Privacy in trajectory micro-data publishing: a survey

Marco Fiore, Panagiota Katsikouli, Elli Zavou, Mathieu Cunche, Françoise Fessant, Dominique Le Hello, Ulrich Matchi Aivodji, Baptiste Olivier, Tony Quertier, Razvan Stanica

TRANSACTIONS ON DATA PRIVACY 13 (2020) 91–149

Privacy in trajectory micro-data publishing: a survey

Marco Fiore*, Panagiota Katsikouli, Elli Zavou***, Mathieu Cunche***, Françoise Fessant****, Dominique Le Hello****, Ulrich Matchi Aivodji*****, Baptiste Olivier****, Tony Quertier****, Razvan Stanica*****

*IMDEA Networks Institute, Spain.

**Technical University of Denmark, Lyngby, Denmark.

***University of Lyon, Inria, INSA-Lyon, CITI, France.

**** Orange Labs,France.

*****Universite du Quebec à Montreal, Canada.

E-mails: marco.fiore@ieiit.cnr.it, panka@dtu.dk, elli.zavou@inria.fr, mathieu.cunche@insa-lyon.fr, francoise.fessant@orange.com, dominique.le.hello@orange.com, aivodji.ulrich@courrier.uqam.ca, baptiste.olivier@orange.com, tony.quertier@orange.com, razvan.stanica@insa-lyon.fr

Received 30 April 2019; received in revised form 26 August 2019 and 16 January 2020; accepted 8 April 2020

Abstract. We survey the literature on the privacy of trajectory micro-data, *i.e.*, spatiotemporal information about the mobility of individuals, whose collection is becoming increasingly simple and frequent thanks to emerging information and communication technologies. The focus of our review is on privacy-preserving data publishing (PPDP), *i.e.*, the publication of databases of trajectory micro-data that preserve the privacy of the monitored individuals. We classify and present the literature of attacks against trajectory micro-data, as well as solutions proposed to date for protecting databases from such attacks. This paper serves as an introductory reading on a critical subject in an era of growing awareness about privacy risks connected to digital services, and provides insights into open problems and future directions for research.

Keywords. Privacy, Trajectory micro-data, Positioning data, Personal data, Data publishing, Re-identification, Pseudonymization, Anonymization

1 Introduction

Our lives are increasingly entangled with ubiquitous communication technologies. Calling someone on a mobile phone, tweeting about an event, browsing the World Wide Web, using a car navigation system, or paying with a credit card are a few examples of situations that create a seamless trail of *digital breadcrumbs* about our daily activities. These actions are easily recorded and persistently stored into databases. Today, the pervasiveness of mobile

Identifiers	Spatiotemporal points *[with non-positioning information]*					Attributes		
Alice +39 320 191 7047	45.061679, 7.677888 2018/01/24 08:05	45.062518, 7.662191 2018/01/24 10:32	45.062288, 7.671960 2018/01/24 14:18	45.058935, 7.686642 2018/01/24 19:41	45.070908, 7.684926 2018/01/24 22:01	Female	Accountant	€ 36,000
Bob +39 339 205 3011	45.068962, 7.698691 2018/01/24 07:10	45.068962, 7.698691 2018/01/24 07:13	45.067780, 7.694743 2018/01/24 09:20	45.079630, 7.671697 2018/01/24 16:18	45.081024, 7.625563 2018/01/24 16:53	Male	Engineer	€ 74,000
Charlie +39 347 772 3345	45.068962, 7.698691 2018/01/24 07:10	45.068962, 7.698691 2018/01/24 07:13	45.033696, 7.675753 2018/01/24 11:53	45.040004, 7.676439 2018/01/24 11:54	45.081024, 7.625563 2018/01/24 16:53	Male	Lawyer	€ 74,000
Dave +39 328 055 4606	45.094756, 7.526836 2018/01/24 12:26	45.090878, 7.528896 2018/01/24 13:13	45.066634, 7.515850 2018/01/24 15:17	45.063240, 7.522717 2018/01/24 17:48	45.092090, 7.524776 2018/01/24 19:33	Female	Consultant	€ 103,000
Erin +39 348 223 1098	45.135216, 7.760983 2018/01/24 07:11 *[incoming SMS]*	45.008652, 7.532330 2018/01/25 03:44 *[outgoing call]*	45.109635, 7.640991 2018/01/24 20:32 *[location area update]*	45.109635, 7.640991 2018/01/24 20:58 *[outgoing SMS]*	45.105176, 7.641850 2018/01/24 21:48 *[outgoing call]*	Male	Plumber	€ 31,000
Frank +39 333 879 4903	45.064937, 7.641850 2018/01/24 23:29 *[outgoing call]*	45.008652, 7.532330 2018/01/25 03:44 *[incoming SMS]*	45.008652, 7.532330 2018/01/24 22:48 *[outgoing SMS]*	45.004768, 7.535076 2018/01/24 23:35 *[incoming SMS]*	45.003069, 7.532673 2018/01/25 00:42 *[incoming SMS]*	NA	NA	NA

Figure 1: Example of database of trajectory micro-data. Each record is composed of an identifier (left), a spatiotemporal trajectory (middle), and additional attributes (right). In this specific example, the person's name and phone address are the identifiers, and spatiotemporal points in the trajectory are GPS locations augmented with non-positioning information, within brackets, about their mobile communication activity. Attributes consist of gender, employment and revenue.

communication technologies allows tracking millions of users simultaneously, leading to the collection of vast amounts of personal mobility data, which are then mined for many and varied purposes, such as location-based marketing, targeted advertising, behavioural profiling, transportation analysis, liability attribution, or security enforcement – just to cite a few relevant applications.

The galloping pace of innovations in this field, along with the increasing trend of digitalization of our lives, suggests that what we are experiencing nowadays is just the tip of the iceberg. In fact, services based on personal data records promise to be life-changers for the newer generations, with a clear trend of innovation happening in the data domain, where social networks (and alike) collect and exploit users' information more and more [106].

