




Privacy Preserving Multidimensional Profiling

Francesca Pratesi^{1,2}(✉) , Anna Monreale², Fosca Giannotti¹,
and Dino Pedreschi²

¹ ISTI-CNR A. Faedo, Pisa, Italy
francesca.pratesi@isti.cnr.it

² Department of Computer Science, University of Pisa, Pisa, Italy
<http://kdd.isti.cnr.it/homes/pratesi/>

Abstract. Recently, big data had become central in the analysis of human behavior and the development of innovative services. In particular, a new class of services is emerging, taking advantage of different sources of data, in order to consider the multiple aspects of human beings. Unfortunately, these data can lead to re-identification problems and other privacy leaks, as diffusely reported in both scientific literature and media. The risk is even more pressing if multiple sources of data are linked together since a potential adversary could know information related to each dataset. For this reason, it is necessary to evaluate accurately and mitigate the individual privacy risk before releasing personal data. In this paper, we propose a methodology for the first task, i.e., assessing privacy risk, in a multidimensional scenario, defining some possible privacy attacks and simulating them using real-world datasets.

Keywords: Privacy risk assessment · Mobile phone data · Retail data

1 Introduction

One of the most pressing challenges of our time is understanding the complexity of our globally interconnected society: the big data originating from the *digital breadcrumbs* of human activities let us observe the ground truth of individuals at an unprecedented detail. Indeed, companies and governments are using this ocean of Big Data to unleash powerful analytic capabilities, connecting data from different sources, finding patterns and generating new insights. This can help transform the lives of individuals and help to solve many of society's challenges [23]. Hence, we want to explore a multidimensional scenario, providing an example of off-line analyses taking advantage of different data sources. The data can be purchasing data, movement tracks, electronic payments, social media data, and so on. The benefits of analyzing multidimensional data can be various, and they can lead to convenience for users, data providers or third parties.

Therefore, having the possibility to collect data from different sources, we want to analyze the implications from the point of view of both possible services

and privacy risk for participants. This means to address the study of quality and privacy in multidimensional data, i.e., data from different kind of sources. Clearly, each of these sources can lead to privacy leaks by itself, but if an attacker, who gain access to the datasets, has information about more aspects of her target’s life (e.g., she knows some visited locations *and* some bought products), she can gain better chances to recognize her among all the users. The aim of this paper is to provide an example of a real-world privacy-preserving multidimensional service and to show what are the advantages and threats related to the use of such kind of data. Besides there are several strategies for the privacy risk mitigation, we illustrate the results using only the simplest solution, i.e., the suppression of risky individuals, showing that it is possible to obtain good results if we use the appropriate level of detail of the data.

The rest of the paper is organized as follows. In Sect. 2, we review some related work. In Sect. 3, we report the preliminaries of our work. In Sect. 4, we illustrate the service, the minimum data formats required for implementing it and the attacks we consider. In Sect. 5, we simulate the attacks and the results obtained at each step of the service. Section 6 concludes the paper.

2 Related Work

In recent years, privacy has been one of the most discussed issues. The aim of the methods proposed in the literature is assuring the privacy protection of individuals during both the analysis and the publishing of human data. The privacy has been studied in several contexts, from location based services [22] to GPS trajectories [1], from mobile phone data [2, 20, 24] to retail data [6, 9, 17]. In particular, our focus is on mobile phone and retail data. Privacy risks in mobile phone data, even in the case of releasing information with not fine granularity, are studied in [24]. In particular, authors consider the top N locations visited by each user. Typical solutions, highlighted by Blondel et al. [2] are to operate small modification of datasets or to change frequently (at least daily) pseudo-identifiers. Unfortunately, this can lead big limitation on analyses and services that can be performed. In [8], it is suggested to use synthetic data, which can reproduce many features of mobility of users. An extension of this work that relies on Differential Privacy can be found in [11]. Retail data are analyzed for maximizing the profit of companies [13] and for potentiating the customer care [18], but they might reveal personal or sensible information about the subject (for example, if he discovers that the subject regularly buys gluten-free pasta, it is easy to infer that the user suffers from celiac disease). To prevent these issues, researchers have developed privacy preserving methodologies, in particular, to extract association rules from retail data [6, 9, 17].

