J Ambient Intell Human Comput (2016) 7:83–96
DOI 10.1007/s12652-015-0303-x

ORIGINAL RESEARCH

Re-identification and information fusion between anonymized CDR and social network data

Alket Cecaj[1] · **Marco Mamei**[1] · **Franco Zambonelli**[1]

Received: 14 January 2015 / Accepted: 29 June 2015 / Published online: 14 July 2015

Abstract The analysis of multiple datasets on users' behaviors opens interesting information fusion possibilities and, at the same time, creates a potential for re-identification and de-anonymization of users' data. On the one hand, this kind of approaches can breach users' privacy despite anonymization. On the other hand, combining different datasets is a key enabler for advanced context-awareness in that information from multiple sources can complement and enrich each other. In this work we analyze different anonymized mobility datasets in the direction of highlighting re-identification and information fusion possibilities. In particular we focus on call detail record (CDR) datasets released by mobile telecom operators and datasets comprising geo-localized messages released by social network sites. Results shows that: (1) in line with previous findings, few (about 4) data points are enough to uniquely pin point the majority (90 %) of the users, (2) more than 20 % of CDR users have a single social network user exhibiting a number of matching data points. We speculate that these two users might be the same person. (3) We derive an estimate of the probability of two users begin the same person given the number of data points they have in common, and estimate that for 3 % of the social network users we can find a CDR user very likely (>90 % probability) to be the same person.

✉ Alket Cecaj
alket.cecaj@unimore.it

Marco Mamei
marco.mamei@unimore.it

Franco Zambonelli
franco.zambonelli@unimore.it

[1] University of Modena and Reggio Emilia, Reggio Emilia, Italy

Keywords Mobility patterns · De-anonymization · Information fusion

1 Introduction

Mobile phones track the activities of their users along many dimensions. For example, telecom operators can access the location of their subscribers at a fine grain scale, e.g., via the use of call detail record (CDR) data. Similarly, the vast majority of applications installed on smartphones (especially those related to social network sites) collects and stores on the Web information about users locations and activities (e.g., by means of geo-localized pictures, or geo-localized messages and tweets).

As location and mobility are primary sources of context-information for several applications, the use of this kind of data is important and can have a strong impact in the fulfillment of the pervasive-anticipatory computing vision as described in Pejovic and Musolesi (2015). Furthermore, in developing countries, there seems to be a direct connection between mobile telephone services penetration and economic development as location and context based services can improve productivity in many sectors of economy as research shows in Abraham (2006) and Blondel et al. (2013).

In this scenario, an important but problematic activity consists in joining datasets by matching different users associated to the same real person. For example, it would be interesting to realize that user X in a CDR dataset is actually the same person as Twitter user Y, and then join the two datasets. Geo-located data presents a natural framework to evaluate users' similarity and enable re-identification (join) across multiple datasets. The matching process across multiple datasets is in fact rather straightforward in principle and consists in identifying whether

CDR user *X* and Twitter user *Y* consistently produce data at the same time and place. Once enough geo-referenced elements overlap, we can be reasonably sure that the two users are actually the same person.

On the one hand, this could raise serious privacy issues, as relations between different types of data can be used to infer information of any kind from socio-economic status, to mobility and shopping patterns, to the user's social graph as illustrated in Wicker (2012). This is particularly problematic once the process of matching users among data sources allows to bypass the anonymization of a given dataset. As work in Wicker (2012) describes it may in fact happen that: "The continued accumulation of location data may reach a point where a marketer can uniquely match an anonymous location trace to a named record in a separate database".

On the other hand—for the same reason—joining different datasets is the key for advanced forms of context awareness that could notably improve pervasive applications and services. In fact on the basis of such a combined dataset, it would be possible to infer what the users were doing in a given location and their general profile.

The contribution of this paper is to conduct analysis and experiments in this direction. In particular:

1. We analyse the uniqueness of mobility traces as a first essential step in order to quantify the re-identification potential. Uniqueness analysis, pioneered in Montjoye et al. (2013), consists in evaluating how many (geo-located) data points are necessary to uniquely identify a user in a dataset.
2. The next step is the re-identification method itself. Specifically, we try to answer the following question: *Can we use data from geo-referenced social networks, to re-identify mobile users from an anonymized call description records dataset?* We provide an answer to this question by using a probabilistic approach, that evaluates the probability that users from multiple datasets are actually the same person.

The content of this article is organized as follows. Section 2 presents researches at the state of the art in entity matching among multiple data sources. Section 3 illustrates the CDR and social network datasets we used for our analysis. In Sect. 4 we explain the uniqueness evaluation analysis performed on mobility traces. Section 5 presents initial re-identification results based on counting the number of matches among events generated by users across the two datasets and some ground truth evidence. Section 6 presents our probabilistic model to assess whether different users are actually the same person and provides experiments in this direction. Finally in Sect. 7 we present our conclusions.

2 Related work

As large-scale mobility and social network data is progressively available to researchers, there is a considerable amount of works on data re-identification as a mean to threaten users' privacy. The vast majority of works deal with the problem from the data uniqueness perspective: *what is the subset of data about someone to make him/her unique and thus re-identifiable among all the other users?*

In Golle (2006), for example, authors analyze census data discovering that the disclosing of gender, ZIP and full date of birth allows for unique identification of 63 % of individuals of the US population. Many studies explore the re-identification of datasets, such as movie ratings as in Narayanan and Shmatikov (2008) or Massachusetts Hospital medical records using publicly available side information. Another interesting case is the re-identification of anonymous volunteers in a DNA study for the Personal Genome Project in Sweeney et al. (2013). In Rossi and Musolesi (2014) are presented trajectory-based and frequency-based (frequency of visit to specific location) tecniques that aim to re-identify users in location based social networks.

More in line with our domain, in Montjoye et al. (2013) authors analyse a large CDR dataset discovering that 4 CDR events are enough to uniquely characterize the traces of mobility of 95 % of the users, whereas 2 CDR events can characterize up to 50 % of the users. A similar work using Markov chains models for the de-anonymization of geolocated data on the basis of the visited points-of-interest and similarity distance metrics can be found in Gambs et al. (2013). In Sharad and Danezis (2013) the authors attempt to de-anonymize the communication subgraph of a CDR dataset, using a social networking analysis approach. Finally, authors in Zang and Bolot (2011) describe how the anonymization tecniques for large datasets can be ineffective.

The main difference between these works and ours is that we do not focus only on uniqueness of data, but we try to evaluate the actual re-identification possibility of a user across multiple datasets. For example, even if it is possible to uniquely pinpoint a single user on the basis of only four CDR events as described in Montjoye et al. (2013), it is not granted that those four points can be identified in another set of data (e.g., geo-referenced Twitter). Should the user always use WiFi to connect to Twitter, those four points would be never found, hampering re-identification. We present a more detailed description of such aspects in Sect. 6.

From a complementary perspective, another stream of works deals with the problem of guaranteeing k-anonymity in spatial databases. The idea, is that a system provides k-anonymity protection if the information for each person contained in a data release cannot be distinguished from at

least k − 1 individuals also appearing in the data release as defined in Sweeney (2002). For example, techniques for k-anonymity in the case of mobility data are presented in Abul et al. (2010). Similarly, in Parent et al. (2013) authors consider two approaches: one based on anonymizing each trajectory as a whole, and another which considers attackers who can link specific pairs of location and time to a person and re-identify him/her using a certain number of points. In this latter case—which is also our set up—the goal of anonymization is to hamper the attacker from associating a sequence of points to less than k individuals. This is done in Parent et al. (2013) by removing those points from a person's trajectories that would allow to single him/her out. These approaches however, often tend to neglect the fact that most solutions for privacy preservation such as data suppression (removing critical data) or generalization (provide clusters of data rather than individual instances) highly compromise or destroy the utility of the dataset itself. Some approaches extending k-anonymity concepts are presented in Dwork (2011) and Zang and Bolot (2011).