A common trait to most of these emerging technologies is that they often build and rely on databases that compose of or include *trajectory micro-data*. As the term indicates, these are micro-data, *i.e.*, information about single individuals, that describe their *spatiotemporal trajectories*, *i.e.*, sequences of geographical positions of the monitored individuals over time. Figure 1 shows a toy example of a typical trajectory micro-data database: each record corresponds to one person, and contains an identifier as well as a set of geo-referenced and time-stamped elements, or *spatiotemporal points*. Depending on the nature of the database, the elements can also include non-positioning (*e.g.*, numerical or categorical) information associated to each spatiotemporal point. Also, the database can present additional fields that map to attributes beyond the spatiotemporal trajectory.

The definition of trajectory micro-data database above encompasses positioning information gathered in a variety of ways, via different platforms and technologies. For the sake of clarity, we illustrate below five prominent examples of trajectory micro-data sources.

- Location-based services (LBS) are implemented as applications running on mobile devices (*e.g.*, smartphones or tablets), which upload user position data as required for service operation. Many extremely popular applications, such as Google Maps, FourSquare, Twitter, Instagram, or Pokemon Go, fall in this category, and relentlessly capture trajectory micro-data of individuals.

- Cellular network operators deploy passive monitoring systems in their networks to collect data about their subscribers' activity, for purposes including billing, traffic engineering or added-value service development. Such data include time-stamped user locations (*e.g.*, the location of the antenna which the user device is associated to, or a triangulated point from signal strength indicators). For instance, research-

favoured call detail records (CDR) allow tracking mobile subscribers every time their devices interact with the network.

- Mobile devices equipped with Wi-Fi interfaces typically broadcast probe messages to discover nearby Access Points (APs). By letting APs (or sniffers, *i.e.*, dedicated devices that passively monitor probe messages) record the unique Medium Access Control (MAC) address of the devices emitting such probes, the Wi-Fi access provider can track users within coverage of the Wi-Fi network. In presence of large deployments, *e.g.*, covering municipalities or urban transportation infrastructures, mobile devices can be potentially followed across a vast portion of their movements.

- Modern car navigation systems have Internet connection capability, thanks to an embedded mobile network interface. This setup allows notifying drivers about road traffic conditions in real time, but also to collect fine-grained positioning data while the vehicle engine is on. Such data are used by navigation system providers to determine the congestion level of roads, and by insurance companies to determine liability in case of accidents or to profile driving styles and associated risk levels.

- Electronic payments are replacing cash in everyday's shopping. The resulting transactions are easily linked to the address of the retailer who accepted the payment, which allows companies in the banking sector to monitor the movements of their customers as they use their debit or credit cards.

In all examples above, new and pervasive technologies allow the collection of trajectory micro-data at very large scales, *i.e.*, enable tracking thousands to millions of users at once.

It is precisely the possibility of knowing the movement patterns of large populations at an individual level that paves the road to a wide range of applications for databases of trajectory micro-data. Such applications encompass a variety of contexts, including intelligent transportation, assisted-life services, city planning, location-based marketing, data-driven decision-making, or infrastructure optimization; as a recent and compelling example, targeted containment measures enacted by many countries during the COVID-19 outbreak in early 2020 have relied on contact information extracted from trajectory micro-data collected by companies in the telecommunication sector [77]. In the light of these considerations, network and service providers are obviously eager to exploit trajectory micro-data in new ways by continuously developing original dedicated platforms [37, 3], and financial forecasts expect trajectory micro-data-driven services to grow into a new multibillion-dollar market in the coming years [1, 44]. Ultimately, the emergence of such a large and innovative technological and business ecosystem creates a whole new need for gathering, storing, provisioning, circulating and trading databases of trajectory micro-data.

1.1 Privacy of trajectory micro-data

Privacy is an obvious major concern at all stages of trajectory micro-data manipulation. This consideration holds no matter whether the ultimate aim of the data processing is the discovery of new knowledge or the monetization of embedded information. As a matter of fact, owing to the nature of trajectory micro-data, incorrect stewardship can easily reveal sensitive personal information about the users. Examples include iOS devices storing their own spatiotemporal trajectories in unencrypted format and transmitting them to Apple [29, 28], US mobile carriers selling real-time personal trajectory micro-data to third party service providers [95], or demand-side platforms for targeted advertising in mobile phone apps

Pseudo-identifier	Spatiotemporal points *[with non-positioning information]*					Attributes		
2BD806C97F0E00	45.061679, 7.677888 2018/01/24 08:05	45.062518, 7.662191 2018/01/24 10:32	45.062288, 7.671960 2018/01/24 14:18	45.058935, 7.686642 2018/01/24 19:41	45.070908, 7.684926 2018/01/24 22:01	Female	Accountant	€ 36,000
81B637D8FCD2C6	45.068962, 7.698691 2018/01/24 07:10	45.068962, 7.698691 2018/01/24 07:13	45.067780, 7.694743 2018/01/24 09:20	45.079630, 7.671697 2018/01/24 16:18	45.081024, 7.625563 2018/01/24 16:53	Male	Engineer	€ 74,000
B9DD960C175345	45.068962, 7.698691 2018/01/24 07:10	45.068962, 7.698691 2018/01/24 07:13	45.033696, 7.675753 2018/01/24 11:53	45.040004, 7.676439 2018/01/24 11:54	45.081024, 7.625563 2018/01/24 16:53	Male	Lawyer	€ 74,000
7CBCCB0C4CAADF	45.094756, 7.526836 2018/01/24 12:26	45.090878, 7.528896 2018/01/24 13:13	45.066634, 7.515850 2018/01/24 15:17	45.063240, 7.522717 2018/01/24 17:48	45.092090, 7.524776 2018/01/24 19:33	Female	Consultant	€ 103,000
61EA0803F88535	45.135216, 7.760983 2018/01/24 07:11 *[incoming SMS]*	45.008652, 7.532330 2018/01/25 03:44 *[outgoing call]*	45.109635, 7.640991 2018/01/24 20:32 *[location area update]*	45.109635, 7.640991 2018/01/24 20:58 *[outgoing SMS]*	45.105176, 7.641850 2018/01/24 21:48 *[outgoing call]*	Male	Plumber	€ 31,000
77646F5A4F3166	45.064937, 7.641850 2018/01/24 23:29 *[outgoing call]*	45.008652, 7.532330 2018/01/25 03:44 *[incoming SMS]*	45.008652, 7.532330 2018/01/24 22:48 *[outgoing SMS]*	45.004768, 7.535076 2018/01/24 23:35 *[incoming SMS]*	45.003069, 7.532673 2018/01/25 00:42 *[incoming SMS]*	NA	NA	NA