In the last year, different techniques for risk management¹ has been proposed, such as the OWASP’s [12] and LINDDUN [4] methodologies, Microsoft’s DREAD [10], SEI’s OCTAVE [7]. Unfortunately, many of them do not consider in deep

¹ The risk evaluation task is compliant with the EU General Data Protection Regulation.

privacy. A methodology that presents an approach for the systematic privacy evaluation is the one presented in [16] and extended in [14, 15]. This represents a starting point for our work and it will be outlined in Sect. 3.

3 Privacy Risk Assessment Framework

In this paper, we consider the work proposed in [16], which allows for the privacy risk assessment of human data, considering a scenario where a Service Developer asks a Data Provider for data to develop an analytical service. The Data Provider must guarantee the right to privacy of the individuals whose data are recorded. In a nutshell, the framework PRISQUIT is based on the Privacy by Design paradigm, so the first specification is to establish the data requirements for the service. Then the Data Provider queries its dataset, producing a set of datasets with different data structures and aggregations and it then: (i) identifies the background knowledge that an adversary might have about her target; (ii) simulates the attack based on that background knowledge, computing the privacy risk values for every individual; (iii) selects the dataset with the best privacy-utility trade-off; (iv) applies a privacy risk mitigation method (e.g., generalization, randomization, suppression) on that dataset; and (v) delivers the sanitized dataset to a third party.

In [16], and consequently in this paper, it has been used as privacy risk the risk re-identification [19], whose 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 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 visited a certain location on a certain day) and B_h to indicate a set of background knowledge of size h (e.g., B_2 can represent all the possible couple of locations visited by an individual).

Let \mathcal{D} be a database, D a dataset derived from \mathcal{D} (e.g., an aggregated data structure on time and/or space), 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 [16]). *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 background knowledge b .*

Note that $PR_D(d=u|b) = 0$ if the user u is not in D . Since each background knowledge b has its own probability of re-identification, we define the risk of re-identification of an individual as the maximum probability of re-identification over the set of possible background knowledge:

Definition 2 (Privacy risk [16]). The risk of re-identification (or privacy risk) of an individual u given a set of background knowledge B_k is her maximum probability of re-identification $Risk(u, D) = \max PR_D(d = u|b)$ for $b \in B_k$. It has the lower bound $\frac{|D_u|}{|D|}$ (a random choice in D), and $Risk(u, D) = 0$ if $u \notin D$.

An individual is hence associated with several privacy risks, each for every background knowledge of an attack.

4 Privacy Preserving Multidimensional Profiling

In [16], we have a scenario where a single Data Provider (DP) and a single Service Developer (SD) interact, using PRISQUIT, to determine the best dataset to be released, regarding the trade-off between privacy and utility. In this paper, instead, we have an ecosystem where several DPs interact with an SD, through a Safe and Trusted Environment (STE) for personal data collection and sharing, which provides a link between DPs and SD and assisting DPs in the use of PRISQUIT. Indeed, the SD asks to the STE for a series of data, and they define together the minimum data format suitable to provide a reliable service. The STE address the study of quality and privacy in multidimensional data, i.e., data from different kind of sources. In order to have a clear view of both the actual privacy risk and the potentiality of services with combined datasets, we should consider that users appear in each source. Only by relying on this basis we can provide a rigorously correct quantification. However, we can analyze a what-if scenario, where users belonging to different datasets are linked to each other, pretending to be the same individual. In the following, we outline a possible service that needs multidimensional sources, analyzing the minimum data format necessary and the attacks that is possible to perform on every single dataset separately or exploiting the presence of same users in more datasets.

4.1 Promotion Service Based on Recurrent Events

We suppose that a marketing manager of a retail chain wants to suggest products related to a recurrent event, such as an annual movie festival or a monthly organic market, to her customers. The marketing expert decides to target the campaign only to customers that actually participated in the event in the past since they will be more likely to respond positively to the advertisements.

In order to implement this service, we need to have access to a dataset from a telco operator to obtain the participants and one from a retail chain to verify their purchases. Indeed, the service is composed of the following steps.

Event Detection and Users Participation. The SD asks to the STE for mobile phone data. Then, the SD can detect relevant peaks representing an event, comparing the density of population within a region in a given moment against the expected density for that area [21]. The SD can also discover the users that participated in that event, discriminating between actual visitors and regular inhabitants [5].