In this work we focus on measuring the re-identification potential in real datasets and, other than evaluating privacy risks, we also emphasize the advantages of such an approach for information fusion aspects.

3 Dataset

In this section, we give a detailed description of the datasets used in this paper and considerations we made about their most interesting characteristics. The data we use for our experiments are anonymized CDR data of mobile users and the publicly available geo-referenced data from Twitter and Flickr during the same period of time and in the same area. For our analysis each record in every mentioned dataset is considered as an *event*.

3.1 CDR data

We got access to two CDR datasets describing mobility traces of a large user population over an extended period of time.

The first dataset (referred to as CDR-DATA1) has a time span of one month. It comprises records of each cell network event whether associated to incoming/outgoing calls, internet connections and text messages made by a mobile device, its timestamp and the geographic coordinates of the cell tower handling the event. This latter information is extracted from a table containing the coordinates of each cell tower and approximate area of coverage. In Fig. 1a—it is shown the structure of a a CDR record. Each record comprises a user (hashed) id, the mobile country code (MCC), the timestamp of the event, the code of the cell tower and the coordinates and coverage radius of the cell tower. Thus, the spatial resolution of CDR localization is the cell radius. Figure 2a illustrates the distribution of CDR-DATA1 events per user. Figure 2b illustrates the radius of gyration for a given percentile of users. The radius of gyration is a synthetic and easy-to-compute parameter describing the spatial extent of user traces. It is defined as the deviation of user positions from the corresponding centroid position. It is given by: $r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - p_{centroid})^2}$ where p_i represents the ith position recorded for the user and $p_{centriod}$ is the center of mass of the users recorded displacements obtained by: $p_{centroid} = \frac{1}{n}\sum_{i=1}^{n}(p_i)$. It is possible to see that the first quartile tends to be associated to sedentary people with $r_g < 5$ km. The bulk of the distribution (25th–75th) percentiles can be associated to urban mobile people as the diameter of peri-urban areas of main cities in the region is about 15 km. Users beyond the 75th percentile are associated to commuters spanning on the wide region area.

The second dataset (referred to as CDR-DATA2) records have the same characteristics as those of CDR-DATA1 and are shown in Fig. 1b. The only difference is that there are not given the radiuses of the cells and the mobile country code of the user. This dataset is not that big compared to CDR-DATA1 as the number of mobile users is smaller. The distribution of the events generated during a period of 2 weeks is given in Fig. 2c as events per user. The radius of gyration is represented in Fig. 2d. With regard to the radius of gyration it is possible to see that some users travel for

Fig. 1 Records for each dataset. **a** CDR-DATA1 dataset, **b** CDR-DATA2 dataset, **c** FT dataset

(a)

User id	MMC	Timestamp	Tower id	Coord	Radius
3dd285b	222	734628648723	123	(41.28,13.92)	450

(b)

User id	Timestamp	Tower id	Coord
12d285b	734628648723	123	(41.38,13.22)

(c)

User id	Pic/Tweet	Timestamp	Coord
12d285b	Text	734628648723	(41.28,13.92)

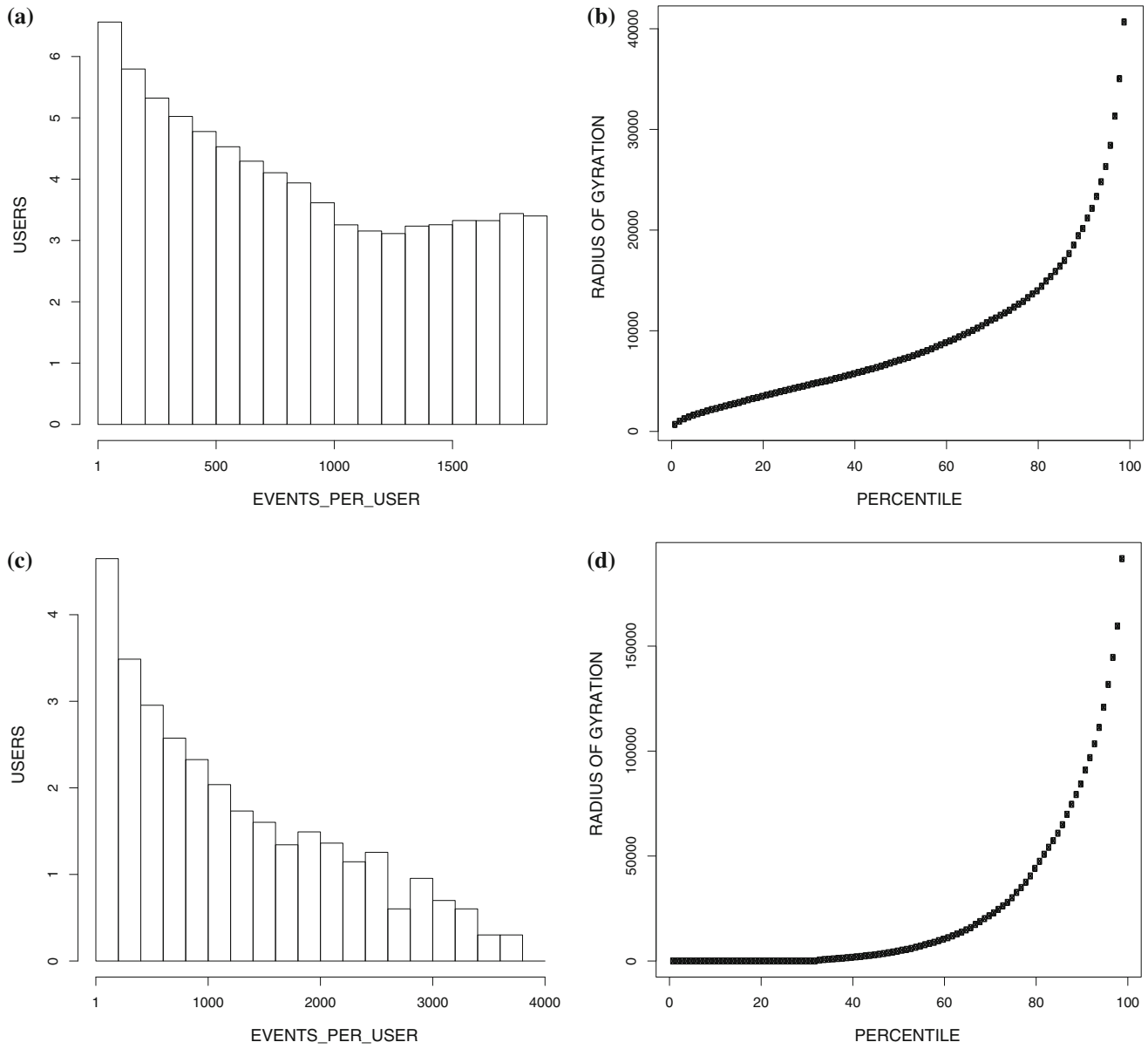

Fig. 2 **a** Events per user in CDR-DATA1 dataset. **b** Radius of gyration for a given percentile of users in CDR-DATA1. **c** Events per user in CDR-DATA2 dataset. **d** Radius of gyration for a given percentile of users in CDR-DATA2 dataset

longer distances than in the CDR-DATA1 dataset. This is because the CDR-DATA2 data spans in a much bigger area than CDR-DATA1.