Figure 2: Example of pseudonymisation of the database in Figure 1. Identifiers (left) are replaced with values that cannot be linked with the person's identity. The spatiotemporal trajectory and additional attributes remain unchanged.

paving the way to uncontrolled collection of personal trajectory micro-data [127]. This is also generating a growing concern in the general public, as awareness is raising about the privacy risks associated with spatiotemporal tracking [125], and about how such personal information is shared in the data market [126].

Situations such as those mentioned above call for *privacy-preserving data publishing* (PPDP) of trajectory micro-data databases in all contexts where this kind of data is stored or shared. PPDP recommends that databases should be transformed prior to publication in potentially hostile environments, so as to grant that the published data remains useful while individual privacy is preserved [52].

The common practice adopted by data collectors and data owners in order to protect the privacy of the individuals they monitor is *pseudonymisation*, also referred to as *de-personification*. This straightforward approach consists in removing all personal identifiers (*e.g.*, information that is directly linked to the person's identity, such as name, telephone number, precise address, plate number, etc.), and replacing them with some *pseudorandom identifier*; the latter can be a keyed hash of the original personal identifiers, or simply a random number that is uniquely associated to the trajectory micro-data of an actual individual. Figure 2 provides an example of pseudonymisation, for the database in Figure 1.

Unfortunately, pseudonymisation only provides a very mild level of protection. A number of experiments, performed in recent times and using large-scale real-world datasets, have repeatedly demonstrated the significant risks associated to pseudonymised trajectory micro-data. In particular, naive cross-correlation of pseudonymised data with named side information (obtained from, *e.g.*, public-access social network data) leads to *re-identification*, *i.e.*, disclosure of the identities of users with high probability, making pseudonymisation basically useless. We will discuss in detail the investigations leading to such conclusions in Section 2. What is relevant here is that, in the light of these findings and quite unsurprisingly, data controllers have nowadays become extremely cautious in opening access to pseudononymised trajectory micro-data. A prominent example is that of TfL, the transport regulator in London, UK, which recently ran a pilot experiment by tracking passengers in the London Underground network via Wi-Fi probes broadcasted by mobile devices. TfL later rejected a Freedom of Information (FOI) request to release the pseudonymised dataset, exactly because of the potential re-identification risks of the data [81].

With such a growth of concerns about risks associated with uncontrolled gathering and mining of trajectory micro-data, regulatory bodies have been working on new legal frameworks dedicated to personal data protection. A leading act in this sense is the General Data Protection Regulation (GDPR) [49], which became effective on May 2018 and applies to all European Union citizens. The GDPR enforces that data controllers shall adopt the best mea-

sures for data protection by design and by default. Such measures include pseudonymisation, as it can reduce the risks for the data subjects concerned and help controllers and processors to meet their data-protection obligations. However, the GDPR makes it very clear that pseudonymisation alone is an insufficient privacy measure when it comes to PPDP. Indeed, the regulation decrees that pseudonymised data has still to be treated as personal data, which must be securely stored and cannot be circulated freely. Instead, the GDPR lays down that a more open publication of data is allowed upon *anonymization*, a process which ensures that the data cannot be any longer linked to an identified or identifiable natural person or data subject. According to the GDPR, anonymized data is not personal anymore, hence is not concerned by the privacy-protection rules it defines.

Legislations such as the GDPR are thus an important part of the solution, as they make procedures ensuring a correct data processing mandatory. However, they must be complemented by sound technical solutions that implement the invoked "best measures" and achieve the privacy goals set by PPDP. In the specific case of trajectory micro-data, developing anonymization algorithms that provably prevent any re-identification or personal information inference from the original spatiotemporal points is extremely challenging. As it is often the case in presence of difficult tasks, the problem has drawn a substantial effort by the research community: a plethora of scientific papers have appeared over the past decade, aiming both at unveiling privacy risks connected with trajectory micro-data, and at proposing solutions to cope with such risks. However, such a large body of works targets heterogeneous types of trajectory micro-data, considers a variety of attacker models, relies on different privacy criteria, and uses disparate data transformation techniques. This substantial diversity makes the literature tangled and complicated to approach, raising questions about where the current state of the art actually stands.

1.2 Objective, positioning and structure of the survey

This survey serves as a comprehensive introduction to the domain of privacy of trajectory micro-data for PPDP. It summarizes almost two decades of research, providing a review of a large number of works that cover all aspects of the problem. These include the assessment of privacy risks in trajectory micro-data, the definition of attacks realizing such risks, and the proposition of solutions that protect user privacy from the aforementioned attacks.

