Towards matching user mobility traces in large-scale datasets

Dániel Kondor, Behrooz Hashemian, Yves-Alexandre de Montjoye, Carlo Ratti

Abstract—The problem of unicity and reidentifiability of records in large-scale databases has been studied in different contexts and approaches, with focus on preserving privacy or matching records from different data sources. With an increasing number of service providers nowadays routinely collecting location traces of their users on unprecedented scales, there is a pronounced interest in the possibility of matching records and datasets based on spatial trajectories. Extending previous work on reidentifiability of spatial data and trajectory matching, we present the first large-scale analysis of user matchability in real mobility datasets on realistic scales, i.e. among two datasets that consist of several million people's mobility traces, coming from a mobile network operator and transportation smart card usage. We extract the relevant statistical properties which influence the matching process and analyze their impact on the matchability of users. We show that for individuals with typical activity in the transportation system (those making 3-4 trips per day on average), a matching algorithm based on the co-occurrence of their activities is expected to achieve a 16.8% success only after a one-week long observation of their mobility traces, and over 55% after four weeks. We show that the main determinant of matchability is the expected number of co-occurring records in the two datasets. Finally, we discuss different scenarios in terms of data collection frequency and give estimates of matchability over time. We show that with higher frequency data collection becoming more common, we can expect much higher success rates in even shorter intervals.

Index Terms—individual mobility traces, privacy, identifiability, data fusion, matchability

1 Introduction

NOWADAYS many service providers routinely collect mobility traces of individuals. These constitute various types of data such as call detail records (CDR) from mobile phone usage [1], smart cards used in public transportation systems and for identification [2], financial transactions such as payments made with bank cards or mobile devices [3], and GPS coordinate updates recorded by smartphone apps [4], [5]. While these provide a valuable data source for researchers [1], [6], [7], [8] and also enable various services [2], the high amount of tracking of individual mobility has raised serious concerns about privacy in several different contexts [9], [10].

This has been emphasized by research that shows that these mobility traces are highly unique, warning that identifying an individual in a mobility dataset based only on their observed records must be considered as a real possibility [11], [12]. The basis of this argument is that since a small number of records uniquely identifies an individual, then reidentification can be achieved based on a relatively small amount of information, e.g. by following someone for only a short amount of time, or by merging with an external dataset even with a short timespan. Furthermore, the possibility of such deanonymization existing *at all* is counter-intuitive to the perception of anonymity achieved in a crowd of strangers that is typically associated with urban life [11], [13].

- *D. Kondor is with the Senseable City Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA and the Future Urban Mobility group, SMART, Singapore 138602*
- *B. Hashemian and C. Ratti are with the Senseable City Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA*
- *Y. A. de Montjoye is with the Dept. of Computing and Data Science Institute, Imperial College London, London SW7 2AZ, UK*

On the other hand, fusing data at an individual-level from different sources is expected to provide valuable new insights for studies in personal mobility and urban planning e.g. by relating mobility and social characteristics [6], helping the development of new security and privacy policies, and benefiting the people involved by offering new services [14], [15]. In accordance with that, previous work has tried to establish methodology for effectively *matching* mobility datasets based on user traces [4], [16], [17], [18].

In this paper, we evaluate *matchability*: the possibility of matching users between large-scale anonymized datasets based on their trajectories. This is in contrast to previous work which has focused on the potential of reidentifiability of anonymized records at the individual level [11], [12], [27] We utilize two data sources, each of them containing mobility traces of millions of people over the course of one week, a statistically representative sample of a metropolitan area. While ground truth data about corresponding traces is not available to perform direct evaluations, we address the main challenges present when dealing with datasets containing fine grain mobility traces of millions of people, as is the case in a major city. Further, we evaluate the expected success rate of a matching procedure in a realistic scenario, providing first results on such a large scale. Our main contributions in this paper are the following:

1: We study the problem of matchability using two datasets which correspond to a significant sample of the population in the area considered. To our best knowledge, this is the first attempt to estimate the potential for merging datasets on this scale. This presents a realistic scenario in terms of computational complexity and data density, i.e. the number of false positives is non-negligible.

2: We evaluate and develop a matching methodology which can handle data of this size; a main objective is to be able to perform the matching without having to evaluate a similarity metric among any pair of users which would present prohibitively high computational complexity. We make our implementation available to the research community as open-source software which performs the search efficiently on datasets consisting of few hundred million records of several million users each.

3: We develop an empirical framework for establishing the *matchability* of the datasets and use it to evaluate the expected success rate of the matching methodology to estimate the required data collection period for successful matching of users given their activity. This work is extensible to more complex search and matching strategies as well.

2 Related work

The problem of *matchability*, along with the related problems of reidentifiability, uniqueness and concerns for privacy has been in the focus of research for several decades. An early systematic treatment was presented by Fellegi and Sunter in the 1960s [19], with the goal of merging incomplete databases obtained from various sources; their work was inspired by the beginning of the large-scale deployment of computer systems for storing records in electronic formats and presenting efficient search capabilities for the first time. At that time, the resources required for acquiring and analyzing large-scale datasets were still limited it to the government, large corporations and a few academic institutions; privacy concerns were already present, at least with respect to what data could be released to the public e.g. from census databases [20], [21], [22], [23].

Since then, especially the past two decades brought the proliferation of data which is collected and shared about a significant fraction of the population. A tremendous amount of information is publicly shared on online social networks (OSNs) [24], while products and services allowing the tracking of individuals have gained a high penetration rate. These include credit cards, mobile phones and smartphones, RFID-based payment or identification systems, and subscription-based online services; all have the possibility of generating a large amount of personal data about their users [1], [2], [8], [9]. While this continues to raise concerns from several parties, there exist many new opportunities using insights based on the data, both commercial and academic, giving operators an incentive for sharing the collected data [10].

Given that the amount of data routinely collected about individuals is increasing rapidly, several recent works have tried to evaluate the possibility of reidentifying records in a large dataset or matching records from different sources. The work of Sweeney established the notion of k-anonymity as a guarantee considering the privacy implications of releasing data with records of individuals [25], [26], formalizing intuitive requirements that records of individuals should not be unique. Further work showed that k-anonymity in many cases can be impractical to achieve, especially in the case of high-dimensional but sparse datasets, which typically exhibit a high level of uniqueness. For example Narayanan and Shmatikov, using data released publicly by Netflix, showed that identifying individuals can be possible based on knowledge of only a small number of their records [27], [28]. Zhang et al. present a more general treatment of data disclosures and provide algorithms for obtaining sub-data suitable for releasing under more general constraints than k-anonymity [29]. On the other hand, Dwork argues that for any data release, the possibility of combining it with external information also needs to be considered and privacy cannot be guaranteed in a general setting; the definition of *differential privacy* is suggested to quantify the risks of sensitive information that can be gained this way [30]. Their approach and motivation is in many ways similar to previous work on statistical databases [22], [23], with considering more general cases. Further work focused on linking Bitcoin addresses to IP addresses, showing the limits of privacy [31], [32], while the problem of record linkage has important applications in disciplines like biology or astronomy, facilitating probabilistic matching of data obtained from different measurements [33], [34].

As location data about individuals is being collected at an increasing pace, several previous studies considered reidentifiability, uniqueness and matchability among mobility traces of individuals. The work of De Mulder et al. was among the first to consider the possibility of identifying anonymized location traces from mobile networks; looking at a dataset containing data recorded on the phones of 100 volunteers, they achieved an identification accuracy of over 80% [35]. Other works showed that mobility traces of individuals are highly regular and predictable [36], [37], and that this regularity can be exploited for identifying people based on knowledge of their historical trajectories [38]. In related work, Crandall et al. look at the problem of inferring friendships form social media mobility traces based on the assumption that friends will have more shared location updates than strangers [14], while Li et al. studied the problem of measuring similarity between different people's location histories with the goal of adapting recommendation systems to include spatial data as well [39].