3.2 Social network data

Social network data has been captured by using the REST API services provided by Flickr (http://www.flickr.com) and Twitter (http://www.twitter.com). Overall we refer to this dataset as FT data. In Fig. 1c is given how the records of this dataset are structured. Starting from a pool of 2456 users and having discarded from this dataset users with only one or two events, we consider for our analysis 965 (231 Flickr users and 739 Twitter ones) of them. We discarded users with less than three events as those provide very little information on their location history. At the end, there is an average number of 21 events per user and a max/min value respectively of 686 and 3 events per user. Fig. 3a, b shows the distribution and the radius of gyration of Twitter users. Figure 3c, d shows the distribution and the radius of gyration of Flickr users. The radius of gyration in both Twitter and Flickr is rather low, since we extracted information only from a limited bounding box around the area of interest.

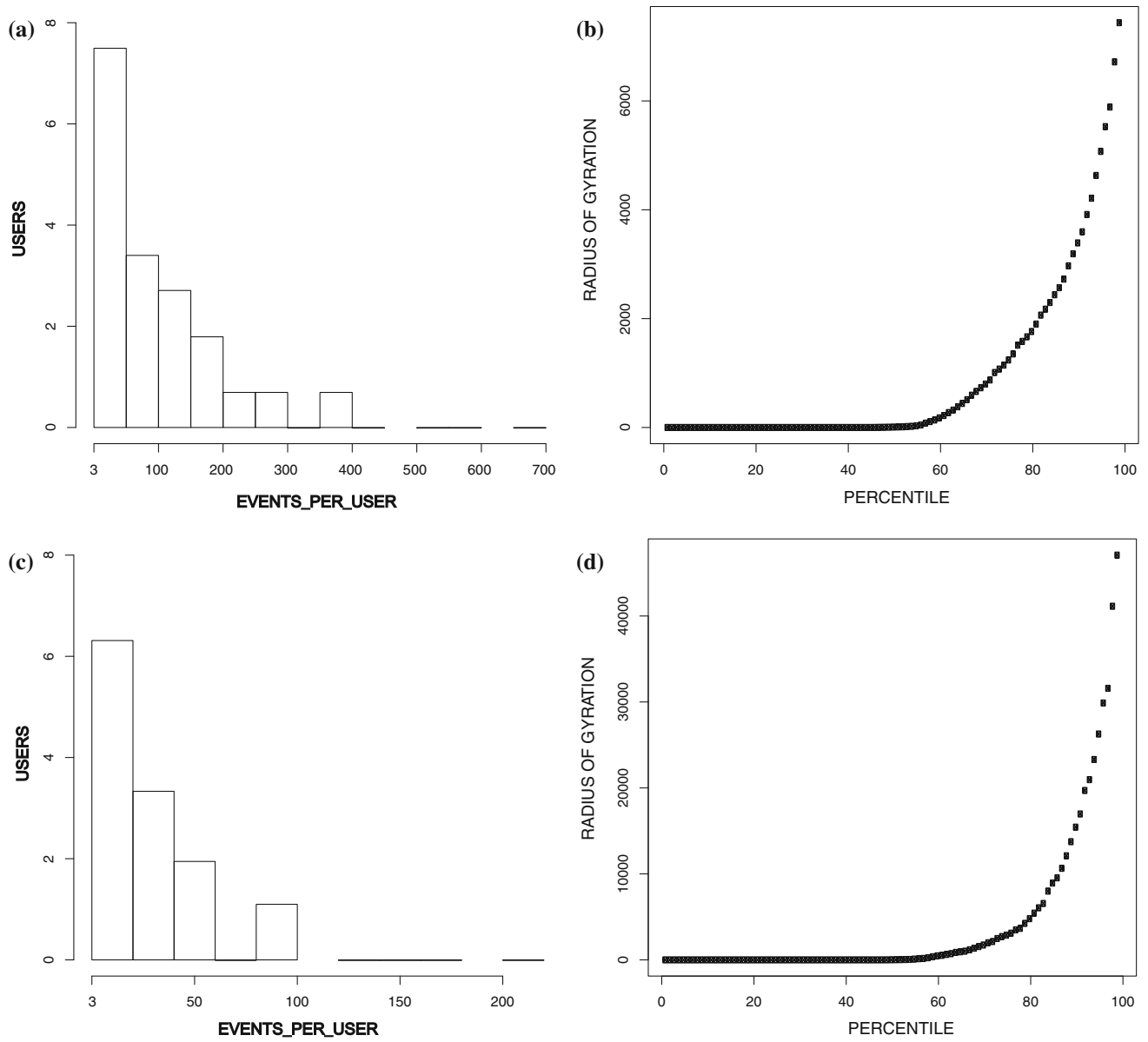

Fig. 3 **a** Twitter events per user. **b** Radius of gyration for a given percentile of Twitter users. **c** Flickr events per user. **d** Radius of gyration for a given percentile of Flickr users

4 Evaluating the uniqueness of mobility traces

In this section, we evaluate the quantity of information, in terms of geo-referenced events, that is necessary to identify a mobility trace as unique. Another way to deal with this problem is to find out for each user, what is the minimum number of events that set him/her apart from all the other users. This basically allows to answer the question: *who is the user who was in these locations at these times?* When the number of locations and times increases there is a *single* user compatible with all the sightings. A simple representation of this idea is given in Fig. 4 where the trace of user A is highlighted. In the figure, it is possible to see that as soon as we consider three events, we can pin point user A as no one of the other users have the same events.

4.1 Uniqueness test on CDR datasets

More formally, for each CDR user u_x we consider their data trace in terms of *id* of the network cell and *time* of the CDR event. We call cdr_j^x the j-th event generated by user x. Given a time interval Δt, we say that two CDR events from users x and y match if $cdr_j^x.id == cdr_i^y.id$ and $|cdr_j^x.time - cdr_i^y.time| < \Delta t$. We call M_j^x the set of users having at least one event (M)atching with cdr_j^x. For each

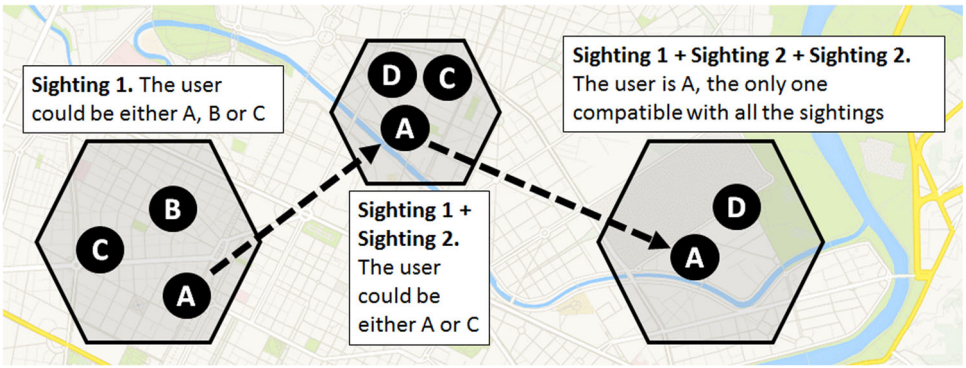

Fig. 4 Uniqueness of mobility traces. After three sightings (geo-referenced data) we can uniquely pinpoint user A

user x, we want to find the minimum number of events *min* so that $\bigcap_{j=0}^{min} M_j^x = \{u_x\}$. This is the minimum number of events necessary to set u_x apart from other users. The idea of a time limit is to take into account uncertainty on when the user created an event. So while there could be always just a single user generating an event at a given millisecond (thus $min = 1$), when we have uncertainty about the time of an event multiple users might be a match.

4.2 Experiments

We perform experiments with the CDR datasets (CDR-DATA1 and CDR-DATA2) to find the average number of points/events to pinpoint a user. To limit the computational effort, we take a sample of 1000 users from CDR-DATA2 and of 1000 users from CDR-DATA1 dataset. The uniqueness test is performed by an algorithm implementing the above definitions. In doing this type of analysis, we are interested basically in two types of results: (1) the average number of points/events needed for a trace to be unique and (2) the percentage of users that could be re-identified in terms of identification as unique. The test for matching the events is repeated with different time limits Δt that go from 2 to 30 min.