Our survey joins a rather small family of reviews that previously explored similar domains. The early works by Decker [41] and Chow and Mokbel [34] review privacy in LBS: as we will explain in Section 1.3, this is an orthogonal problem with respect to that of PPDP of trajectory micro-data. Garfinkel [55] provides a general overview of personal information de-identification across a wide range of database types, including geographic and map ones. The portion of the work dedicated to trajectory micro-data is necessarily limited in such a holistic document, and does not cover the subject in depth. Haris *et al.* [89] discuss privacy leakages, associated risks and potential remedies in the broad context of mobile computing. Their review of the literature has a very wide breadth, which is fully orthogonal to ours: indeed, their work targets applications and services rather than data. Christin [35] focuses on mobile participatory sensing, discussing privacy threats relevant to the different phases of the sensing process. Mobile participatory sensing can generate in some cases trajectory micro-data, hence some reviewed works are also covered in our document. However, the overlap is marginal, and the overall context, discussion of challenges and conclusions by Christin [35] are related to data and processes of a different nature from those of interest to us. The overlap is also minimal with the short review by Al-Azizy *et al.* [8], who review the literature on data de-anonymisation: on the one hand, they con-

sider any class of data instead of the more specific trajectory micro-data; on the other hand, they focus on de-anonymisation, whereas we cover all aspects of the problem, from risk assessment to solutions for data protection. Similar considerations hold for the recent book edited by Gkoulalas-Divanis and Bettini [57], which provides a comprehensive overview of current challenges and solutions in mobile data privacy; the book covers a wide range of subjects, from data collection to management and analysis, but does not include a dedicated discussion on mobile data PPDP. The work that is the closest in spirit to ours is that by Bonchi *et al.* [19], who discuss a selected set of seven papers that propose techniques to anonymise trajectory micro-data: this is a small subset of the studies we survey, which cover a much larger literature on both attack and protection techniques.

As a result, none of the existing surveys provides a literature overview that comprehensively addresses trajectory micro-data privacy. This paper aims at closing this gap – a significant one in the light of the rapid emergence of real-world services that heavily rely on trajectory micro-data. The document is structured into two main Sections, respectively dedicated to *attacks* against trajectory micro-data, and *anonymisation* of trajectory micro-data. The former, in Section 2, reviews the body of works that assess the privacy risks associated with trajectory micro-data, by devising, implementing and evaluating attacks that allow re-identifying users in a trajectory micro-data database. The latter, in Section 3, surveys countermeasures proposed to protect trajectory micro-data from the aforementioned privacy threats. The contents of such Sections allows us to draw considerations, as well as present open issues and research opportunities in Section 4.

1.3 Remarks

Before proceeding further, the following three important remarks are in order.

First, as anticipated above, our focus is on privacy-preserving publication (PPDP) of trajectory micro-data, which is an entirely different problem than privacy in LBS. Indeed, the two scenarios entail non-comparable system models. In the case of trajectory micro-data publishing, databases of millions of records are mined offline, and the challenge is ensuring that their circulation does not pose a threat to user privacy, but retains data utility. In the case of LBS, single (geo-referenced and time-stamped) queries generated by mobile devices must be processed in real-time, and the objective is *location privacy*, *i.e.*, ensuring that such a process preserves users' privacy by preventing the service provider from locating users. These considerations make PPDP of large-scale datasets the relevant problem in the context of trajectory micro-data, while LBS are more concerned with the real-time anonymization of small sets of spatiotemporal points; ultimately, this difference entails attacker models and anonymization techniques that are very diverse for the two scenarios. Indeed, Xiao and Xiong [129] and Bindschaedler *et al.* [16] have shown that individual spatiotemporal points anonymized via solutions for location privacy are still vulnerable to attacks when their time-ordered sequence is considered, *i.e.*, when they are treated as a spatiotemporal trajectory. As our focus is on trajectory releasing, the vast body of literature on privacy of LBS is out of the scope of the present document; we refer readers with an interest on privacy in LBS to the dedicated surveys by Decker [41], Chow and Mokbel [34], and Bettini [12].

Second, consistently with the scope of the survey, we are interested in attacks and anonymization solutions that target trajectory micro-data. Therefore, we do not consider in our review attacks against metadata that contain spatiotemporal information but where such information is not factually exploited, as in the case of the mobile phone call graphs considered by Sharad and Danezis [108], or of the mobile subscriber communication history studied by Mayer *et al.* [85]. Similarly, we target individual trajectories intended as se-

quences of spatiotemporal points, while we do not consider aggregate forms of such data, like people counts or density. Hence, we do not review attacks against aggregate data from trajectories, such as those designed, *e.g.*, by Xu *et al.* [130]; and, we do not delve in the details of privacy-preserving techniques for aggregate statistics, such as those proposed by Liu *et al.* [80], or those briefly mentioned at the beginning of Section 3.4.

Third, in the remainder of the document we will not make any distinction between original and pseudonymised data, and just refer to both as trajectory micro-data. This is because pseudonymisation does not provide any significant layer of privacy protection, and the vast majority of studies assume that all their input data are already pseudonymised.

2 Attacks on trajectory micro-data

The first part of our survey is dedicated to the body of works on attacks against trajectory micro-data. The objective of these studies is to propose techniques that enable the re-identification or inference of personal data from datasets of trajectory micro-data, implicitly revealing privacy risks associated to the inconsiderate publication of this class of databases. To structure our discussion, we present in Section 2.1 an original taxonomy of attack strategies against trajectory micro-data. Then, in Sections 2.2 to 2.8, we review relevant studies, separating them into classes based on the proposed taxonomy.

2.1 A taxonomy of attacks

Attack models are typically defined by a precise objective, and by the background knowledge that the adversary can exploit towards attaining such an objective. The taxonomy we adopt builds on three orthogonal dimensions that fully capture these features: *(i)* the objective of the attack, presented in Section 2.1.1; *(ii)* the format of the adversary's knowledge and *(iii)* its origin, which are introduced in Section 2.1.2 and Section 2.1.3, respectively. The classification of the literature that results from the intertwining of these three dimensions is finally presented and discussed in Section 2.1.4.

2.1.1 Attack objective

While the overall aim of an adversary remains that of re-identifying individuals in the trajectory micro-data, or more generally acquiring sensitive information about them from the trajectory micro-data, different approaches can be leveraged to those purposes. Each approach translates into a specialized ***attack objective*** (denoted by **O** in the rest of the paper), which is the first dimension of our taxonomy. The attack objective is in fact a rather standard way to categorize privacy threats on generic micro-data datasets, and we borrow from the classic codification by Fung *et al.* [52] to organize the literature along this dimension.