Considering the problem of identifying users based on a sample of their trajectories in truly large-scale datasets, de Montjoye et al. introduce the concept of *unicity* and find that mobility traces in mobile phone network usage and credit card transaction data are highly unique: even in datasets containing the mobility traces of more than a million people, only four randomly picked records from a person's trace uniquely identifies a large majority of traces [11], [12]. To mitigate these concerns, He et al. suggested an advanced anonymization methodology based on applying the concepts of differential privacy to location data to achieve a form of k-anonimity [40]. Nevertheless, these results suggest the possibility of performing a systematic reidentification among two different large-scale data sources, effectively *matching* traces of users present in both datasets merely based on their records of movement. In line with this, several studies have performed experiments to match users in distinct datasets based on trajectories. In an early work, Malin and Sweeney suggest the possibility of deanonymizing genetic data (i.e. DNA sequences), by comparing trajectories of patients obtained from medical records with matching DNA sequences they left at several different institutions [41]. Looking at the problem of matching trajectories in a more typical setting, Cecaj et al. [16] perform a search for the trajectories of a sample of about one thousand social media users in a mobile network dataset and present an estimate for the number of users matched based on a probabilistic model as their data source does not include ground truth information. Riederer et al. [17] developed a probabilistic matching algorithm based on bipartite graphs on pairs of datasets with readily available ground truth information, while Cao et al. [4] define a signal based similarity measure to integrate data collected via various mobile apps. Recently, Basık et al. [18], [42] performed a large-scale study testing matching algorithms between CDR, social media and synthetic data.

Most of the above studies considering matching traces in mobility data [4], [16], [17] share the limitation that search and matching is only performed for a limited set of sample users in the range of few tens of thousands; while this provides a test-case for development of algorithms, the density of the dataset (i.e. the number of records occurring in a given space and time) will highly affect the methodology applied for matching and the scalablility and validity of algorithms. Our datasets allow us to study the problem of matchability on the scale of millions of users. As this presents new challenges in terms of computational performance, we only consider a matching strategy which performs an efficient search instead of evaluating any possible pair of users as a match candidate. However, given the scale of the dataset, we will have a realistic estimate on the probabilities of finding false positive matches, highly unlikely when using a dataset containing only a few thousand trajectories. In this respect, our work is most similar to Basık et al. [18], [42], who utilize two large datasets as well (containing of several hundred thousand and a few million users respectively), but do not present relevant statistics about the matches between these two datasets due to the lack of ground truth data. They do present an extensive analysis on the quality of matching between a synthetic dataset generated from CDR data and the original CDR dataset. Using a different approach, we now present a probabilistic framework to assess the possibility of matching real-world mobility traces between two large-scale datasets.

3 Data description

In this work, we utilize one week of mobile communication and transportation data from the city state of Singapore recorded during the spring of 2011. The mobile communication dataset was provided by Singtel, the largest mobile network operator (with a market share of over 45 %) and contains 485,237,708 individual records of 2,844,721 users, where one record represents the start or the end of a call (either placed or answered) or sending or receiving a text message and includes the timestamp and the geographic coordinates of the antenna the user was connected at the time. The transportation data comes from the Singapore Land Transportation Authority (LTA) and is based on the smart cards used by the electronic fare system on buses and trains. This dataset contains 71,319,524 individual records produced by 3,348,628 unique smart cards where one record corresponds to either boarding or exiting a bus or train and includes the timestamp and the coordinates of the corresponding stop. Train rides always include both the start and the end of the journey with the possibility of transfers in between not necessarily recorded, i.e. the start and end stations can be on different lines. Bus rides include the end of the journey only when the passenger performs an additional *tap-out* while exiting the vehicle. Doing so is optional and incentivized by providing fare discounts (i.e. the fare is billed based on the actual travel distance instead of a flat fee); this is highly effective as evident by the fact that about 94 % of bus rides in our dataset record the end of the trip as well.

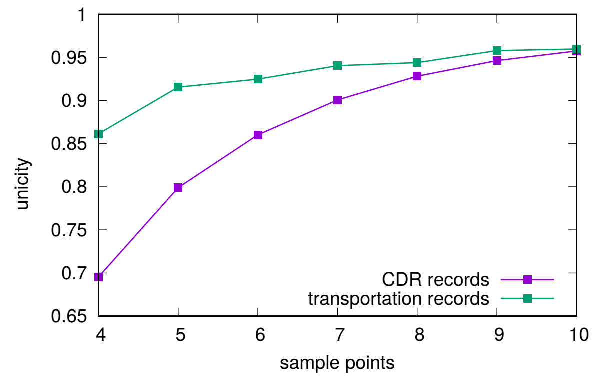

Figure 1. Evaluating unicity with $d = 500\,\mathrm{m}$ and $\tau = 5\,\mathrm{min}$.

As a first step to characterize the data, we look at the temporal distribution of records as well as basic statistics of the number of records per user (summarized in Fig. S3 in the Supplementary Material). It is clear that both data sources are relatively low-density as especially in the case of the transportation dataset, most people are expected to travel by public transportation only a few times per day and we only have records at the start and end of their trips.

To better characterize temporal distributions of activity during the week, in Fig. S2 in the Supplementary Material we display activity distributions for transportation and mobile network users. We do this by grouping users based on their level of activity. The distributions are remarkably similar among the groups of users with different level of activity for the mobile network dataset, but they are significantly different in the transportation dataset. This can be explained by noting that the less active user groups are more likely to be casual users, i.e. people who do not use public transportation as their primary means for commuting to work; this seems to be the largest difference between the groups. Also, in Fig. S4 in the Supplementary Material we display the distribution of average travel speeds in the transportation dataset, used later as the basis of defining spatiotemporal neighborhoods of transportation events.

3.1 Evaluating unicity

Before proceeding to estimate the matchability of the two datasets, we first evaluate the uniqueness of mobility traces using unicity as the measure [11]. We select random samples from users' records and test whether these uniquely identify them in the dataset. We show results for *unicity* (i.e. the ratio of users uniquely identified) in Fig. 1. These are similar to that of previous work [11], [16], although slightly lower. This is possibly the result of our datasets being denser (both in space and in time and also in the case of public transportation use, we expect the train stations to be more crowded). Furthermore, in the case of the communication dataset, instead of grouping records by antennas, we evaluate spatial proximity based on the Voronoi-polygons centered on antenna locations. This is a difference from previous studies [11], [16] which we expect to result in lower values of unicity, but model the process of matching records in two different datasets better as well. We point out a fundamental difference from matching two datasets: when evaluating unicity, we know that there is a match for each record (i.e. the record itself), and then increasing the search radius (either in space or time) adds potential false positive matches, giving rise to decrease in unicity [11], [12]. On the other hand, when matching records from different datasets, most of the records are likely to not have a match in the other dataset from the same user (see also Fig. 3 and the related discussion in Section 5); increasing the search radius will increase both the chance of a record of the same user to be matched and the number of false positives.

4 Spatiotemporal matching

The main goal of this work is thus to study matchability as a function of the length of the data collection period and the amount of data available per individual. We emphasize that to do this reliably, it is necessary to have a realistic estimate of the levels of activity both in terms of spatiotemporal distribution and density. While our datasets do not include ground truth information on matching trajectories, they provide a significant sample of all activities in a compact metropolitan area, allowing to do such estimation more reliably. However, having such datasets presents obvious challenges in terms of data handling and computational complexity, limiting methodology to those that scale up efficiently to these large scales.

To be able to evaluate matchability, we first need to define the matching process whose expected success rate we will estimate. In this section, we present a simple choice for matching users where expected matching performance can be evaluated probabilistically, without the need for ground truth data.

Our methodology is a spatiotemporal search which yields candidate user pairs and can be carried out efficiently using standard indexing techniques without having to perform a comparison of all possible pairs of trajectories. We search for matching points in the spatiotemporal neighborhood of each record of every user, excluding candidates who have records temporally close, but spatially distant. For each user we select the one in the other dataset with the highest number of matching records as a candidate.

While our matching procedure is relatively simple, it scales well to datasets of several hundred million points each and can be easily adapted to estimate matchability based on the probability of finding false positive matches. It would also allow further extensions (e.g. with weighting matches based on local density) if ground truth data were available to perform training on. In this sense, we expect our results on matchability to be lower bounds on the success rate of any matching algorithm developed with real-world data.

