Human mobility modeling at metropolitan scales, in: Proceedings of the 10th international conference on Mobile systems, applications, and services, Acm, 2012, pp. 239–252.

[22] A. Cavoukian, Privacy by design: The 7 foundational principles, Information and Privacy Commissioner of Ontario, Canada, 2009. https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf.

[23] A. Monreale, S. Rinzivillo, F. Pratesi, F. Giannotti, D. Pedreschi, Privacy-by-design in big data analytics and social mining, EPJ Data Science (2014). 2014:10.

[24] B. C. Fung, K. Wang, A. W.-C. Fu, S. Y. Philip, Introduction to privacy-preserving data publishing: Concepts and techniques, CRC Press, 2010.

[25] C. Dwork, Differential privacy: A survey of results, in: International Conference on Theory and Applications of Models of Computation, Springer, 2008, pp. 1–19.

[26] C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in: Theory of Cryptography Conference, 2006, pp. 265–284.

[27] F. McSherry, K. Talwar, Mechanism design via differential privacy, in: 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS), 2007, pp. 94–103.

[28] G. Acs, C. Castelluccia, A case study: Privacy preserving release of spatio-temporal density in Paris, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2014, pp. 1679–1688.

[29] D. J. Mir, S. Isaacman, R. Cáceres, M. Martonosi, R. N. Wright, Dp-where: Differentially private modeling of human mobility, in: Big Data, 2013 IEEE International Conference on, IEEE, 2013, pp. 580–588.

[30] G. Acs, G. Biczók, C. Castelluccia, Privacy-Preserving Release of Spatio-Temporal Density, Springer International Publishing, Cham, 2018, pp. 307–335. URL: https://doi.org/10.1007/978-3-319-98161-1_12. doi:10.1007/978-3-319-98161-1_12.

[31] P. Samarati, L. Sweeney, Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, Technical Report, SRI International, 1998.

[32] P. Samarati, L. Sweeney, Generalizing data to provide anonymity when disclosing information, in: Principles Of Database Systems (PODS), volume 98, 1998, p. 188.

[33] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, Mondrian multidimensional k-anonymity, in: Proceedings of the 22Nd International Conference on Data Engineering, ICDE '06, IEEE Computer Society, Washington, DC, USA, 2006, pp. 25–. URL: https://doi.org/10.1109/ICDE.2006.101. doi:10.1109/ICDE.2006.101.

[34] J.-W. Byun, A. Kamra, E. Bertino, N. Li, Efficient k-anonymization using clustering techniques, in: Proceedings of the 12th International Conference

on Database Systems for Advanced Applications, DASFAA'07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 188–200. URL: http://dl.acm.org/citation.cfm?id=1783823.1783848.

[35] J.-L. Lin, M.-C. Wei, An efficient clustering method for k-anonymization, in: Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society, PAIS '08, ACM, New York, NY, USA, 2008, pp. 46–50. URL: http://doi.acm.org/10.1145/1379287.1379297. doi:10.1145/1379287.1379297.

[36] B. Furletti, R. Trasarti, P. Cintia, L. Gabrielli, Discovering and understanding city events with big data: The case of Rome, Information 8 (2017) 74.

[37] A. Cavoukian, Privacy by design [leading edge], IEEE Technology and Society Magazine 31 (2012) 18–19.

[38] L. Sweeney, Achieving k-anonymity privacy protection using generalization and suppression, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10 (5) (2002) 571–588.

[39] R. Agrawal, R. Srikant, Privacy-preserving data mining, in: SIGMOD Conference, 2000, pp. 439–450.

[40] J. L. Bentley, Multidimensional binary search trees used for associative searching, Commun. ACM 18(9) (1975) 509–517.

[41] J. Domingo-Ferrer, V. Torra, Disclosure control methods and information loss for microdata, Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies (2001).

[42] M. Hay, V. Rastogi, G. Miklau, D. Suciu, Boosting the accuracy of differentially private histograms through consistency, Proceedings of the VLDB Endowment 3 (2010) 1021–1032.

[43] C. Dwork, F. McSherry, K. Nissim, A. Smith, Differential privacy — a primer for the perplexed, Joint UNECE/Eurostat work session on statistical data confidentiality 11 (2011).

[44] R. Chen, B. Fung, B. C. Desai, N. M. Sossou, Differentially private transit data publication: a case study on the montreal transportation system, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2012, pp. 213–221.

[45] D. Huang, S. Han, X. Li, Achieving accuracy guarantee for answering batch queries with differential privacy, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2015, pp. 305–316.

[46] F. McSherry, R. Mahajan, Differentially-private network trace analysis, in: ACM SIGCOMM Computer Communication Review, volume 40, ACM, 2010, pp. 123–134.

[47] G. Forman, Quantifying counts and costs via classification, Data Mining and Knowledge Discovery 17(2) (2008) 164–206. doi:https://doi.org/10.1007/s10618-008-0097-y.