The vast majority of attacks against trajectory micro-data investigated in the literature belong to the category of ***record linkage attacks*** (**O.1**) as defined by Fung *et al.* [52], which are often simply referred to as *linkage attacks*. As illustrated in the toy example in Figure 3, record linkage attacks aim at mapping records in the *target* trajectory micro-data with *side information* owned by the adversary. The side information must include personal identifiers, without which re-identification is not possible, as well as some (possibly limited) data about the mobility of (a subset of) users in the target database. These can be collected in a variety of ways: examples include directly observing target individuals (*e.g.*, by physically meeting or following them and recording their movements), mining suitable open

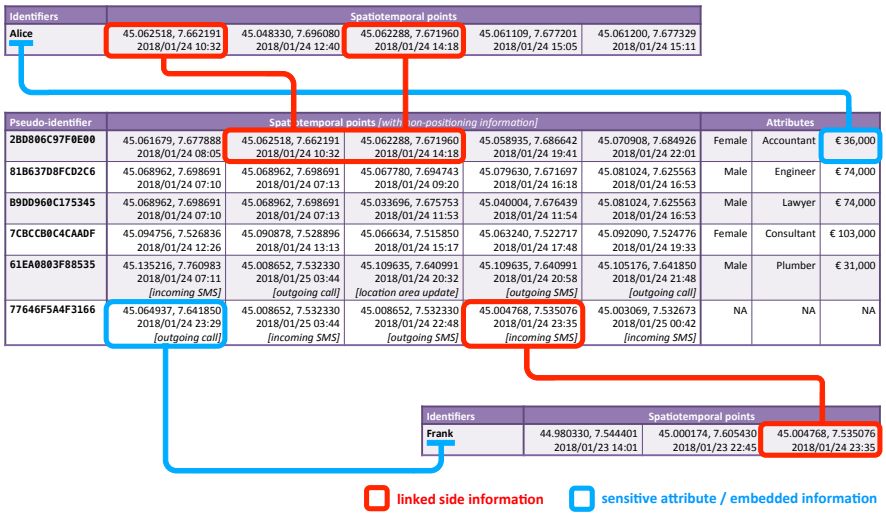

Figure 3: Examples of record linkage attacks against the pseudonymised database in Figure 2. The side information (top and bottom) contains personal identifiers (the identities of *Alice* and *Frank*, respectively) and some spatiotemporal points. Record linkage exploits matches in the trajectory data (*e.g.*, identical GPS locations in the side information and target database, in red) to associate the identifiers to one record. It links the person to sensitive data (in blue), be those attributes (*e.g.*, *Alice*'s salary), or information embedded in the spatiotemporal data itself (*e.g.*, *Frank* visiting a gentlemen club at $45.064937, 7.641850$).

data (*e.g.*, via crawling of geo-referenced social network metadata), or gaining access to samples of the actual trajectory micro-data (*e.g.*, by leveraging a security breach).

A successful attack allows associating an identity to records in the target database. Establishing such a link represents a privacy breach when the records of the target database contain *sensitive attributes*. In the context of trajectory micro-data, there are two different situations where this happens, both exemplified in Figure 3.

- The typical assumption made in most of the works we will review is that the database of trajectory micro-data also includes additional, separated sensitive attributes. For instance, each record could include the spatiotemporal points as well as personal data about the individual such as gender, age, address, employment, or accounting information. In this case, mobility data allows linking personal identities with the non-positioning sensitive data, as in *Alice*'s case in Figure 3.

- A second, subtler perspective is that trajectory micro-data embed information that is potentially sensitive per-se. Gaining access to a large amount of timestamped locations visited by an individual may allow an adversary understand where the individual lives, where she works, and which kind of Points of Interest (PoIs) she visits. The latter can reveal important locations (*e.g.*, home and workplace addresses), commuting patterns (*e.g.*, periodic times of visits to public transportation hubs), religious and political views (*e.g.*, regular visits to places of worship or party meetings), or health conditions (*e.g.*, frequent visits to healthcare structures), just to mention a few relevant examples. Extracting PoIs from trajectory micro-data is in fact a fairly easy task. The vast literature on knowledge discovery from spatiotempral trajectories proposes a variety of techniques to infer home, work, and other relevant locations from

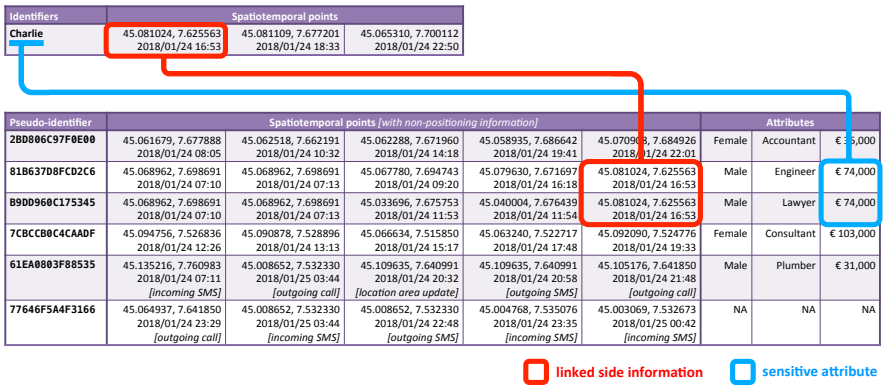

Figure 4: Example of attribute linkage attacks against the pseudonymised database in Figure 2. The side information does not allow linking personal identifiers to one specific record, since two different records contain the single matching spatiotemporal sample. However, the two records yield the same revenue value, hence the attacker can still link the identity of *Charlie* to the sensitive information of a $74,000$-Euro yearly income.

this kind of data [92]. This would be the case of the record linkage against *Frank*'s trajectory micro-data in Figure 3.

Classes of attack beyond record linkage have been rarely investigated in works related to trajectory micro-data. Notable exceptions have explored two additional attack categories.