4.1 Notations and preliminaries for spatiotemporal matching

We first need to define when we consider two records to be matching. Let C and T denote two mobility datasets (i.e. communication and transportation), with n_C and n_T individual users respectively. We denote the records of user i in either dataset as $x_{ik}^{\alpha} = (\mathbf{x}_{ik}^{\alpha}, t_{ik}^{\alpha})$, where $\alpha = C, T$, $i = 1, 2, \ldots n_\alpha$ and $k = 1, 2, \ldots r_i^\alpha$ where r_i^α is the number of records of user i in dataset α In the case of the transportation dataset ($\alpha = T$), for each record, we will further use a flag $S_{ik}^T = 0, 1$ which indicates whether that the record corresponds to the start ($S_{ik}^T = 1$) or the end ($S_{ik}^T = 0$) of a journey. Using these notations, we define two points to be a *spatial match* if they are close in space and time:

$$|\mathbf{x}_{ik}^C - \mathbf{x}_{jl}^T| \leq d \quad \text{and} \quad |t_{ik}^C - t_{jl}^T| < \tau \tag{1}$$

for some parameters d and τ which define spatiotemporal neighborhoods. Further, we define two points to be an *impossible match*, if they are close in time, but separated in space:

$$|\mathbf{x}_{ik}^C - \mathbf{x}_{jl}^T| > d \quad \text{and} \quad |t_{ik}^C - t_{jl}^T| < \tau\,. \tag{2}$$

Following Basık et al., we also use the term *alibi* to refer to impossible matches [18]. We use the term *temporal match* for pairs of points which are either a spatial or impossible match (i.e. any pair of points separated by less than τ time irregardless of their spatial distance). Essentially these definitions mean that we consider two points to possibly belong to the same user if they are separated by maximum τ in time and d in space, while we consider them to certainly belong to distinct users if the temporal separation is less than τ and the spatial separation is more than d. Note that pairs of points with temporal separation larger than τ are not considered in any way.

To perform dataset matching, we then need to choose the d and τ parameters such that they are consistent with typical mobility patterns of individuals. To better accommodate for the characteristics of urban movements, we further refine these thresholds by differentiating between walking and traveling with transit, based on the transportation records. We thus use separate parameters d_w, τ_w for walking and d_t, τ_t for transit. We refine the definition of the parameters used for establishing temporal, possible and impossible matches as following:

S_{ik}^T	t order	τ	d
1 (start)	$t_{ik}^C < t_{jl}^T$	τ_w	d_w
0 (end)	$t_{ik}^C > t_{jl}^T$	τ_w	d_w
1 (start)	$t_{ik}^C > t_{jl}^T$	τ_t	d_t
0 (end)	$t_{ik}^C < t_{jl}^T$	τ_t	d_t

(3)

In practice, we chose the parameters as $d_w = 500\,\mathrm{m}$, $\tau_w = 10\,\mathrm{min}$, $d_t = 2\,\mathrm{km}$ and $\tau_t = 5\,\mathrm{min}$ according to the typical travel speeds we found in the data. Note that this implies a typical average transit velocity of $24\,\mathrm{km/h}$ during this $5\,\mathrm{min}$ period; for bus rides, typical average travel speeds are below this (see Fig. S4 in the Supplementary Material), while for train rides, one has to consider that this time interval includes the time needed to reach and enter or exit the train from the station entrance where the smart card is validated. Looking at the distribution of distances spanned by trips shorter than 5 minutes (as shown in the inset of Fig. S4 in the Supplementary Material) and between 5 and 10 minutes, we estimate that only 1% of trips shorter than 5 minutes has a distance larger than $2\,\mathrm{km}$, while only 2% of trips between 5 and 10 minutes spans a distance larger than $4\,\mathrm{km}$, meaning that our choice of spatiotemporal neighborhoods is realistic for a large majority of trips.

We finally need to take into account the spatial uncertainties of the data. While in the case of transportation records, the location of stops is exact (i.e. a record implies that the corresponding user was present at the exact location at that exact time), for the cell phone data, antenna locations are only an approximation of the users' location as they could be anywhere in the antenna's corresponding reception area or even in a neighboring region if the antenna closest to them is experiencing heavy traffic. To take this into account, we calculate the Voronoi tessellation of unique antenna locations, and consider the user to be possibly present anywhere within the Voronoi cell which corresponds to a particular antenna. For a certain public transportation stop and cell phone antenna, records at these two which match temporally will be considered as *spatial matches* if a circle of d radius around the transportation stop and the Voronoi-polygon associated to the antenna have any overlap, while these records will be considered *impossible matches* if there is no overlap. We display an example of evaluating such overlaps in Fig. S1 in the Supplementary Material.

4.2 Matching users

Based on the previous considerations, we perform a search procedure among the two datasets which results in a list of candidate matching pairs. While the size of the data is fairly large, we exploit the fact that there is a limited number of possible matching mobile network antenna – transportation stop pairs, pre-compute the list of these based on the Voronoi-polygons (see the previous sections) and use an indexing in time and by the antenna or stop IDs. This allows us to avoid performing a spatial search and instead use a range search (in time) along with a dictionary search (among the possible antenna – stop pairings). Using this strategy, the search for possible match candidates can be performed in the matter of 40 hours using a mid-grade server with 18 virtual cores and 96 GB available memory. We estimate that evaluating all temporal matches for each user pairs would take approximately 12 days. Note that this latter computation requires considering all $n_C \times n_T$ user pairs, while during the matching procedure, this complexity is significantly reduced by only considering candidate pairs that have co-occurring data points, resulting in the shorter computational time. We make the source code of all programs utilized in these calculations available at `https://github.com/dkondor/matching`. A formal overview of the matching procedure is presented in Algorithms 1 and 2.

We consider a pair of users $i \in T$ and $j \in C$ as a candidate if any of their record pairs x^T_{ik} and x^C_{jl} are spatial matches and none of their possible record pairs are alibis. In this case, we define the number of matches m_{ij} between the two users as the maximum number of possible match pairs such that each record is used once at maximum (i.e. we exclude multiple matches for a single record). In the case of any alibis, we define $m_{ij} \equiv 0$. For each transportation user, we select the CDR user with the highest number of matches as a candidate to be its counterpart. We then refer to such user pairs as a *pairing* to distinguish in language between the case of matching points and matching trajectories. We display the distribution of the number of matches found in Fig. 2. We note that while the importance of using alibis can be easily understood, not all previous work have utilized it. The probabilistic models in studies [4], [17] only deal with co-occurring events, disregarding the possibility of using such negative evidence to prune candidate matches; on the other hand the work of Cecaj et al. [16] perform a similar filtering (using the term "exclusion condition" for it), while Basık et al. [18] define the term alibi, explicitly test for the importance of such filtering and find that doing so significantly increases the quality of matching.

Algorithm 1 Basic algorithm to calculate matching user pairs. The input is the two datasets, sorted by user ID and time. The second dataset (the communication dataset C in this case) is also stored in an index which allows quickly finding points spatiotemporally close to a given query, described in the main text in more detail. The main loop selects candidate matching pairs based on spatiotemporal proximity which are then evaluated using the more detailed `CompareUsers` function, listed separately as Algorithm 2. Performing this comparison in this separate function is necessary so that we can ensure that each point from either trajectory is considered at most once as a match. Impossible matches for each user are kept track in the A set, while potential matches are kept track in the M result set. The M set can be pruned after processing each user to only retain the top match (or top few matches), limiting the size of the output. Computational complexity scales with the number of candidate point pairs found, but the M and A sets ensure that the `CompareUsers` function is called at most once for each candidate user pair. As the spatial search is implemented by a simple index lookup, its complexity scales with the total number of points found.

```
T = {i, {x^T_ik}_{k=1}^{r_i^T}}_{i=1}^{n_T} transportation dataset
C = {i, {x^C_ik}_{k=1}^{r_i^C}}_{i=1}^{n_C} communication dataset
M = {empty set of match candidates}
for all i ∈ {1, 2, ... n_T} do
    A = {empty set of users from C with alibis}
    for all k ∈ {1, 2, ... r_i^T} do
        S = search for spatial matches around x_ik
        U_C = distinct users from S
        for all j ∈ U_C do
            if j ∉ A and (i, j) ∉ M then
                m = CompareUsers(i, j)
                if m is alibi then
                    add j to A
                else
                    add (i, j, m) to M
                end if
            end if
        end for
    end for
    optionally: only keep top few matches of i in M
end for
```