The summary of this analysis is given in Fig. 5 which provides results for the CDR-DATA1 and CDR-DATA2 datasets. These results are in line with those obtained by related studies summarized in Montjoye et al. (2013)and Gambs et al. (2013). Looking at Fig. 5-left it is possible to see that the number of CDR events to uniquely identify a user slowly grows with time interval. This is a rather natural phenomenon in that increasing the time limit loose the constraint about matching CDR. In the extreme case of time limit $\Delta t = 0$, all the users would be uniquely identified with only one event. As pointed out also by other authors in Montjoye et al. (2013), the slow growth in the number of points required for uniqueness illustrates that anonymization techniques based on blurring the time of the

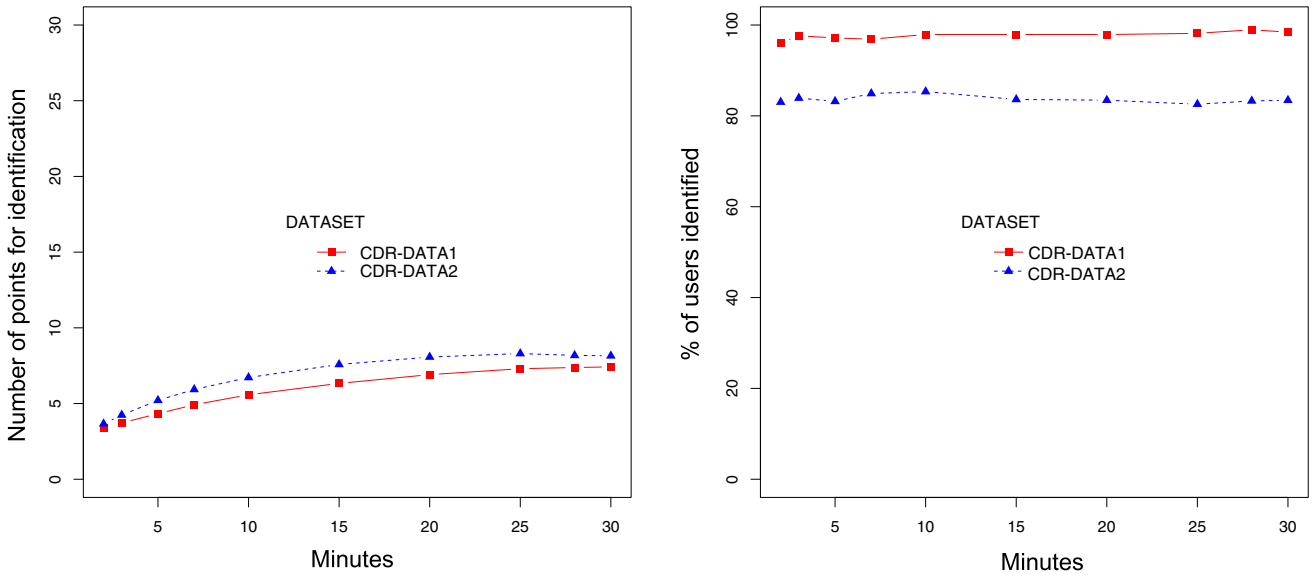

Fig. 5 Uniqueness of mobility traces. *Left* number of events needed for identification. *Right* percentage of users re-identified

events are not effective. These considerations are also represented in Fig. 5-right: the percentage of users that can be uniquely pin-pointed remains high (over 80 %) independently of the limit interval. In both the graphs, it is possible to see that it is slightly easier to uniquely identify users in the CDR-DATA1 dataset. This is probably because of the denser cell network in CDR-DATA1, rather than in CDR-DATA2—(denser cell means higher spatial resolution in the dataset). Also in this case, however, the difference between the two results is not dramatic, showing that the spatial resolution does not impact uniqueness a lot as stated in Montjoye et al. (2013).

In general these results are important to highlight issues related to CDR datasets with regard to users' privacy. However, to evaluate the potential for joining different datasets, this is not enough: even if a user is uniquely found in a dataset on the basis of their traces, it might be still difficult to find the same users also in other datasets, thus enabling the join.

5 Matching users among datasets

5.1 Matching CDR and social network data

As already introduced, matching users on the basis of mobility and geo-referenced data is in principle straightforward: it consists in identifying couples of users in the two datasets that often produce data at the same time and at the same place. Once enough geo-referenced elements overlap, we can be reasonably sure that the two users are actually the same person. Figure 6 exemplifies the matching process: Flickr/Twitter user FT_a is compared with CDR users u_x, u_y and u_z from the CDR-DATA1. The most probable match is $FT_a \equiv u_x$ as the two users appear almost always together.

The key novelty when considering multiple datasets is that the events from the same user in different datasets are *not* always in a 1-to-1 relationship. In contrast to the previous scenario, we can consider users in different datasets to be the same even if not all the events in one dataset have the corresponding events in the other dataset. In more technical terms, the intersection we used in Sect. 4 to uniquely identify users can produce the empty set. Still, given a sufficient overlaps among the data we will consider two users to be the same.

More in detail, for each Flickr/Twitter user FT_a and for each CDR user u_x we consider their respective data traces with a geographic and temporal reference. We confront these traces in order to find matches among couples of events. More formally, we call ft_i^a the events generated by user FT_a, and cdr_j^x the events generated by user u_x. We also call r_j^x the radius of the network cell associated to the event cdr_j^x.

Two events ft_i^a and cdr_j^x match if $s_{dist}(ft_i^a, cdr_j^x) < r_j^x$ and $t_{dist}(ft_i^a, cdr_j^x) < \Delta t$, where s_{dist} and t_{dist} are the spatial and temporal distance respectively.

We set the threshold for the spatial distance to r_j^x as the radius of the cell where the CDR event originates (we conducted experiments with the CDR-DATA1 dataset where this radius is reported. In other situations similar measures might be derived from the Voronoi tessellation of the network cells as described in Blondel et al. (2013). In particular, for each couple of users we count how many matching events are there. Since events are often produced in bursts, a naive matching approach between the traces could result in matching a number of CDR events with a single FT event, or vice versa. This would over-count the number of matching pairs as a single CDR/FT event could match with multiples FT/CDR events. To avoid such an issue, after recording an event cdr_j^x we skip all the events from the same users closer in time than Δt from cdr_j^x. In this way any given FT event can match with only one CDR event of a single user.

5.1.1 Exclusion condition

It is also important to consider the situation in which a CDR user u_y generates an event cdr_j^y that is close in time to an event ft_i^a generated by the Flickr/Twitter user FT_a, but the two events are far away in space. More formally

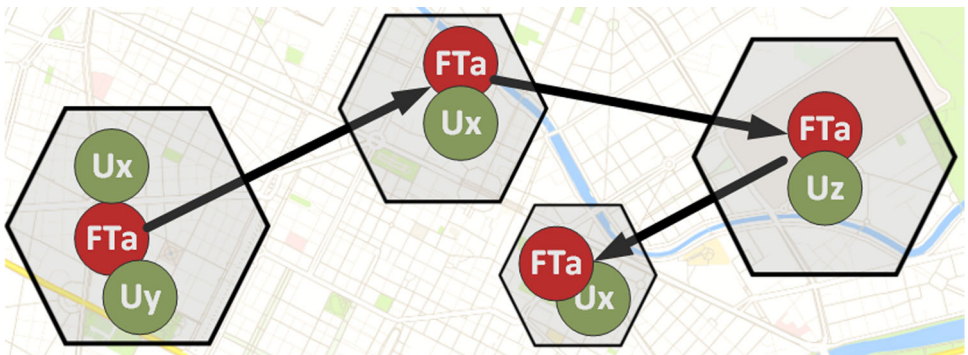

Fig. 6 Flickr/Twitter user FT_a (in *red*) is compared with CDR users u_x, u_y and u_z (in *green*). The most probable match is $FT_a \equiv u_x$ as they are consistently producing data at the same time and at the same place (color figure online)

$t_{dist}(ft_i^a, cdr_j^x) < \Delta t$ and $s_{dist}(ft_i^a, cdr_j^x) > 2 \cdot r_j^x$. In this case, it is impossible that the two users are the same in that we know that on that time they were physically apart. We call this case the *exclusion condition* as these CDR users can be excluded from the analysis. It is worth noticing that we considered $2 \cdot r_j^x$ in order to have a buffer-zone separating a match from a certain exclusion.