The first is that of ***attribute linkage attacks*** (**O.2**) in the terminology by Fung *et al.* [52], which are also known as *homogeneity attacks*. In this case, the objective of the adversary is to link its side information with sensitive attributes (rather than specific records) in the target trajectory micro-data. The typical scenario envisioned for attribute linkage is one were the attacker's side information maps to multiple records (hence preventing record linkage), but all such records share the same sensitive attributes (which are thus re-indentified by attribute linkage). Figure 4 depicts a toy example of database configuration that is prone to a homogeneity attack on trajectory micro-data. The privacy risks associated with a successful attribute linkage attack are the same as for record linkage: indeed, record linkage yields a privacy breach because the adversary can infer the sensitive information within a record, and not the specific record itself. Yet, homogeneity attacks can be successful in cases where record attacks are not possible, and thus pose a greater risk to privacy.

The other category of threats considered in the literature is that of ***probabilistic attacks*** (**O.3**) in the categorization by Fung *et al.* [52], which are also referred to as *inference attacks* in the literature. The goal of a probabilistic attack is increasing the adversary's knowledge by accessing the target database. This is equivalent to generalizing the notion of sensitive attribute to any information contained in a record: in the context of trajectory micro-data, learning any additional, non-negligible portion of the mobility of a user beyond the original side information already makes the attack successful. Figure 5 shows an example of probabilistic attack on trajectory micro-data. Preventing probabilistic attacks is more challenging than countering record or attribute linkage, since the adversary's goal is much broader.

2.1.2 Format of the side information

To achieve any of the attack objectives above, an adversary must leverage some background knowledge. This side information can have different formats, which drive the implementa-

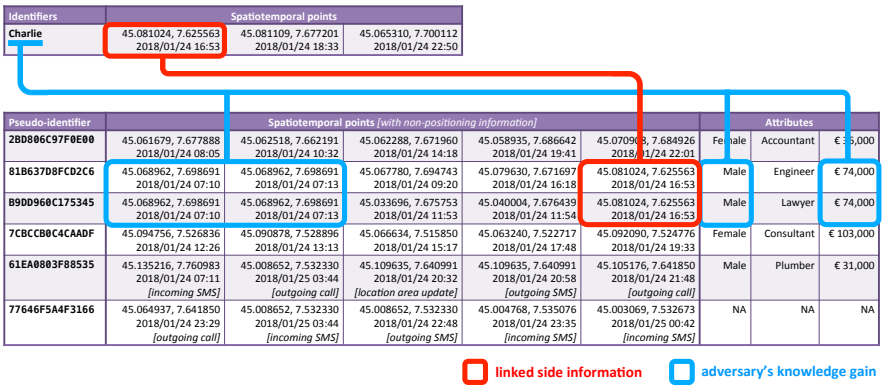

Figure 5: Example of probabilistic attacks against the pseudonymised database in Figure 2. The side information allows the adversary to increase their knowledge, by discovering two new spatiotemporal samples, as well as the gender and revenue of the target individual, *Charlie*. All such additional information is considered sensitive in probabilistic attacks.

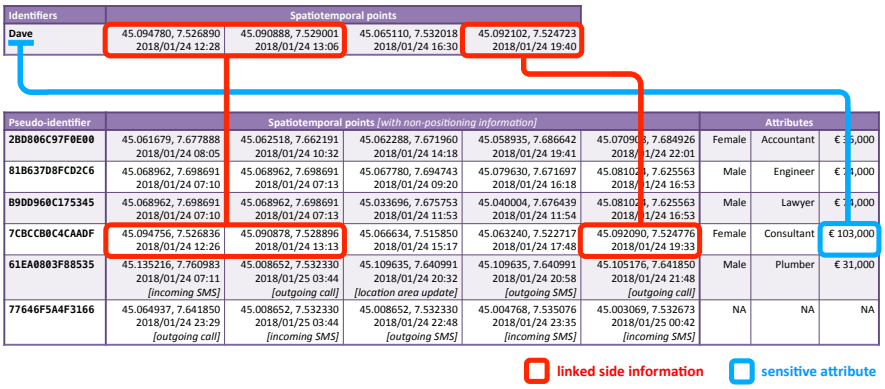

Figure 6: Example of record linkage attack against the pseudonymised database in Figure 2. The side information includes a set of spatiotemporal points that does not perfectly match any of those in the database. The adversary can still map his knowledge to the most similar trajectory: multiple points in his possession are very close to those in one specific record, which allows linking *Dave*'s identity to the sensitive revenue attribute.

tion of the attack. Thus, the second and third dimensions considered in our taxonomy relate to the nature of the background data of the adversary. Specifically, we tell apart the two distinguishing characteristics of the side information: *(i)* its format, *i.e.*, the actual content of the side information, discussed next; and, *(ii)* its source, *i.e.*, how the side information is gathered, which is instead presented in Section 2.1.3.

Concerning the ***format of side information*** (**F**), we identify three classes from studies on trajectory micro-data privacy. In the baseline case, the side information has the exact same format of the mobility data contained in the target database, *i.e.*, a sequence of ***spatiotemporal points*** (**F.1**). Two situations can occur under this format, as follows.