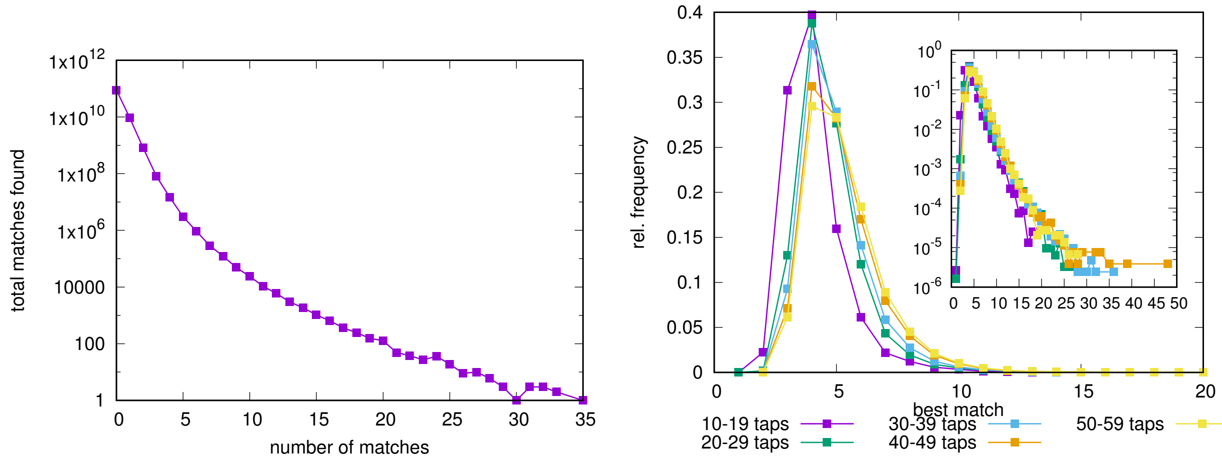

Figure 2. **Distribution of spatial matches.** Left: distribution of the number of spatial matches found between any two users in the dataset. Alibis are included as zero matches in accordance with the definition of P_s in Section 5). Right: Distribution of the best spatial match (i.e. user pairings with the maximum number of matches for each user) when selecting transportation users and searching for matching users in the CDR data, grouped by activity. Each distribution is normalized among the activity group. The inset shows the same distribution on a logarithmic scale.

Algorithm 2 The `CompareUsers` function from Algorithm 1. This function checks that the trajectories of users $i \in T$ and $j \in C$ are compatible (contain no alibis) and counts the number of matching points. It employs the constraint that each point is counted at most once as a match. This is achieved by iterating over the points in time order in both trajectories and keeping track of which points have been already recorded as part of a match. The computational complexity scales with the number of temporal matches between the two trajectories, as all of these need to be checked for spatial consistency. Note that in practice, this function will receive records that are already sorted by time, so a sort step is not necessary and is included only for the clarity of presentation.

function COMPAREUSERS(i,j)
  x_{ik}^T records of user i in T
  x_{jk}^C records of user j in C
  ensure that x_{ik}^T and x_{jk}^C are sorted by time
  $M = \{$ empty set for matched points from C $\}$
  $m = 0$ result: number of matches
  for all $k \in \{1, 2, \dots r_i^T\}$ **do**
    $m_k =$ `False`
    for all l s.t. x_{ik}^T and x_{jl}^C are a temporal match **do**
      if x_{ik}^T and x_{jl}^C are spatially inconsistent **then**
        return alibi
      else if $m_k =$ `False` and $x_{jl}^C \notin M$ **then**
        $m = m + 1$
        $m_k =$ `True`
        add x_{jl}^C to M
      end if
    end for
  end for
  return m
end function

5 Estimating Matchability

Having presented a matching methodology, we proceed with estimating its expected rate of success based on considerations of the statistical properties of matches. The main question we evaluate in the rest of the paper is whether a user pairing produced by the matching algorithm really corresponds to trajectories of the same person or if it is a false positive, i.e. two people who happened to appear in the dataset together m_{ij} times at random.

We proceed in three steps: first, we estimate the probability distribution of having a certain number of temporal matches between the records of any two users in the two datasets. Second, we use this as a basis for estimating the probability of an individual having a certain number of true positive matches among their two trajectories in the two datasets. Finally, we use the observed distribution of spatial matches to estimate the probability of obtaining more false positive matches randomly than the two number of positive matches estimated from the temporal match distribution.

5.1 Preliminary assumptions

We begin by defining probability distributions for obtaining a specific number of matches among user pairs from the two datasets and then estimating these from our data. We use the following notations:

- $P_t(m|i,j)$ is the probability distribution of users i and j having exactly m temporal matches in the data.
- $P_s(m|i,j)$ is the probability distribution of users i and j having exactly m spatial matches in the data and no impossible matches. We define $P_s(0|i,j)$ to include both the case when the two users have zero temporal matches and the case when they have > 0 temporal matches of which at least one is an alibi, i.e. spatially inconsistent.
- $P_s(m|i)$ is the probability distribution of user i having exactly m possible matches with *any* user in the other dataset.

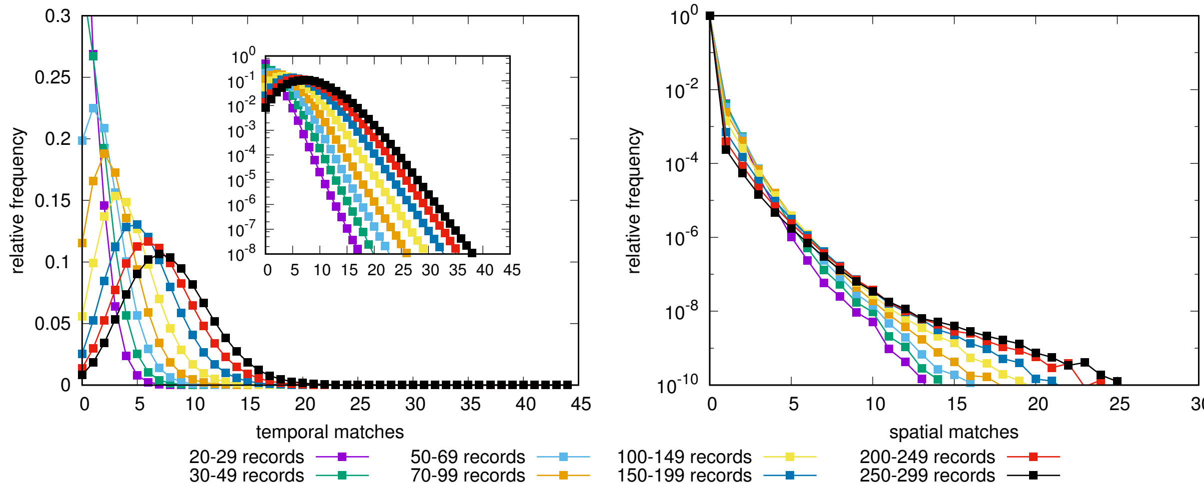

Figure 3. **Probability distributions for matches** compiled for LTA users with activity between 40 and 49 taps and Singtel users grouped by the number of calls. Left: distributions of temporal matches, i.e. $P_t(m|r_i, r_j)$, which we use to estimate the probability of a person having a certain number of real matches among their records in the two datasets; the inset shows the same distributions with a logarithmic scale. Right: distributions of possible matches, i.e. $P_s(m|r_i, r_j)$, which we use to estimate the probability of getting false positive matches.