5.2 Experiments

5.2.1 Experiment 1: matching

In a first experiment we tested the above approach to find, for any FT user, the CDR user having the largest number of matching events. We say we re-identified the user if there is a *single* CDR user with such a number of events. The naive assumption is that such a single best-match is the same person of the FT user (hence the re-identification)—more on these aspects is in Sect. 6. In Fig. 7 we present the average number of points required for a re-identification and the percentage of the users than can be re-identified. We experiment with time intervals Δt from 10 s to 30 min. In general the figure shows that about 20 % of the FT users can be re-identified with about 4–5 matching events.

To support our study it is also interesting to consider the smartphones' usage pattern focusing on sms/calls and Internet connection requests for uploading a picture to Flickr or sending a Twitter message. The transition from one activity to the other should be happening in a short interval of time that would permit us to link a CDR with a FT event. Of course, the presence and frequency of this behavior is very important for the results of our study. Research in Verkasalo (2010) shows that voice call/sms functions and Internet browsing are the two most used functionalities, and the largest part of Internet activity is spent on social network sites and is mainly conditioned from the places and social context in which the user happens to be as in Do et al. (2011). According to statistics reported in Verkasalo (2010) the voice call/sms functions are used 24–27 days per month by 90 % of users. However, these data correspond approximately also to our CDR dataset statistics and we find an average of three events per day. On the other hand internet browsing is used 13–18 days per month by 85 % of users spending 60 % of the time in social network activities. Considering the above statistics, we assume that the call/sms-social network pattern, happens very often even in a short interval of time, thus enabling our approach to detect correspondences.

5.2.2 Experiment 2: matching statistics

We run experiments to count, for each FT user, the number of CDR users having 1, 2...n matching events with FT. Rather naturally, the number of CDR users that match with a FT user diminishes as we look for a larger number of events in common between them. In many cases, it converges to one after a certain number of points n which means that there is only one CDR user that has n points in common with the FT user. As reported in Montjoye et al. (2013) and Zang and Bolot (2011), the CDR user count (especially for a high-number of matching events) do not change significantly by changing Δt or by using a different threshold for the spatial distance.

Figure 8 (left) illustrates the statistics of the matching process. For a given number of matching events, we create a box plot. Each box plot describes the distribution of the number of CDR users having n matching events with the FT users. In the box plot, we report 25th percentile, median, 75th percentile. The boxplots extends 1.5 times the interquartile range. Such counts are in the log-scale to better appreciate the differences. It is easy to see that there are a lot of CDR users sparsely matching with FT users, while the number decreases for multiple matches.

Figure 8 (right) shows the percentage of cases in which those correspondences converge to $1, 2, \ldots, n$. In particular 23 % of the FT users have in common a certain number of points with only a single CDR user. *Does it mean that the two users are actually the same?* Of course, if the number of matching events would be large, we could be confident of an affirmative answer. Otherwise, such coincidences could be made by chance. We will discuss further this question in the next section making some more considerations.

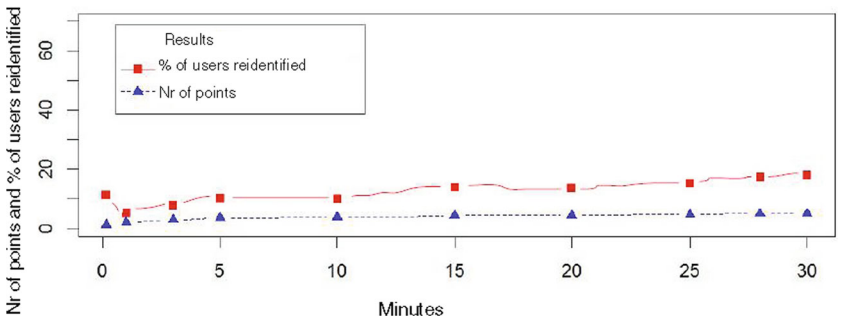

Fig. 7 Percentage of users re-identified and number of points needed for re-identification

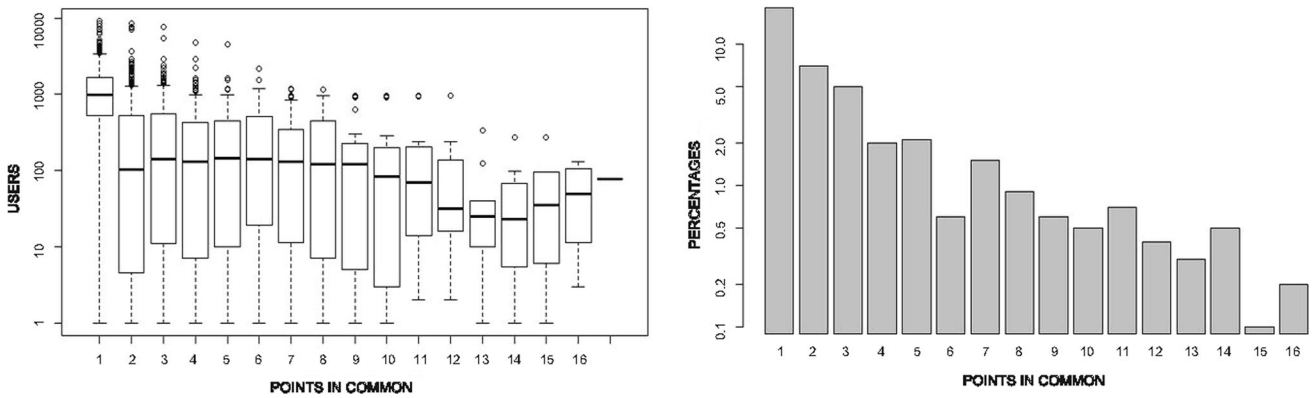

Fig. 8 *Right* boxplot diagram summarizing the statistic of the number of CDR users having X matching events with FT users. *Left* percentage of FT users for that can be associated to X number of CDR users

5.2.3 Experiment 3: matching by chance

We tried also to understand if the matching events can happen only by chance (this of course would invalidate our analysis). In particular, the Bonferroni's principle illustrated in Rajaraman and Ullman (2011) states that in a large dataset, like the one we consider, there are *always* some events matching a given signature, even if the events are produced randomly. In our settings, this means that even considering random FT users there will be CDR users that will resemble (by chance) FT users.

In order to avoid incorrect conclusions, and to validate our results we made a comparative experiment. For this purpose, a new dataset of mock-up FT users has been created on the basis of an artificial distribution of values (geographic coordinates and timestamps).

1. We created a first mock-up dataset by adding some noise to the original FT users' traces. In particular, we create a clone of each FT by displacing both in space (± 1 km) and time (± 30 min) the original user's events.
2. We created a second mock-up dataset by just randomly generating events within the area and time frame of the original data.

Testing with these fake users we find significantly less correspondences and less traces of series of points overlapping. More specifically, in the first case, we find 40 % less correspondences per user and in the second up to 60 % less correspondences than in the case with the real data. The fact that our data exhibits much larger correspondences means that most of the matches cannot be explained by chance only.