Marketing Campaign Definition. Then, the SD can ask for the purchases, related to a specific time window before the event, of the participants. The size of the time window is strictly related to the event: it could be a day or a week time window, or even a monthly one if the event needs some preparation, like Halloween or Christmas. The SD defines what the products related to the event are, and he checks if users bought those products. If not, the SD can contact the retail chain marketing responsible (through the STE, which is responsible for the match of users among the different data sources), suggesting how to spread a targeted campaign before the following occurrence of the event.

Minimum Data Formats

Since we rely on the data minimization principle² and on the Privacy-by-Design model [3], we define the minimum data format necessary to develop this service.

Minimum Data Format (Mobile Phone Data). We want to obtain a density map of a territory, so we only need an aggregation like the one presented in [5]. Therefore, the data format is a profile, i.e., a matrix P where i denotes the time slot (e.g., morning, afternoon, night) and j denotes the day (if the event is a daily one, we will remove the time slots). Since we do not need the precise activities of users, the profile contains only 1 (indicating a presence) and 0 (absence).

$$P_{ij}^u = \begin{cases} 0 & \text{if } u \text{ is not present or he performs no calls at day } j \text{ and slot } i \\ 1 & \text{if } u \text{ performs at least one call in the area at day } j \text{ and slot } i \end{cases} \quad (1)$$

Minimum Data Format (Purchasing Data). Here, the data format is the set of distinct items bought in a determined time window:

$$\{i_1, i_2, \dots, i_n\} \quad (2)$$

Clearly, considering a product at the detail level of the bar code is usually unnecessary, so we can climb the product taxonomy and establish the essential level of detail, like categories or subcategories, suitable for the service realization.

Attack Models

We point out that each attack assumes the adversary gains access to the dataset. Performing an attack means finding a set C of possible matches for a target, given a certain background knowledge. The probability of re-identification of the user u is $\frac{1}{|C|}$. A greater number of candidates implies a better privacy protection.

Background Knowledge on Mobile Phone Data. We suppose that the attacker knows part of the data format (1), e.g., she knows that her target was in Florence on Monday and Friday of a specific week. Thus, the adversary can build a partial (but exact) profile b , where $b_{ij} = -1$ if the attacker does not have any

² Art. 5 EU GDPR.

information about the period (i, j) . We suppose that h represents the number of weeks in which the attacker knows information about the calls of her target, so B_h corresponds to the possible combination of weeks of the profile P .

Attack on Mobile Phone Data. The attacker uses the background knowledge b on the user u to match all the profiles that include b . The set of matched profiles is $C = \{P \in \mathcal{P} | \forall b_{ij} \geq 0. b_{ij} = P_{ij}\}$.

Background Knowledge on Retail Data. We suppose that the attacker has as background knowledge a subset of products bought by her target, in the format (2); for example, the attacker once saw the shopping chart of her target. The subset of items known has size q , so we have as background knowledge $b \in B_q$.

Attack on Retail Data. The attacker uses the background knowledge b on the user u to match all the set of items that include b . Given $D(u_i)$ the set of items of the user $u_i \in \mathcal{D}$, the candidate set is computed as $C = \{u_i | b \subseteq D(u_i)\}$.

In the following, we define a possible multidimensional background knowledge and the correspondent attack. Note that, in order to execute this attack, it is necessary to have a link between users: for example, we could know that the caller 100 in the mobile phone dataset is the customer 30 in the retail dataset.

Background Knowledge on Mobile Phone and Retail Data. Here, we suppose that the attacker has as background knowledge a subset of call activities and products bought by her target, i.e., a combination of the two previous attacks. The subsets of items and call activities may have variable size q and h , respectively.

Attack on Mobile Phone and Retail Data. Given the sizes q and h , we denote by B_q and B_h the two set of background knowledge. Let $D_p(u_i)$ the set of items of the user u_i and $D_m(u_i)$ the weeks of call activities of u_i in the format (2) and (1), respectively. For each instance b' of B_q and b'' of B_h , the candidate set is computed as: $C = \{u_i | b' \subseteq D_p(u_i) \wedge b'' \subseteq D_m(u_i)\}$ The probability of re-identification given the background knowledge $b = \{b' \cup b''\}$ is $\frac{1}{c}$.