We note that for known trajectories, the exact number of matches can be calculated. In our case however, we do not know which trajectory belongs to which person, thus we will use empirical estimates of these distributions based on the counts of matches among groups of trajectories. To be able to do so and use these distributions to calculate reidentification probabilities we employ some assumptions. First, we assume that $P_t(m|i, j)$ depends neither on whether users $i \in T$ and $j \in C$ represent the same person, nor on whether the two trajectories in the data are spatially consistent (i.e. they could be alibis); thus we can use an empirical estimation of it based on the data without the need for ground truth on user pairings. We note that this assumption means that when considering the trajectories of the same person in the two datasets, $P_s(m|i, i) = P_t(m|i, i)$, as in this case, all matches are spatially consistent. Second, we group users together by activity and estimate P_s and P_t for each groups empirically, i.e. $P_t(m|i, j) = P_t(m|r_i, r_j)$, where r_i and r_j are the number of records we have about them in the dataset. This includes the assumption that matches are independently and identically distributed among any pair of users in these groups. To improve the statistics and limit the complexity, we use moderate-sized subgroups of user activities instead of estimating the distributions for every possible (r_i, r_j) pair. As an example, we display the obtained P_t and P_s distributions for transportation users between 40 and 49 records (between 20 and 25 trips) and several groups of mobile network users in Fig. 3. The P_t distribution of temporal matches shows that our dataset has the limitation that for the most typical combinations (3 trips per day and 5-10 calls or texts per day, resulting in about 42 taps and 35-70 CDR records in our data), the expected number of matches is still relatively small (1-5). As expected, the P_s distribution decreases rapidly as well, based on the high unicity in the data. We display average distributions (i.e. for transportation users between 40 and 49 records and any mobile network user) in Fig. 4. We further display the observed ratio of P_s/P_t, which we can interpret as the ratio of probabilities having a given number of spatially consistent matches when considering the trajectories of different people vs considering the trajectories of the same person.

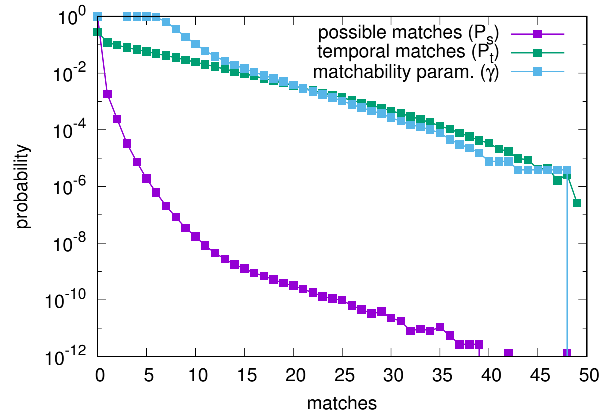

Figure 4. **Estimating expected success ratios.** Aggregated probability distributions of matches for transportation users with activity between 40 and 49 taps and any mobile network user, displayed along with the calculated matchability parameters calculated from the possible match distributions.

5.2 Matchability estimate

Using our previous assumptions, we estimate the probability of successfully reidentifying a user with given activities (r_1, r_2) in the two datasets. In accordance with the previous assumptions, we use $P_t(m|r_1, r_2)$ for estimating the probabilities of getting a certain number of real matches among the two traces: we thus assume that the real number of matches is m^* which is drawn from the probability distribution P_t. We then assume that the reidentification is successful if there is no other user with possible matches of m^* or more occurring randomly. Using our previous definitions, the possibility of this for one user is given by $\sum_{m' \geq m^*} P_s(m'|r_1)$. As this can occur for each user (a total of $n_C \approx 2.84M$ in our communication dataset), we need to calculate the probability that none of them has such a match,

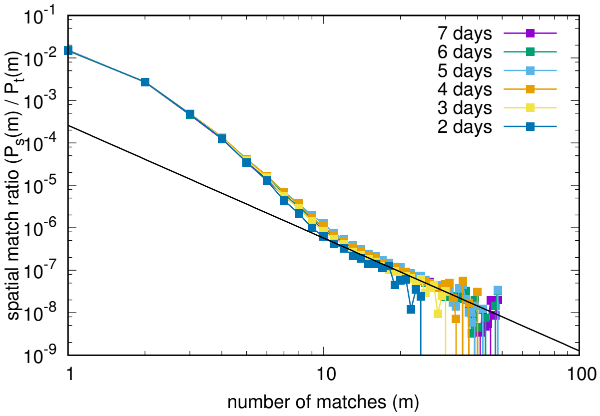

Figure 5. **Ratio of spatially consistent matches to temporal matches**, i.e. $P_s(m)/P_t(m)$. The data in this figure was created after including all user activities in our dataset, representing the distributions to be found among any pair of records of user trajectories. The different lines correspond to different subsets of the data; we can see that all of these behave quite similarly, supporting the assumption that the success ratios will also behave similarly regardless of the time period considered. The black line shows a power-law fit to the last section. The fitted function is given as $P_s(m)/P_t(m) = 8 \cdot 10^{-6} m^{-1.52}$.

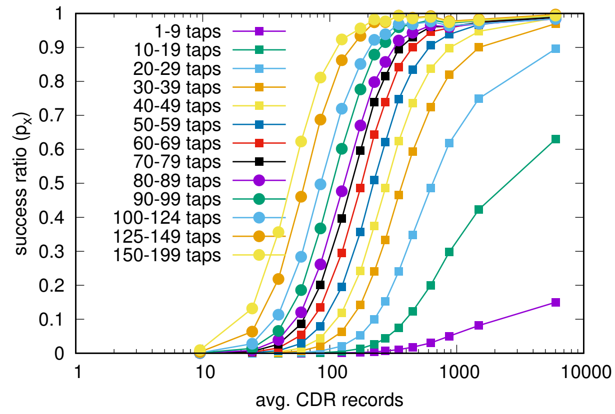

Figure 6. **Estimated success ratios** for the one week long dataset for different activity level of users. The x-axis corresponds to activity in the communications dataset, while the different lines correspond to different level of activity in the transportation dataset.

given by $\gamma(m^*|r_1) \equiv \left(1 - \sum_{m' \geq m^*} P_s(m'|r_1)\right)^{n_C}$ which we denote as the *matchability parameter* for a given user activity r_i and number of matches m^*. Since we consider the number of real matches (m^*) as a random variable as well, we then proceed by calculating the probability of successful reidentification, denoted by $p_x(r_1, r_2)$ as the expected value of γ:

$$p_x(r_1, r_2) = \langle \gamma(m|r_1) \rangle = \sum_m P_t(m|r_1, r_2) \gamma(m|r_1) =$$
$$= \sum_m P_t(m|r_1, r_2) \left(1 - \sum_{m' \geq m} P_s(m'|r_1)\right)^{n_C} \quad (4)$$

We note that in practice, we do not know the real values of the (r_1, r_2) pair of activities. Instead, we calculate $p_x(r_1, r_2)$ for different groups of r_1 and r_2 values and then calculate weighted averages assuming the r_1 and r_2 are selected independently random from the empirical distribution of activities in our data.

We display results among different groups of activity in Fig. 6 and also in Table S1 in the Supplementary Material. We see that high success ratios require relatively high activity; activities which we might consider typical, e.g. between 30 and 39 or 40 and 49 records in the transportation data (corresponding to 2-4 trips every day; recall that most trip results in two records in our dataset) and between 150 and 199 records in the phone data (one call or text per hour on average) only lead to $14\% - 24\%$ success rates. Weighted average of estimated success rates of user groups with these activity (i.e. with any random activity level in the other dataset) are in a similar range, while weighted average success ratio for the whole dataset (i.e. choosing two trajectories from the two datasets randomly and assuming they would be a real match, i.e. belong to the same person) is only around 8.1%. We also see in Fig. 6 that success ratios sharply increase as the number of records in the trajectories increase. Leaving out very inactive users from the averages (people with less than 10 transportation events and less than 20 mobile communication records) increases the weighted average success ratio to 14.8% already. Looking at users with very high activity in the communication dataset having over 1000 records and still focusing on typical transportation users whose weekly number of records is between 30 and 49, our estimation yields success ratios of over 90%. We note that only considering calls and text messages, these number of activities can be considered unrealistically high for typical people. Such frequencies could be achieved by using detailed CDRs including data communication and cell handover information, suggesting that achieving such success rates could be feasible with data that is currently already available to operators.