5.2.4 Experiment 4: validating results

Since groundtruth about the real identity of the users is missing, sound validation of the results is difficult. Partial validation evidence can be found in matching Flickr social users with Twitter social users. We have 231 users from the first social network and 734 from the second, and during the matching we notice that about 6 % of Flickr users, other than having similar traces have the same name or the same very singular username with Twitter users. Moreover those users with the same username often match with a single CDR user which is the same for Flickr and for Twitter users. Figure 9 (top) shows the traces of two exemplary social users with the same username. Figure 9 (bottom) shows that a single CDR user well matches will both of them. The three users are thus likely to be the same person.

A similar analysis can be conducted by comparing the CDR mobile county code (MMC), typically indicating the nationality of the user, with the language used in Twitter messages and Flickr tags. This type of analysis can also give an idea of the number of tourists that visit the region during the period of observation. In our CDR-IT datset about 15 million events have a mobile country code different from that of Italy and a time window permanence of 1–2 weeks (see Fig. 10-left). In the social dataset there are more than 200 users that use a language different from Italian. Figure 10 (right) shows along the y axis in logarithmic scale, the number of events and along the x axis the different languages we found in analyzing the text. To extract in an automatic way the languages of the social users we use an open Java library for language detection.[1] A common problem with this approach is that for example users from Switzerland can tweet or describe pictures in different languages such as German, French or Italian as well as people from nordic countries tend to tweet in English. Thus only 3 % of the social users could be re-identified supported by the match between Mobile Country Code and social users'language. Despite the scarce efficiency, this method functions well for the re-identification of users comming

[1] https://code.google.com/p/language-detection/.

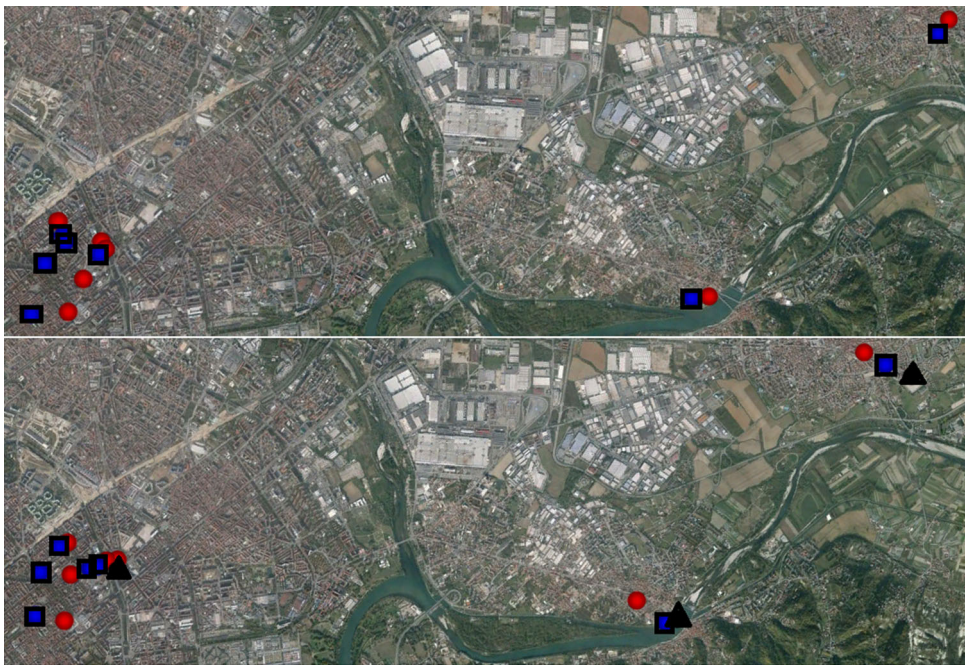

Fig. 9 *Top* matching mobility traces between Flickr events in *blue squares* and Twitter events in *red round shape* of the same social user. *Bottom* CDR events from a user in *black triangle*. The social user in its both profiles as Twitter and as Flickr user matches with the same CDR (color figure online)

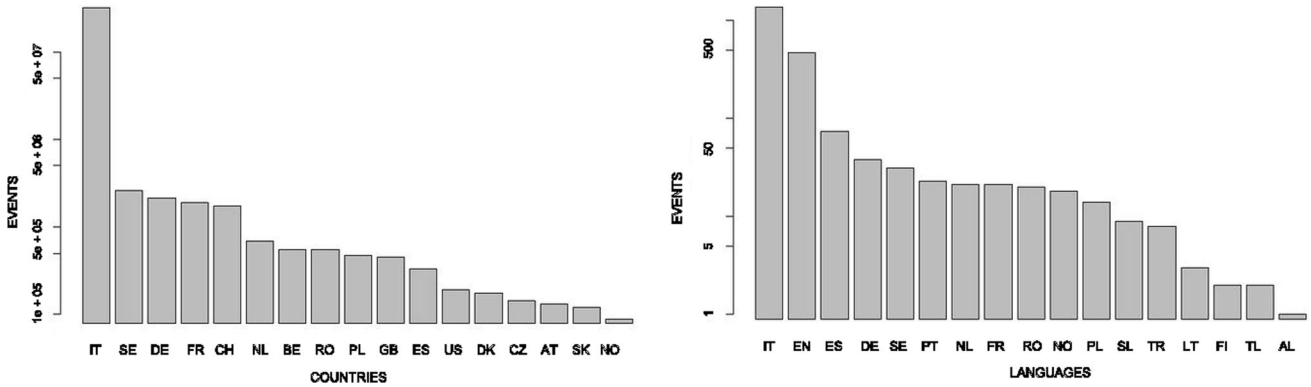

Fig. 10 *Left* distribution of events per mobile country code. *Right* distribution of social users languages

from small countries that remain faithful to their language while tweeting or describing pictures. From our point of view this provides one more clue in supporting ground truth evidence for the matching process.

5.2.5 *Experiment 5: reidentification with CDR-DATA2 dataset*

In analysing the CDR-DATA2 dataset we realize that it presents a challenge for our re-identification method. The dataset contains different subsets created in subsequent periods of time where in each subset there is a new random numeric identifier for the users. We choose 1500 users from the first subset and confront these users with all the users in the other subsets. Using the same approach as in the previous section, we repeat the matching process by time and tower id. While in the case of matching CDR-DATA1 and FT users we had data from the same period in this case we are matching data from different subsequent periods of time so we use time slots instead. This means that every time two users—the one from the 1500 sample and every other user in the other subsets—made an event with the same tower id and in the same time slot we had a match. In Fig 11 is represented an error bars plot about the matching process up to 50 points/events in common, while the percentage of users we can re-identify this way is 88.7

6 Probabilistic approach

In this section, we present a model trying to answer the following question: *given that the CDR user u_i has n_i events matching with the events of the Flickr/Twitter user*