5 Experiments

Here, we present the simulation of the attacks presented in Sect. 4. Firstly, we illustrate the datasets; then, we provide the results of the simulation of privacy risk and an indication of the output service-side, i.e., from the SD's point of view. We provide the simulations for all the attacks; however, since we do not have the actual link between users in the two datasets, we report a what-if analysis for the multidimensional attack, imagining that the link between callers and customers is known by the STE. With this aim, we randomly selected customers, and we forcibly assigned them to some of the callers (in particular, we assumed that STE knows the purchases of one-third of the mobile phone population).

5.1 Datasets Presentation

The mobile phone dataset is provided by one of the major Italian mobile operators. It regards the territory of a great part of Tuscany (106 municipalities out

of 276), for a period from February 17, 2014, to March 23, 2014, and it reports the activities of around 858k individuals, for a total of 51 million call records. The retail dataset is provided by Unicoop Tirreno, one of the major retail distribution companies in Italy. It regards a time window spanning from January 1, 2007, to June 30, 2014. The active and recognizable customers, i.e., users with a loyalty card and with at least a purchase in the period, are about 800,000.

5.2 Privacy Analysis

As illustrated in Sect. 4, we start from mobile phone data. The SD asks for this kind of data among the Tuscan municipalities, in order to detect events. The STE builds the profiles aggregated in days (i.e., from midnight to midnight) of all the people present in each area. At a certain point, the STE analyzes the Viareggio municipality, i.e., it simulates the attack on mobile phone data presented in Sect. 4.1, obtaining the cumulative distribution of probabilities of the privacy risk shown in Fig. 1(a). Here, we can find the percentages of users having at most a certain risk of re-identification, obtained varying the background knowledge from 1 to 4 weeks. As we can see, due to the aggregated nature of the data, if we hypothesize that the attacker knows 1 week of calls of her target, the probability she succeeds in re-identification is extremely low, since for 95% of users the privacy risk is below 0.005 (i.e., they are indistinguishable from at least 199 others). As we can expect, if we increase the number of weeks known by the attacker, the risk increases too. However, it is never dramatically high: knowing 2 weeks, we have that only 10% of users have a risk greater than 0.2, while knowing 3 weeks this risk is associated with 35% of individuals. It is interesting to note the “knee” in the curves: its presence indicates that until that point it is possible to obtain a lower risk renouncing to relatively few users.

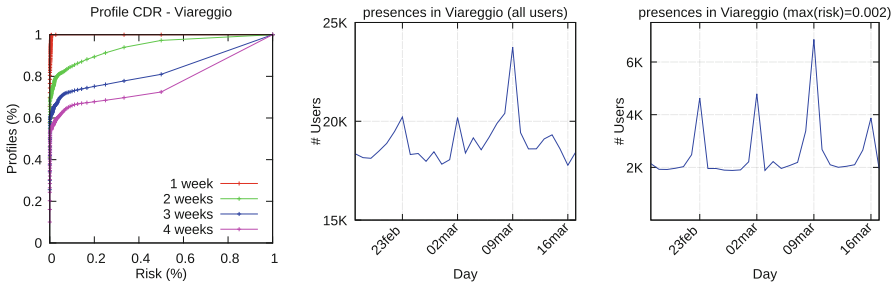


Fig. 1. (a) CDF of the privacy risk, varying the background knowledge. (b–c) The timeseries of presences in Viareggio, using mobile phone data: (b) timeseries computed with all the original users; (c) timeseries computed with only safe users, fixing the maximum risk of re-identification to 0.002

Since we are mainly interested in visitors, i.e., exceptional presences w.r.t. the routine, we can imagine that they are individuals present only for the event.

So, they are less problematic concerning privacy w.r.t. commuters and residents, since, likely, they are naturally similar to other visitors. For this reason, we can decide to use a quite extreme privacy threshold, expecting more or less the same efficacy in terms of quality. In particular, we chose to guarantee a maximum risk of re-identification among users of 0.002 (that corresponds to a group size of at least 500), releasing 48% of profiles (i.e., 41K users out 86K).