To better characterize how matchability depends on activities, in Fig. 7 we display success ratios among all user pair groups as the function of the expected number of temporal matches among those groups, i.e. $< P_t(m|r_i, r_j) >$. We see that all different cases follow the same relation, suggesting that the number of expected matches (or the number of real matches in case of looking at the representation of a specific person in the data) is the variable determining matchability. Using this as a working hypothesis, we can extrapolate to longer time intervals by employing the assumption that the dependence of success ratios on the expected number of matches does not change significantly with the time interval considered. We test this by performing the same analysis for shorter sub-intervals of the one week data that is available for us, and displaying the dependence of success ratios on the average number of temporal matches among different groups of users in Fig. 7. We see that for all cases with intervals ranging from 2 to 7 days, we have a very similar behavior, with somewhat larger variations for shorter time intervals as it is expected as a consequence of more limited amount of available data. This relation can be fitted well with the analytical form of $p_x(r_i, r_j) = \frac{1}{1 + A\bar{m}^{-b}}$, where $\bar{m} \equiv < P_t(m|r_i, r_j) >$ is the expected number of matches and the fitted parameters are $b = 2.993$ and $A = 434.69$. For high values of $\bar{m}$, we use

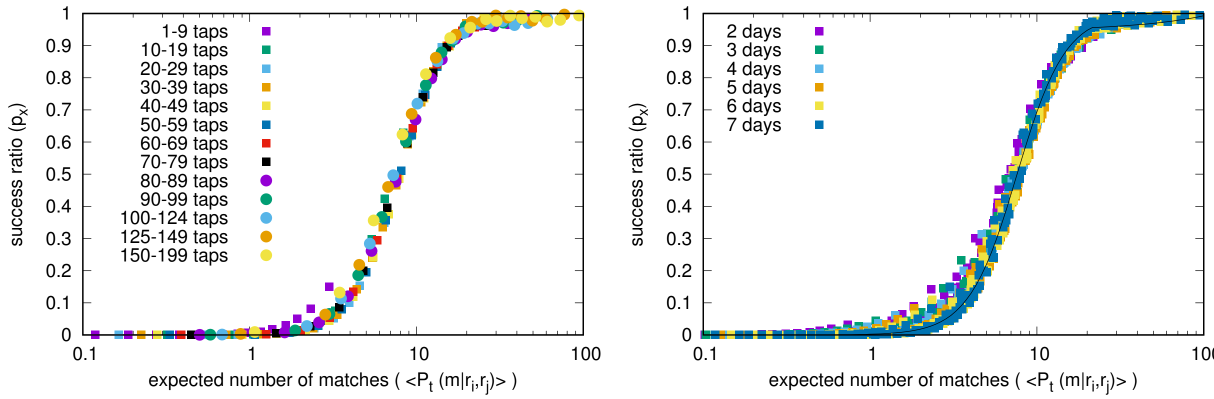

Figure 7. **Estimated success ratios as a function of the expected number of matches.** The expected number of matches was estimated from the data, as the average number of temporal matches found between the different user groupd. Left: data is displayed for the whole week long dataset, colors and shape of points correspond to different activity in the transportation dataset. We see that all of them follow the same relation approximately. Right: data is displayed for the one week long dataset along with subsets of shorter time intervals and colors correspond to different time intervals (e.g. the dark blue points include all data from the left panel). Again, we see that all different intervals follow approximately the same relation, allowing us to extrapolate to longer time intervals where P_s is not known, only the expected number of matches, $\bar{m} \equiv < P_t(m|r_i, r_j) >$ can be extrapolated. The black line displays a fitted analytical function we use for this as explained in the main text.

a more conservative estimate, given as a linear function: $p_x(r_i, r_j) = 4.66 \cdot 10^{-4}\bar{m} + 0.946$ for $\bar{m} > 21.09$, so as not to overestimate success ratios in this area where points in the one week long dataset are more scattered. We note that this choice does not alter the end results significantly. Furthermore, to show that the expected number of false positive matches will not increase significantly as we increase the observation interval, in Fig. 5, we show the ratio of spatially consistent and total temporal matches among any two users in the dataset for these different time intervals. We see that this behavior is very similar for different time intervals as well, with a sharp decrease in spatially consistent trajectories as the number of matches increases, suggesting that as the expected number of matches among users increases, the number of false positive matches will continue to decrease. This supports our extrapolation methodology by helping to establish that it will not overestimate success ratios.

We then perform the extrapolation by assuming that the expected number of temporal matches among the groups scales linearly with time (which essentially corresponds to a convolution of the P_t distributions for the longer time intervals), and interpolate the expected success ratios as a function of the expected number of temporal matches using the previous simple functional form which fits the data well. Based on this, we calculate similar measures as in the case of the original one week long dataset up to four weeks; we display the individual values among different groups of users in Tables S2, S3 and S4 in the Supplementary Material, while summarize the results in Table 1.

We can make several observations based on these results which we can use to project the possibility of reidentification in several different scenarios in terms of data collection methodology and data density. While readers are encouraged to look into Tables S1–S4 in the Supplementary Material for more insights, here we summarize the cases we find most relevant:

1) *Matching transportation and CDR data.*: This is essentially the case of datasets we have at hand, and here we focus on regular transportation users (people with 30 – 39 taps per week, which corresponds to 2 – 3 trips taken per day). If we match this with a typical number of phone calls and messages (5 – 10 per day, resulting in 30 – 69 records per week), success ratios are generally low, $< 1\%$ for the one week long dataset, and only increase to between $18\% to 43\%$ when considering a four week long period. Nevertheless, if phone activity increases to between 21 – 28 records per day (150 – 199 records per week, i.e. one record per hour on average), success of matching is already 14% for one week and reaches over 92% for a four week data collection interval. We emphasize that about 34% of all phone users in our dataset have at least 150 records per week, suggesting that such considerations are reasonable for a significant portion of the population. Furthermore, allowing even longer data collection intervals would result in identifying even less active users as well. In the case of people who we consider typical transportation users and moderate phone users (30 – 39 records per week in the transportation dataset and 50 – 69 records per week in the mobile phone dataset), we estimate that after 11 weeks, success rate would reach around 95%.

2) *Matching transportation or similar dataset with detailed CDR or GPS traces.*: With the proliferation of smartphones and data connections, people generate much higher activity in the mobile network than it used to be the case. Even when not actively using a smartphone, apps running in the background periodically check for updates, generating data traffic which is logged by the network. Furthermore, many apps record location periodically (as reported by the phone based either on GPS or wireless signal) and report it to the app developers. Both cases allow a much higher quality reconstruction of the people's trajectory during the day. Using a conservative estimate, we can expect this case to correspond to data collection with at least one point per hour, yielding a similar result as for people considered "active" phone users in the previous case, i.e. an expected success of matching with typical transportation users of at least 14% for one week and over 92% for a

four weeks. Making similar estimates with doubled data collection rate (once per half an hour, or about 300 – 399 records per week), we have an expected success rate of 46% for only one week already, almost 90% for two weeks and over 95% for three or four weeks. We note that these are still conservative estimates, as both network operators and app developers can easily detect users' movements and implement adaptive data collection, allowing them to reconstruct trajectories in good quality with less data points. We further note that beside transportation data, credit and debit card usage can generate similar amount of records in developed countries where a major portion of payments is made electronically, as well as geo-tagged social media posts of people actively maintaining a presence on microblogging services.

3) *Matching two datasets with increased density.*: Based on the previous discussion that data traffic and smartphone apps running in the background can easily generate at least one record per hour (i.e. between 150 and 199 records per week), we can argue that having access to two such datasets should be considered possible. Looking at results between two groups both having between 150 and 199 records per week, we see that the expected success ratio is already 95.6% even for only one week of data collection. Assuming somewhat fewer records (between 100 and 124), success ratio for one week is 72% and reaches almost 95% already after two weeks, establishing that the data collection procedures easily implementable for any smartphone app developer already generate data which allows reidentification only after a very short data collection interval.

We emphasize that the main basis for matchability is that the probability of a temporal match with *one* user (P_t, which we then consider to describe the distribution of real matches) should be higher than the probability of a random spatial match occurring with *any* of the ~ 2.8 million users in the other dataset (described by the γ matchability parameter estimated from P_s). Whether this holds true depends on the statistics of the dataset and estimating these correctly requires that the mobility traces of all of the affected population be present so as to be able to obtain a realistic estimate of false positive spatial matches. In our case, Figs. 3 and 4 illustrate that the ratio of probabilities of spatial and temporal matches decreases to small values quickly and that these probabilities are indeed comparable.