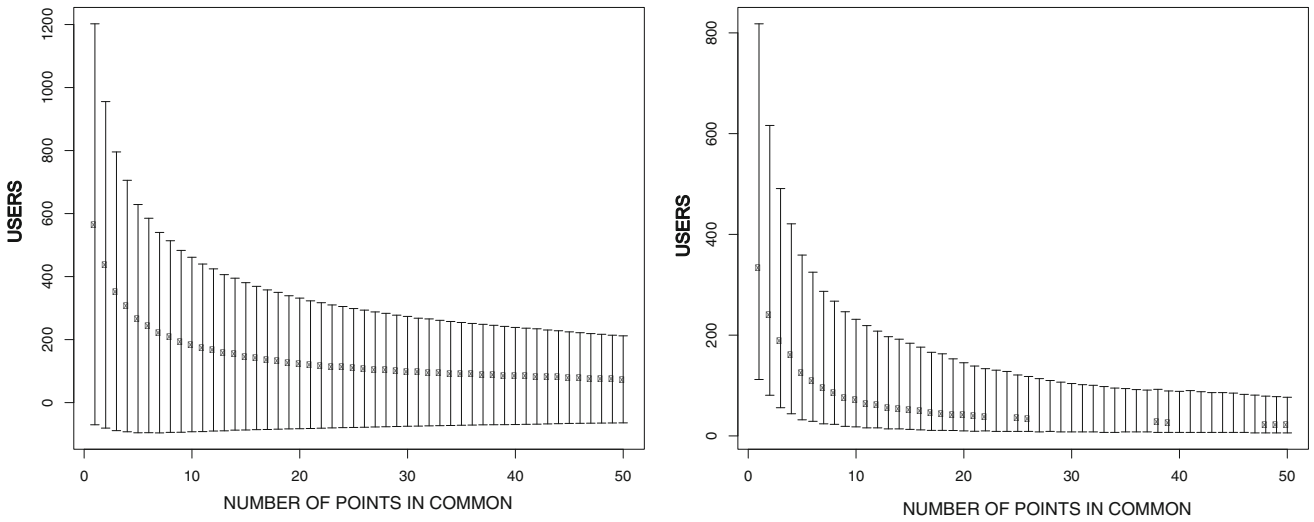

Fig. 11 *Left* error bars plot of the number of CDR-DATA2 users from the sample chosen having X matching events with all other users. *Right* the quartiles of the same distribution, 25th and 75th percentiles

FT_i, how likely it is that the two users are the same? In other words, how likely it is that we actually de-anonymized CDR user u_i?

This kind of question is both novel and fundamental for the practical applicability of our approach. In previous seminal works, as for example in Montjoye et al. (2013),this issue is not addressed. In Montjoye et al. (2013), it is assumed that for each FT event there is *always* the corresponding CDR event. Under such an assumption, re-identification reduces at finding the only CDR user that is *always* around the FT user. In other words, once we found that a *FT* user has n matching events with a *single* CDR user, we are 100 % sure that the two users are actually the same, even if the number of matching events n is 1. In our setting, since we removed the assumption that for each FT event there is *always* the corresponding CDR event, such a result is unreasonable in that a single common data point might be due to chance (as described in Sect. 5.1).

To answer this question we take inspiration from the work in Narayanan and Shmatikov (2008) in which they tackle the problem of re-identifying users in movie-ratings databases. Our approach starts by estimating the probability of a CDR and a FT users being the same person *given* they have n matching elements—$p(CDR \equiv FT|n)$. By definition of conditional probability, $p(CDR \equiv FT|n) = p(CDR \equiv FT, n)/p(n)$.

We compute $p(n)$ by counting in our dataset how many couple of users (CDR/FT) have n matching events. We divide this count by the total number of couples (CDR/FT).

Let us illustrate our approach with a simple numerical example. Suppose that in the the whole dataset there are only 3 CDR users (A, B, C) and 3 FT users (X, Y, Z). Assume that they have a number of matching events as in the following table:

Matches	A	B	C
X	30	1	3
Y	1	3	1
Z	1	1	3

For example CDR user A has 30 matching events with FT user X, 1 matching event with FT user Y and 1 matching event with FT user Z. In this setting we have:

$p(n = 1) = 5/9$, $p(n = 3) = 1/3$ and $p(n = 30) = 1/9$.

Computing $p(CDR \equiv FT, n)$ is more difficult as we do not have groundtruth information on the cases in which $CDR \equiv FT$. To estimate such a probability we *assume* that whenever a CDR user matches with a FT user such that there are not other CDR users having the same or greater number of matches, then the CDR user and the FT user are the same person (This is the same assumption made in Sect. 5.2—experiment 1).

Accordingly, $p(CDR \equiv FT, n)$ is the number of couples of users (CDR/FT) having n matching events in common *and* such that there are not other CDR users with a greater number of matches for the same FT user, *divided* by the total number of couples (CDR/FT).

In the example, $p(CDR \equiv FT, 30)$ = 1/9. In fact, looking at the table, 30 appears once associated to the couple $(A - X)$ and there are not other CDR users with a better match with X. $p(CDR \equiv FT, 3)$ = 2/9. In fact, looking at the table, 3 appears twice in situations in which there are not better matches (i.e., 3 appears in $B - Y$, $C - Z$).

We can then estimate $p(CDR \equiv FT|n) = p(CDR \equiv FT, n)/p(n)$.

In the example, $p(CDR \equiv FT|n = 30) = 1$, $p(CDR \equiv FT|n = 3) = (2/9)/(1/3) = 2/3$.

In the example it is possible to see that although B and C are the users that best match Y and Z the re-identification probability is *not* 1. This is because matches with 3 events are rather common in the dataset, so we are less confident of the re-identification. Considering the real dataset with a large number of CDR and FT users having multiple matches among each other, the approach becomes much more effective. Since there are *a lot* of CDR users having few events matching with some FT users, the re-identification probability on the basis of few events drops almost to zero. Only users with an unusually high number of matching events would be considered properly re-identified.

6.1 Experiments

We applied the above model to answer our motivating question: *given that the CDR user u_i has n_i events matching with the events of the Flickr/Twitter user FT, how likely it is that the two users are the same?*. For each Flickr/Twitter user, we recorded all the CDR users that produced events nearby and we computed the number of matches n_i among them. We run the above approach to estimate $p(CDR = FT|n)$. Figure 12-left shows the result we obtained. We run experiments changing Δt (i.e., the maximum tolerable time-distance between matching events).

The figure shows that for very small Δt (<2 min), or large Δt (>10 min) the $p(CDR = FT|n)$ remains low and drops to zero as the number of matches increases. This is highly counterintuitive, we would expect that the more the matches, the more the probability of the two users being the same person. This can be explained considering that for very small Δt there are 0 users generating n matches with any FT users (for $n > 3$). For large Δt, it is possible to have multiple matches also with different users. So, while $p(n)$ increases, $p(CDR = FT, n)$ does not keep up because (under our computational assumptions) there are few cases in which a single CDR user rises as the single most compatible one. For $\Delta t > 2$ min and $\Delta t < 10$ min we have the expected behavior: the higher the number of matches the greater the probability. Probability that quickly rises to almost one after 4–5 matching events (as expected). Accordingly, we assume this is the proper range to consider for further experimentation.

It is worth noticing that this kind of measure "solves" the issue raised at the beginning of this section, even if there is a *single* CDR user with n matching events with a FT users, if n is small, the chance ($p(CDR = FT|n)$) of the two users being the same person is low. Viceversa, if there are *multiple* CDR users with n matching events with a FT users, if n is large, the chance of all the users being the same person is high (e.g., a user with multiple phones).

Finally, we run another experiment to measure the fraction of users in our CDR-IT dataset that can be reliably matched with some FT users. Figure 12-right shows the maximum probability obtained for a given user percentile (we do not plot below-50 percentile as the probability is almost constant 0). It is possible to see that only the top few percentiles (still amounting at a lot of users) can actually be reliably matched with a corresponding FT user.

This is rather expected given the inherent difficulty in matching among different sparse datasets having non-deterministically correlated entries. It also highlights the difference between this task and the uniqueness identification task addressed in the related works and in Sect. 4.