- In a simpler case, the spatiotemporal points in the side information are a ***subset*** (**F.1a**) of those contained in records of the target trajectory micro-data. This is the situation portrayed in the previous examples in Figures 3 through 5. It assumes that the adversary gathers side information using the exact same tracking technology employed to

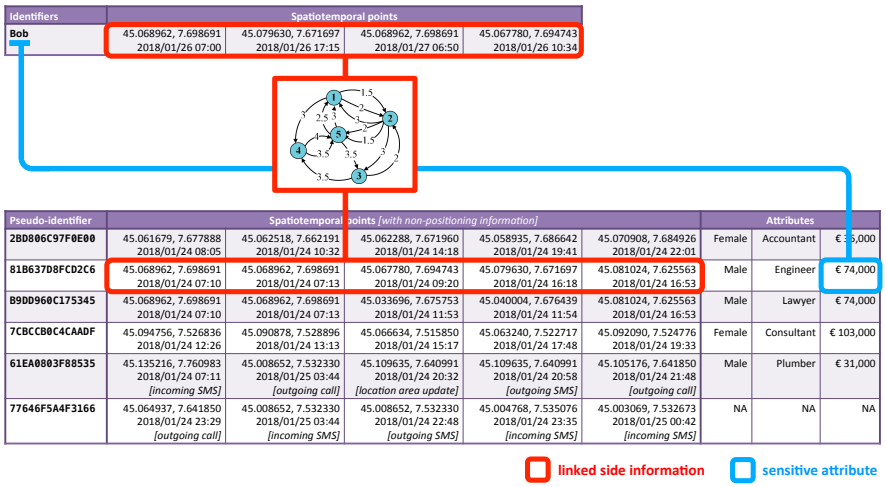

Figure 7: Example of record linkage attack against the pseudonymised database in Figure 2. The side information includes a set of spatiotemporal points that does not perfectly match any of those in the database. The adversary can build a profile of the victim (*e.g.*, a probabilistic mobility model of transitions among locations) from their knowledge, and match this to similar profiles of all records in the database. A very similar profile is identified in the second record, allowing linking *Bob*'s identity to the sensitive revenue attribute.

build the target database. This can be regarded as a best-case scenario for the attacker, which is however unlikely to occur in practical settings.

- In many practical scenarios, the adversary does not possess a perfect subset of the trajectory micro-data. Instead, the spatiotemporal points in the side information are obtained from a different source than that generating the trajectory micro-data, hence they represent a ***diverse sampling*** (**F.1b**) of the underlying spatiotemporal trajectories of the users. Figure 6 exemplifies this situation.

A second type of side information format is represented by indirect knowledge inferred from trajectory micro-data. The adversary does not have access to precise spatiotemporal points of his victim's trajectory micro-data, but knows instead some high-level ***profiles*** (**F.2**) that characterize the movements of the target individuals. The notion of profile is general and can accommodate a wide range of scenarios, among which we identify the following four prominent situations:

- ***mobility models*** (**F.2a**) are mathematical representations that summarize the complete movement behavior of the target individuals, as illustrated by Figure 7;

- ***important locations*** (**F.2b**) are places frequently visited by or especially significant for the target individuals;

- ***mobility features*** (**F.2c**) are specific properties found in the movement patterns of the target individuals;

- ***social graphs*** (**F.2d**) are structures that describe the social relationships of the target individuals with other users whose trajectories are also present in the database.

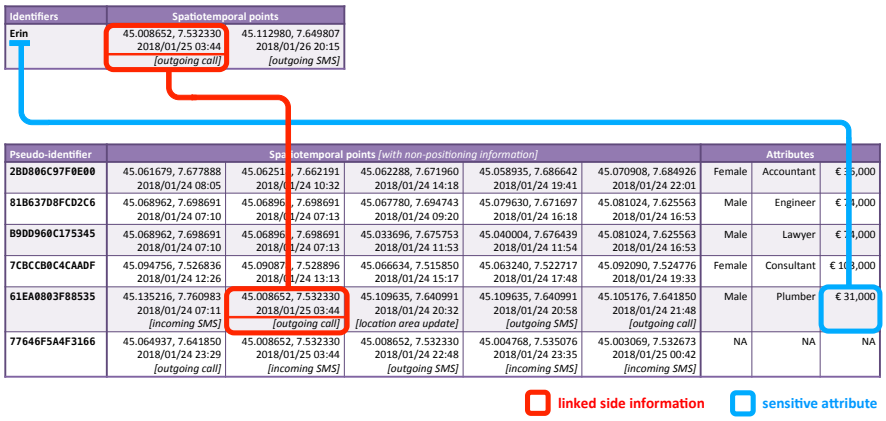

Figure 8: Example of record linkage attack against the pseudonymised database in Figure 2. The side information includes a set of spatiotemporal points, each associated with auxiliary data, within brackets, on mobile communication activities. While the positioning data alone matches two records with different sensitive attributes, the additional knowledge provided by the auxiliary data allows the adversary to tell apart the records, and link *Erin*'s identity to the sensitive revenue attribute.

The third format of side information is a combination of trajectory micro-data and some ***auxiliary data*** (**F.3**) that is not related to the mobility of the target individuals. We can see this format as an augmented version of the two above, thanks to the addition of the auxiliary data. Under the assumption that records on the target database also contain fields related to the auxiliary data, such reinforced side information grants an additional degree of freedom to carry out attacks. An intuitive example is provided in Figure 8.

2.1.3 Origin of the side information

The third dimension we consider in our taxonomy of attacks on trajectory micro-data is that of the ***source of side information*** (**S**) owned by the adversary. Side information can be retrieved from sources belonging to one of two categories, as follows.

- The vast majority of works in the literature directly extract the side information from the target trajectory micro-data. We refer to this approach as ***intra-record*** (**S.1**), since the source of the side information is the data contained in the database records themselves. There exist two subcategories of intra-record sources. So-called ***intra-record subsampling*** (**S.1a**) leaves the original trajectory micro-data unmodified once the side information is extracted: therefore, the side information is necessarily present in the target database. Instead, ***intra-record training*** (**S.1b**) removes from the target trajectory micro-data the side information, separating the original database into training (used as the side information) and test (regarded as the target trajectory micro-data).

- A more realistic approach, which we name ***cross-database*** (**S.2**), consists in considering a side information source that is entirely different from the target trajectory micro-data: indeed, in practical cases, an adversary would derive his background knowledge from direct observations of his victims' movements, or from external datasets that are fully disjoint from the target one. However, a cross-database approach requires suitable side-information databases that contain mobility data for (a subset

of) the users in the target trajectory micro-data, and that are collected via a different technology. Acquiring such databases can be complicated, which is why only a small number of works in the literature adopted this strategy to date.