6 Discussion

In this paper we considered the problem of *matching* mobility datasets on a realistic scale in an urban setting. We developed methodology for handling the problem of comparing users' trajectories on the scale of several hundred million records of several million users and applied our solution for a dataset of one week of mobility traces of mobile communications and transportation logs from Singapore. We presented an empirical framework for calculating the estimated success ratio of matching users based on co-locations. Our results suggest that trajectories of people who have typical transportation usage patterns and are relatively active phone users could be matched based on a few weeks of data, while matching two datasets where data is collected more regularly (i.e. having one record per hour on average, which is easily achieved by network operators logging data communications or smartphone applications regularly querying device location) can be easily possible based on only one week of data collection. As the trend of collecting spatial traces of people continues with many service providers, we expect the possibility of matching people in anonymized datasets based on their trajectories to become increasingly easy in the near future.

Comparing our work to previous studies, we believe that this is the first study which tried to estimate the chances of successfully matching trajectories involving two large-scale dataset (on the order of millions of people in each) from realistic sources. We believe that using data that covers a significant portion of the population is important for this kind of work as data density is a main contributing factor to the success of matching, as with larger datasets we can expect significantly more false positive matches. In our analysis, the probability of finding a certain number of false positive matches is represented by the probability P_s for one user pair and with the γ matchability parameter for the whole dataset. While we saw that P_s is a quickly decreasing function of the required number of matches, γ depends exponentially on the number of users n_C and thus the density of data, meaning that the probability of finding false positive matches can drastically differ between datasets of different density. Furthermore, most of the methods employed for calculating matches in previous work (e.g. in Refs. [17], [38]) employ a metric which needs to be calculated for each possible pair of users, thus being prohibitably computationally intensive in the case of realistic data densities as they result in $O(n_T \times n_C)$ complexity. For the aforementioned reasons, we believe that scaling down our methods for the smaller data sizes used in these works is not expected to give realistic estimates. This way, a more direct comparison of our work with these papers is not easy to achieve.

The work of Cao et al. [4] presents a computationally efficient implementation, but disregards information about temporal matches, thus actually considering a much simpler version of the problem than what we focus on. The work of Cecaj et al. [16] is more similar to ours, but is also applied to a more limited data source, limiting the potential statistical analysis. We believe that our work is most similar to that of Basık et al. [18] who describe a very similar matching procedure augmented by an efficient spatial pre-filtering to reduce complexity. In our case, we believe such pre-filtering would not be applicable as our data comes from a compact but densely populated area, thus all users' trajectories would span the same units when constructing a coarse spatial index for such pre-filtering. The complexity of the main matching procedure is very similar; while they compute a different metric to evaluate potential candidates, enumerating matches is based on similar computational steps. There is a significant difference however in the primary focus of our current study and Ref. [18]. The authors in Ref. [18] primarily focus on evaluating the matching algorithm in terms of computational performance on real data, while they only test the quality of matching on synthetic data. On the contrary, our main focus is estimating the expected success of merging two real-world large scale datasets based on the statistical analysis of the matches found among them. To this end, we used a simpler definition of matching

	success ratio (all groups)	success ratio without inactive users	success ratio for transportation users with 30–49 records	success ratio for transportation users with 30–49 records without inactive CDR users
1 week	0.0805	0.1484	0.1677	0.2181
2 weeks	0.1657	0.3062	0.3718	0.4882
3 weeks	0.2225	0.41	0.4827	0.6348
4 weeks	0.2588	0.4766	0.551	0.7248

Table 1

Average expected success ratios for the study data and extrapolated to longer intervals. The left column shows results from any possible combination of activity pairings (i.e. assuming that the number records corresponding to a person's activities in the two datasets are randomly selected among all possible users activity levels in the two dataset). The right column shows averages with leaving out users with very low activities (less than 10 taps in the case of the transportation dataset and less than 20 records in the case of the mobile communication data) and also users with unrealistically high activities (125 or more taps in the transportation dataset or 2000 or more records in the mobile phone dataset).

which allows the derivation of probabilities for finding false positive matches which we believe is highly determinant in matching perfomance. The main difference is that the authors of Ref. [18] not only count the number of matches, but also the number of distinct locations such matches occur at. In our framework, that would mean an additional random variable for the P_s probability distribution, which would make the empirical estimation much more unreliable. This way, we believe our results represent only a lower bound on the success rates which could be improved upon if training on ground truth data becomes possible. Thus our results are not directly comparable to the matching quality quality metrics in Ref. [18], since those are influenced by the procedure used to generate the synthetic dataset.

Future work can extend the matching procedure employed, i.e. instead of just selecting the candidate with the highest number of matches, a more sophisticated approach could take into account the uneven nature of urban movements and calculate for each matching pair of points a weight representing the importance of it (e.g. a match at a crowded subway station could be considered less important than a match at a remote bus stop) or take into account the number of distinct places trajectories are matched similar to Ref. [18]. We believe such refinements would require at least some ground truth data such that quality of matches can be readily evaluated and optimized. Our matchability estimate could be adapted to a such scenario as long as statistics of matching points can be calculated similarly to the P_t and P_s distributions, i.e. assuming independence and identical distributions for large subsets of users allowing empirical estimate of such distributions. Using a dataset with available ground truth information could also be utilized to calculate measures of individual matchability and establish a connection with the entropy and predictability measures defined in previous work of Song et al. [36].

We believe that the possibility of matching mobility traces can open up new potential for understanding urban human mobility and providing better services for urban residents. Utilizing this while also providing adequate guarantees of privacy of the affected individuals should be in the focus of future interdisciplinary research including urban planning, algorithmic, security and legal perspectives.

Acknowledgements

The authors thank Ali Farzanehfar and Arnaud Tournier for reading the manuscript and providing valuable comments and suggestions.

The authors thank all sponsors and partners of the MIT Senseable City Laboratory including Allianz, the Amsterdam Institute for Advanced Metropolitan Solutions, the Fraunhofer Institute, Kuwait-MIT Center for Natural Resources and the Environment, Singapore-MIT Alliance for Research and Technology (SMART) and all the members of the Consortium.

References

[1] V. D. Blondel, A. Decuyper, and G. Krings, "A survey of results on mobile phone datasets analysis," *EPJ Data Science*, vol. 4, 2015.

[2] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transportation Research Part C*, vol. 19, no. 4, pp. 557–568, 2011.

[3] S. Sobolevsky, I. Sitko, R. Tachet Des Combes, B. Hawelka, J. M. Arias, and C. Ratti, "Cities through the Prism of People's Spending Behavior," *PLoS ONE*, vol. 11, no. 2, p. e0146291, 2016.

[4] W. Cao, Z. Wu, D. Wang, J. Li, and H. Wu, "Automatic user identification method across heterogeneous mobility data sources," in *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*. IEEE, 2016, pp. 978–989.

[5] M. Lenormand, M. Picornell, O. G. Cantú-Ros, A. Tugores, T. Louail, R. Herranz, M. Barthélemy, E. Frias-Martinez, and J. J. Ramasco, "Cross-checking different sources of mobility information," *PLOS ONE*, vol. 9, no. 8, p. e105184, 2014.

[6] M. Lenormand, T. Louail, O. G. Cantú-Ros, M. Picornell, R. Herranz, J. M. Arias, M. Barthelemy, M. S. Miguel, and J. J. Ramasco, "Influence of sociodemographic characteristics on human mobility," *Scientific Reports*, vol. 5, p. 10075, 2015.

[7] G. Goulet Langlois, H. N. Koutsopoulos, and J. Zhao, "Inferring patterns in the multi-week activity sequences of public transport users," *Transportation Research Part C: Emerging Technologies*, vol. 64, pp. 1–16, 2016.

[8] J. L. Toole, Y.-A. de Montjoye, M. C. González, and A. Pentland, "Modeling and Understanding Intrinsic Characteristics of Human Mobility," in *Social Phenomena*, B. Gonçalves and N. Perra, Eds. Cham: Springer International Publishing, 2015, pp. 15–35.

[9] L. Taylor, "No place to hide? The ethics and analytics of tracking mobility using mobile phone data," *Environment and Planning D: Society and Space*, vol. 34, no. 2, pp. 319–336, 2016.