7 Conclusion

Our intent in this paper is to present the potential and/or limits of a technique to re-identify users across multiple mobility datasets. Results illustrates that it is possible to identify some users across different datasets that are likely

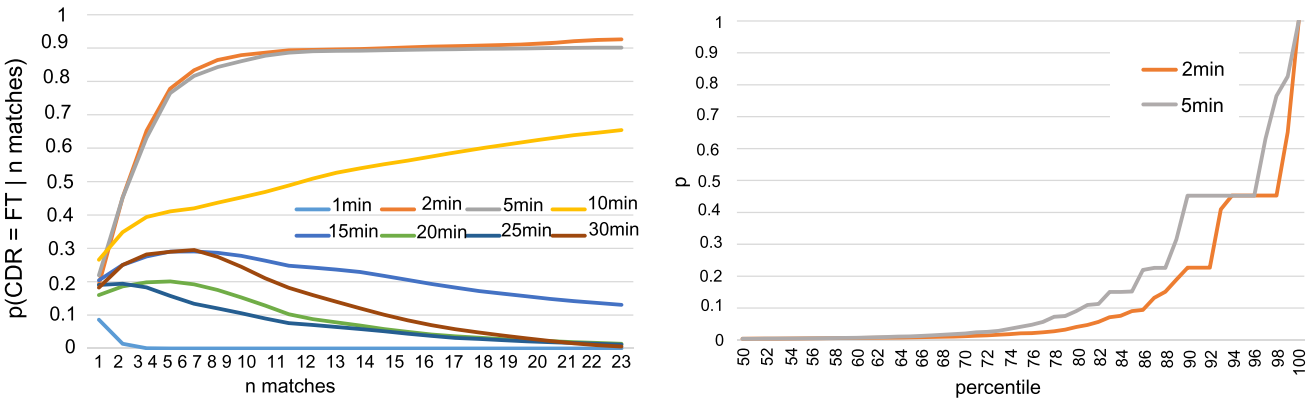

Fig. 12 *Left* probability $p(CDR = FT|n)$. Plots for different values of Δt used in matching process. *Right* maximum probability obtained for a given user percentile (we do not plot below-50 percentile as the probability is almost constant 0)

to be the same person, in the sense that the match among them is largely more probable than the match with any other person. While correlation among datasets (i.e., CDR and social network data) represent a fundamental issue with regard to privacy issues, it also represents a big opportunity to enrich the information available to a pervasive applications. In fact correlation is also the basic building block to fuse different perspectives and information together in order to obtain a multi-faceted representation of phenomena and events influencing the application. For example fusing CDR and Flickr/Twitter data it is possible to better pinpoint where the user was at a given time and what (s)he was doing in there. Moreover, such information could provide a much more fine grained view on the user profile enabling high forms of personalization and context awareness. Our future research on this topic will continue into three main directions:

The probabilistic model presented in Sect. 6 is rather simple. On the one hand, it is based on a number of independence assumptions that can be hardly justified in the real data. On the other hand, we think that further analysis on the eccentricities of the resulting probability distribution could give further insights on the re-identification process as research shows in Narayanan and Shmatikov (2008). With regard to this latter aspect, we plan to investigate and adopt standard privacy measures like k-anonimity concepts presented in Sweeney (2002) and Zang and Bolot (2011) and differential privacy described in Dwork (2011). Another interesting aspect are social ties as work in Crandalla et al. (2010) which is worth investigating for our social users.

As privacy concerns are the main impeding factors preventing CDR data (and pervasive/mobility data in general) to be applied to pervasive applications, it would be fundamental to develop anonymization and privacy-preserving mechanism that preserve data usefulness in the context of pervasive application. There are a number of researches addressing such issues as for example in Brickell and Shmatikov (2008) and Zang and Bolot (2011). However, most general approaches conclude that: *even modest privacy gains require almost complete destruction of data-mining utility* as stated in Brickell and Shmatikov (2008) and Zang and Bolot (2011). We think that a viable approach could be based on application-specific anonymization mechanisms: mechanisms preserving privacy and data mining utility for a *single* aspect (the one useful of the specific application). We will try to investigate these issues in our future work.

Finally, we will try to leverage the proposed re-identification approach to join multiple complementary datasets together to develop advanced context-awareness services in smart city scenarios. In particular a good source of inspiration, of possible applications as well as advances on balancing risks and rewards on data-driven applications is given in Pentland (2014).

Overall, we think that the above research directions will have a strong role in increasing even further the impact of human mobility data on the achievement of the pervasive computing vision.

Acknowledgments Work supported by the SOMUS project (POR-FESR 2007–2013) and by the ASCENS project (EU FP7-FET, Contract No. 257414).

References

Abraham R (2006) Mobile phones and economic development: evidence from the fishing industry in india. In: The international conference on information and communication technologies and development, ICTD 2006, IEEE

Abul O, Bonchi F, Nanni M (2010) Anonymization of moving objects databases by clustering and perturbation. Inf Syst 35(8):884–910

Blondel VD, Esch M, Chan C, Fabrice Clerot PD, Huens E, Morlot F, Smoreda Z, Ziemlicki C (2013) Data for development: the d4d challenge on mobile phone data. Orange Data Dev Chall. Scientific Reports 3, Article No: 1376. doi:10.1038/srep01376

Brickell J, Shmatikov V (2008) The cost of privacy: destruction of data-mining utility in anonymized data publishing. In: International conference on knowledge discovery and data mining, New York, NY, USA

Crandalla DJ, Backstromb L, Cosleyc D, Surib S, Huttenlocherb D, Kleinbergb J (2010) Inferring social ties from geographic coincidences. In: Proceedings of the National Academy of Sciences, vol 107, issue 52, pp 22436–22441. doi:10.1073/pnas.1006155107

Do TMT, Blom J, Gatica-Perez D (2011) Smartphone usage in the wild: a large scale analysis of applications and context. In: ICMI '11 Proceedings of the 13th international conference on multimodal interfaces, Alicante, Spain, pp 353–360

Dwork C (2011) The promise of differential privacy: A tutorial on algorithmic techniques. In: IEEE symposium on foundations of computer science, Palm Springs, CA, USA

Gambs S, Killijian MO, del Prado Cortez MN (2013) De-anonymization attack on geolocated data. In: The 12th IEEE international conference on trust, security and privacy in computing and communications (IEEE TrustCom-13), Melbourne, Australia

Golle P (2006) Revisiting the uniqueness of simple demographics in the us population. In: 5th WPES workshop on privacy in electronic society, Alexandria, VA, USA

Montjoye Y, Hidalgo A, Verleysen M, Blondel V (2013) Unique in the crowd. the privacy bounds of human mobility. Sci Rep 3:161–180

Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. In: IEEE symposium on security and privacy, Las Vegas, Nevada, USA

Parent C, Spaccapietra S, Renso C, Andrienko G, Andrienko N, Bogorny V, Damiani ML, Gkoulalas-Divanis A, Macedo J, Pelekis N, Theodoridis Y, Yan Z (2013) Semantic trajectories modeling and analysis. J ACM Comput Surv (CSUR) 45(42):161–180

Pejovic V, Musolesi M (2015) Anticipatory mobile computing: a survey of the state of the art and research challenges. ACM Comput Surv 47(3), Article No. 47. doi:10.1145/2693843

Pentland A (2014) Big data: Balancing the risks and rewards of data-driven public policy. 2014 World Economic Forum The Global Information Technology Report 2014

Rajaraman A, Ullman JD (2011) Mining of massive datasets. Springer, Berlin

Rossi L, Musolesi M (2014) It's the way you check-in: identifying users in location-based social networks. In: Proceedings of the second edition of the ACM conference on Online social networks, Dublin, Ireland

Sharad K, Danezis G (2013) De-anonymizing d4d datasets. NetMob, Cambridge

Sweeney L (2002) k-Anonymity: a model for protecting privacy. Int J Uncertainty Fuzziness Knowl-Based Syst 18(10):557–570

Sweeney L, Abu A, Winn J (2013) Identifying participants in the personal genome project by name. White Paper 1021–1, Harvard University Data Privacy Lab

Verkasalo DH (2010) Analysis of smartphone user behavior. In: Ninth international conference on mobile business/2010 ninth global mobility roundtable

Wicker S (2012) The loss of location privacy in the cellular age. Commun ACM 55(8):60–68

Zang H, Bolot J (2011) Anonymization of location data does not work: a large-scale measurement study. In: MobiCom11, Las Vegas, Nevada, USA