As soon as the SD receives these profiles, it performs a peak detection, analyzing the presences, day by day, in the considered month. In order to provide a fair comparison of the utility, we show in Fig. 1(b) the results of the peak detection analyzing the profiles of all the 86 K users (so, *without any regards for privacy*) and in Fig. 1(c) the same analysis applied only to released data (i.e., the profiles of users with the maximum risk of re-identification equal to 0.002). Even if the scale between the two plots differs from almost an order of magnitude, we can clearly recognize peaks on Sundays. In particular, Fig. 1(b) and (c) have an anomalous peak on Sunday, March 9, so the SD presumes that something different occurred that day. Indeed, on that Sunday there was a popular Carnival Parade. Using data depicted in Fig. 1(c), the SD can also discriminate between people that regularly live the area (i.e., residents and commuters), who represent the baseline of about 2,000 persons, and real visitors, who define the peak. The SD is interested in asking for the expenses of these 3,500 users (the peak is around 5,000 users higher than the average, but we can assume, for the sake of simplicity, that 1,500 individuals visited Viareggio every Sundays, so we remove them from the count). At this point, we rely on the hypothesis that the STE can have access to the purchases of one-third of these individuals, so we randomly select 1,000 customers from the retail dataset.

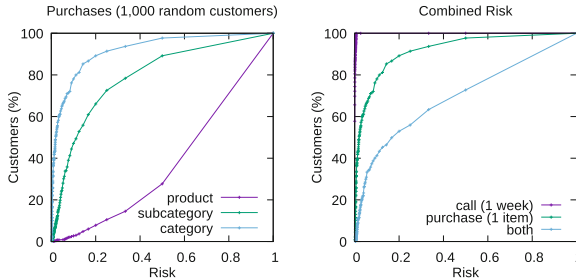


Fig. 2. Simulation of the attack on retail data varying the level of detail of items (Left) and of the attack on call activity and retail data using a random matching between callers and customers (Right)

The second step of the analysis is assessing the privacy risk of these 1,000 customers. STE analyzes the purchases in a time window covering the period in which the supermarket chain sells products related to Carnival, i.e., from January 1 to March 9, 2014. The risk of re-identification of these users, fixing the number of known products $h = 1$, is reported in Fig. 2 (Left). From this

plot, it is immediate to see that using a detailed level leads to a too high risk. However, if we use less detailed information, we obtain better results. In order to clarify the taxonomy, we provide a simple example describing the three levels of detail: we have a specific milk (e.g., “500 ml UHT skim milk of brand Coop”) at the *product* level; as *subcategory* we have “UHT” or “fresh” milk; at the *category* level we only know that the item is “milk”. In our case, at the category level, we can still search for Carnival related items. Thus, releasing the categories, we have that 90% of users have 0.25 as maximum risk, while 78% have 0.1 and 65% have 0.05. The STE can decide to release the purchases of the customers having a maximum risk of 0.05, i.e., who are in groups of at least 20 indistinguishable users. The released customers are 655. The SD analyzes their purchases and discovers that only 37 of them bought in Unicoop stores products related to the event. This means that SD can inform Unicoop that, among its customers, 618 individuals very likely participated in Viareggio Carnival in 2014, but they did not buy anything in the month preceding the event; maybe with focused offers, they will be inclined to buy in Unicoop stores before the following Carnival event.

Lastly, if we consider the multidimensional background knowledge attack on mobile phone data and retail data, which is based on a very strong knowledge, we obtain the simulation depicted in Fig. 2 (Right). Here, we can see that, even if knowing only one aspect of the users’ life does not compromise the possibility to release a portion of data without high privacy concerns, when we combine the two dimensions and the certainty to know who is each individual in the two datasets, the risk of re-identification is substantially higher. Indeed, if we want to provide the same maximum risk level we used before, i.e., 0.05, we can release only 299 individuals’ data.

6 Conclusion

In this paper, we envision a data sharing scenario, where a trusted component called Safe and Trust Environment (STE) offers an interface between different Data Providers (DP) and potential Service Developers (SD). The STE receives queries by SDs and asks for the correspondent data to specific DPs, defining the minimum data format necessary and analyzing the privacy risk of each individual whose data belong, *before releasing the data*. We defined some attacks related to two kinds of data, in order to analyze the risk of sharing data with some background knowledge about them both separately and together. We simulated these attacks on real-world datasets, discovering that using the minimum information required it is possible to achieve a good trade-off between privacy protection and quality of service. However, combining little information from different sources can lead to risk sensibly higher than a larger amount of knowledge related only to one source. As future work, we would analyze different services, including other kinds of data, like GPS tracks or credit card logs, in order to provide realistic examples of privacy-preserving multidimensional data sharing.