[10] S. A. Golder and M. W. Macy, "Digital Footprints: Opportunities and Challenges for Online Social Research," *Annu. Rev. Sociol*, vol. 40, pp. 129–52, 2014.

[11] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the Crowd: The privacy bounds of human mobility," *Scientific Reports*, vol. 3, p. 1376, 2013.

[12] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. S. Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, no. 6221, pp. 536–539, 2015.

[13] S. Milgram, "The Experience of Living in Cities," *Science*, vol. 167, no. September, pp. 1461–1468, 1970.

[14] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg, "Inferring social ties from geographic coincidences," *Proceedings of the National Academy of Sciences*, vol. 107, no. 52, pp. 22 436–22 441, 2010.

[15] P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[16] A. Cecaj, M. Mamei, and F. Zambonelli, “Re-identification and information fusion between anonymized CDR and social network data,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 7, p. 83, 2015.

[17] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi, “Linking users across domains with location data: theory and validation,” *Proceedings of the 25th International Conference on World Wide Web*, pp. 707–719, 2016.

[18] F. Basık, B. Gedik, Ç. Etemoğlu, and H. Ferhatosmanoğlu, “Spatio-Temporal Linkage over Location-Enhanced Services,” *IEEE Transactions on Mobile Computing*, vol. 17, no. 2, pp. 447–460, 2017.

[19] I. P. Fellegi and A. B. Sunter, “A theory for record linkage,” *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1183–1210, 1969.

[20] I. Fellegi, “On the question of statistical confidentiality,” *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 7–18, 1972.

[21] T. Dalenius, “Towards a methodology for statistical disclosure control,” *Statistik Tidskrift*, vol. 15, pp. 429–444, 1977.

[22] D. Dobkin and A. K. Jones, “Secure Databases: Protection Against User Influence,” *ACM Transactions on Database Systems*, vol. 4, no. 1, pp. 97–106, 1979.

[23] D. E. Denning and J. Schlörer, “A fast procedure for finding a tracker in a statistical database,” *ACM Transactions on Database Systems*, vol. 5, no. 1, pp. 88–102, 1980.

[24] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, “Measurement and analysis of online social networks,” *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement - IMC '07*, pp. 29–42, 2007.

[25] L. Sweeney, “k-anonymity: a model for protecting privacy,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, p. 557, 2002.

[26] ——, “Achieving k-anonymity privacy protection using generalization and suppression,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, p. 571, 2002.

[27] A. Narayanan and V. Shmatikov, “Robust De-anonymization of Large Datasets (How To Break Anonymity of the Netflix Prize Dataset),” in *2008 IEEE Symposium on Security and Privacy*. IEEE, 2008, pp. 111–125.

[28] ——, “De-anonymizing Social Networks,” in *30th IEEE Symposium on Security and Privacy*. IEEE, 2009, pp. 173–187.

[29] L. Zhang, S. Jajodia, and A. Brodsky, “Information disclosure under realistic assumptions: Privacy versus optimality,” in *Computer and Communications Security*, 2007, pp. 573–583.

[30] C. Dwork, “Differential Privacy,” *Proceedings of the International Colloquium on Automata, Languages and Programming, Part II (ICALP)*, pp. 1–12, 2006.

[31] A. Biryukov, D. Khovratovich, and I. Pustogarov, “Deanonymisation of clients in Bitcoin P2P network,” in *ACM SIGSAC*, 2014, pp. 15–29.

[32] P. L. Juhász, J. Stéger, D. Kondor, and G. Vattay, “A bayesian approach to identify Bitcoin users,” *arXiv preprint*, p. arXiv:1612.06747, 2016.

[33] P. J. Green and K. V. Mardia, “Bayesian alignment using hierarchical models, with applications in protein bioinformatics,” *Biometrika*, vol. 93, no. 2, pp. 235–254, 2006.

[34] T. Budavári and T. J. Loredo, “Probabilistic record linkage in astronomy: Directional cross-identification and beyond,” *Annual Review of Statistics and Its Application*, vol. 2, no. 1, pp. 113–139, 2015.

[35] Y. De Mulder, G. Danezis, L. Batina, and B. Preneel, “Identification via location-profiling in GSM networks,” in *WPES 08*, 2008, pp. 23–32.

[36] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, “Limits of Predictability in Human Mobility,” *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[37] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell, “NextPlace: A spatio-temporal prediction framework for pervasive systems,” in *Pervasive Computing. Pervasive 2011. Lecture Notes in Computer Science*, K. Lyons, J. Hightower, and E. Huang, Eds., vol. 6696. Springer, 2011, pp. 152–169.

[38] S. Gambs, M. O. Killijian, and M. Núñez Del Prado Cortez, “De-anonymization attack on geolocated data,” *Journal of Computer and System Sciences*, vol. 80, no. 8, pp. 1597–1614, 2014.

[39] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W. Ma, “Mining user similarity based on location history,” *ACM SIGSPATIAL 08*, no. c, p. 34, 2008.

[40] X. He, G. Cormode, A. Machanavajjhala, C. M. Procopiuc, and D. Srivastava, “DPT: Differentially Private Trajectory Synthesis Using Hierarchical Reference Systems,” *Proceedings of the VLDB Endowment*, vol. 8, no. 11, pp. 1154–1165, 2015.

[41] B. Malin and L. Sweeney, “How (not) to protect genomic data privacy in a distributed network: Using trail re-identification to evaluate and design anonymity protection systems,” *Journal of Biomedical Informatics*, vol. 37, no. 3, pp. 179–192, 2004.

[42] F. Basik, “Scalable Linkage across Location Enhanced Services,” in *VLDB*, 2017.

Dániel Kondor Daniel Kondor received his MSc and PhD degrees in physics from the Eötvös Loránd University in Budapest, Hungary. In the past two years he has been a postdoctoral researcher at the Senseable City Lab at MIT in Cambridge, MA USA, while he recently joined SMART FM as a postdoctoral associate. His research focused on social and economic networks during his PhD studies, working on data from Twitter and Bitcoin. During his time at SCL, he participated in several research projects focusing on urban human mobility and related topics, making use of cell phone, transportation and social media data.

Behrooz Hashemian Behrooz Hashemian, PhD, is researcher and chief data officer at Massachusetts Institute of Technology (MIT), Senseable City Lab. He investigates the innovative implementation of big data analytics and technology in smart cities, finance, and healthcare. He is a data scientist with expertise in developing predictive analytics strategies, machine learning solutions, and data-driven platforms for informed decision making. His work endeavors to bridge the gap between academic research and industrial deployment of big data analytics and internet of things.

Yves-Alexandre de Montjoye Yves-Alexandre de Montjoye is an Assistant Professor at Imperial College London's Data Science Institute and Department of Computing where he heads the computational privacy group. He obtained an M.Sc. from Louvain in Applied Mathematics, an M.Sc. (Centralien) from Ecole Centrale Paris, an M.Sc. from KULeuven in Mathematical Engineering as well as his B.Sc. in engineering at Louvain. He completed a PhD at MIT working with Alex (Sandy) Pentland at the Media Lab. Following graduation, Yves was a post-doctoral researcher at Harvard University working with Gary King and Latanya Sweeney. He was also a research scientist at the Media Lab working with Iyad Rahwan. He was also a researcher at the Santa Fe Institute, worked for the Boston Consulting Group and acted as an expert for both the Bill and Melinda Gates Foundation and the United Nations. His research aims at understanding how the unicity of human behavior impacts the privacy of individuals – through re-identification or inference – in large-scale metadata datasets such as mobile phone, credit cards, or browsing data.

Carlo Ratti An architect and engineer by training, Professor Carlo Ratti teaches at Massachusetts Institute of Technology, where he directs the Senseable City Lab, and is a founding partner of the international design office Carlo Ratti Associati. He graduated from the Politecnico di Torino and the École Nationale des Ponts et Chaussées in Paris, and later earned his MPhil and PhD at the University of Cambridge, UK. Prof. Ratti has co-authored over 500 publications and holds several patents. His work has been exhibited worldwide at venues such as the Venice Biennale, the Design Museum in Barcelona, the Science Museum in London, and The Museum of Modern Art in New York. Prof Ratti is currently serving as both a member of the World Economic Forum Global Future Council on Cities and special adviser on Urban Innovation to the European Commission.