Acknowledgment. Funded by the European project SoBigData (Grant Agreement 654024).

References

1. Abul, O., Bonchi, F., Nanni, M.: Anonymization of moving objects databases by clustering and perturbation. *Inf. Syst.* **35**(8) (2010)
2. Blondel, V.D., Decuyper, A., Krings, G.: A survey of results on mobile phone datasets analysis. *EPJ Data Sci.* **4**(1), 10 (2015)
3. Cavoukian, A.: Privacy by design the 7 foundational principles, August 2009
4. Deng, M., Wuyts, K., Scandariato, R., Preneel, B., Joosen, W.: A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. *Requir. Eng.* **16**(1) (2011)
5. Gabrielli, L., Furletti, B., Trasarti, R., Giannotti, F., Pedreschi, D.: City users' classification with mobile phone data. In: *IEEE Big Data 2015* (2015)
6. Giannotti, F., Lakshmanan, L.V., Monreale, A., Pedreschi, D., Wang, H.: Privacy-preserving mining of association rules from outsourced transaction databases. *IEEE Syst. J.* **7** (2013)
7. Institute, C.S.E.: Octave. <http://www.cert.org/octave/>
8. Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., Willinger, W.: Human mobility modeling at metropolitan scales. In: *MobiSys 2012* (2012)
9. Le, H.Q., Arch-int, S., Nguyen, H.X., Arch-int, N.: Association rule hiding in risk management for retail supply chain collaboration. *Comput. Indus.* **64** (2013)
10. Meier, J., Corporation, M.: *Improving Web Application Security: Threats and Countermeasures*. In: *Patterns & Practices*, Microsoft (2003)
11. Mir, D.J., Isaacman, S., Cáceres, R., Martonosi, M., Wright, R.N.: Dp-where: differentially private modeling of human mobility. In: *IEEE Big Data 2013* (2013)
12. OWASP: Risk rating methodology. https://www.owasp.org/index.php/OWASP-Risk_Rating_Methodology
13. Pauler, G., Dick, A.: Maximizing profit of a food retailing chain by targeting and promoting valuable customers using loyalty card and scanner data. *EJOR* **174** (2006)
14. Pellungrini, R., Pappalardo, L., Pratesi, F., Monreale, A.: A data mining approach to assess privacy risk in human mobility data, ready to appear in *ACM TIST*
15. Pellungrini, R., Pratesi, F., Pappalardo, L.: Assessing privacy risk in retail data. In: *PAP@ECML-PKDD 2017* (2017)
16. Pratesi, F., Monreale, A., Trasarti, R., Giannotti, F., Pedreschi, D., Yanagihara, T.: Prisquit: a system for assessing privacy risk versus quality in data sharing, Technical report 2016-TR-043. ISTI - CNR, Pisa, Italy (2016)
17. Rizvi, S.J., Haritsa, J.R.: Maintaining data privacy in association rule mining. In: *VLDB 2002* (2002)
18. Rygielski, C., Wang, J.C., Yen, D.C.: Data mining techniques for customer relationship management. *Technol. Soc.* **24** (2002)
19. Samarati, P., Sweeney, L.: Generalizing data to provide anonymity when disclosing information (abstract). In: *PODS*, p. 188 (1998)
20. Song, Y., Dahlmeier, D., Bressan, S.: Not so unique in the crowd: a simple and effective algorithm for anonymizing location data. In: *PIR@SIGIR 2014* (2014)
21. Trasarti, R., Olteanu-Raimond, A.M., Nanni, M., Couronné, T., Furletti, B., Giannotti, F., Smoreda, Z., Ziemlicki, C.: Discovering urban and country dynamics from mobile phone data with spatial correlation patterns. *Telecommun. Policy* **39**(3–4) (2015)

22. Xiao, Y., Xiong, L.: Protecting locations with differential privacy under temporal correlations. In: ACM CCS 2015 (2015)
23. World Economic Forum: Rethinking personal data: Strengthening trust. http://www3.weforum.org/docs/WEF_IT_RethinkingPersonalData_Report_2012.pdf
24. Zang, H., Bolot, J.: Anonymization of location data does not work: a large-scale measurement study. In: MobiCom. ACM (2011)