2.1.4 Literature classification

We can now classify the existing works in the literature based on the three-dimensional taxonomy proposed in the Sections above. Table 1 summarizes how attacks against trajectory micro-data proposed in the literature are positioned according to our taxonomy. We note that record linkage attacks (**O.1**) have been thoroughly investigated, whereas very little attention has been paid to other types of attacks (**O.2**, **O.3**). This also results in that a variety of side information formats have been considered for record linkage; instead, attribute linkage and probabilistic attacks have been only evaluated with baseline format (**F.1a**).

The intersections of attacker objective and side information format (on rows) with the side information source (on columns) also deserve attention. An important remark is that a single type of source can generate multiple formats of side information. For instance, let us look at the case of intra-record subsampling (**S.1a**): it results in a spatiotemporal subset format (**F.1a**), if the points extracted from the target trajectory micro-data are used as they are by the adversary; it can be cast to a spatiotemporal diverse sampling format (**F.1b**), if the extracted points are perturbed in time and space; or, it can lead to any profile formats (**F.2**), if the extracted points are post-processed to infer, *e.g.*, important locations or specific mobility features. Conversely, intra-record training (**S.1a**) and cross-database (**S.2**) sources feature inherently diverse samplings with respect to the target data, hence cannot generate a spatiotemporal subset format (**F.1a**), and the corresponding table cells are grayed out. However, these sources can still produce all other side information formats.

Interestingly, our taxonomy highlights how some patterns are more frequent than others. As an example, side information in the format of a mobility model (**F.2a**) typically requires that the model is trained and tested on different datasets, making an intra-record training (**S.1b**) the most appropriate type of source. Or, important locations (**F.2b**) are by far the most popular type of profile considered in attacks against trajectory micro-data, and have been tested with all kinds of sources of side information. Also, there exist substantial gaps in the literature when it comes to practical attacks that leverage a cross-database source (**S.2**) and exploit profiles in the form of either mobility models (**F.2a**) or features (**F.2c**).

Overall, Table 1 offers an outlook on well-explored as well as less investigated attack surfaces against trajectory micro-data. It also motivates us to opt for a structure of the next Sections that follows the rows of the table: indeed, it is along the objective dimension that we identify most of the diversity among the reviewed studies; moreover, the format of the background knowledge is what really guides the design of the attack strategy, whereas the side information source is easily shaped into different formats, and does not allow for a rigorous classification. Therefore, in the remainder of this Section we first present works that investigate record linkage attacks based on different side information formats, and then discuss studies on attribute linkage attacks and probabilistic attacks.

2.2 Record linkage via subset of trajectory micro-data

Record linkage attacks are straightforward when the side information database stores a subset of the same spatiotemporal points present in the trajectory micro-data, as in class **O.1**/**F.1a** in Table 1. As shown in Figure 3, it is sufficient that the adversary looks for records in the trajectory micro-data that include the points he owns.

Table 1: Classification of the literature of attacks on trajectory micro-data based on our proposed taxonomy. The three dimensions of the taxonomy are highlighted in bold. The leftmost column separates rows according to different classes of attacker objective (**O**) as per Section 2.1.1. The subsequent two columns refine rows according to different categories and subcategories of side information format (**F**) as per Section 2.1.2. The last three columns distinguish types and subtypes of side information source (**S**) as per Section 2.1.3. Grayed-out cells denote unfeasible combinations of side information format and source.

				Side information source (S)		
				Intra-record **S.1**		Cross-database **S.2**
		Side information format (F)		Subsampling **S.1a**	Training **S.1b**	
Attacker objective (O)	Record linkage **O.1**	Points **F.1**	Subset **F.1a**	Bettini *et al.* [14] De Montjoye *et al.* [38] Rossi *et al.* [102] Sapiezynski *et al.* [105] De Montjoye *et al.* [39]		
			Diverse sampling **F.1b**	Ma *et al.* [82] Rossi *et al.* [102]	Rossi and Musolesi [101]	Tockar [117] Cecaj *et al.* [25, 26] Kondor *et al.* [74] Riederer *et al.* [100] Wang *et al.* [128]
		Profile **F.2**	Mobility model **F.2a**		De Mulder *et al.* [40] Shokri *et al.* [109] Gambs *et al.* [53] Murakami *et al.* [90]	
			Important locations **F.2b**	Freudiger *et al.* [51] Zang and Bolot [134]	Unnikrishnan and Naini [123] Naini *et al.* [93] Rossi and Musolesi [101]	Krumm [76] Goga [58]
			Mobility features **F.2c**	Rossi *et al.* [102] Zan *et al.* [133]		
			Social graph **F.2d**			Srivatsa and Hicks [112] Ji *et al.* [69, 70]
		Auxiliary **F.3**		Zang and Bolot [134] De Montjoye *et al.* [39]		Goga [58]
	Attribute linkage **O.2**	Points **F.1**	Subset **F.1a**	Sui *et al.* [113]		
	Probabilistic **O.3**	Points **F.1**	Subset **F.1a**	Gramaglia *et al.* [62]		

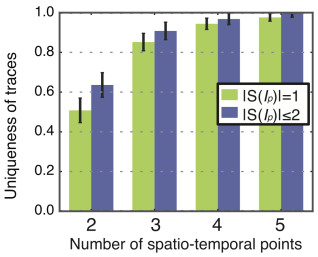

Figure 9: The unicity of trajectories with respect to the number of random spatiotemporal points known to the adversary, denoted by I_p. The green bars represent the fraction of unique trajectories, while the blue bars show the fraction of cases where the random points identify one or two trajectories. Knowledge of four random points pinpoints 95% of the trajectories in a 1.5-million-record database. Reproduced from De Montjoye *et al.* [38].

The key question here is: how much such side information is required to perform a successful linkage? The answer roots in the concept of *unicity*, which is a measure of the diversity that characterizes the movement patterns of different individuals. The higher the unicity of trajectory micro-data, the higher the actual privacy risk connected with them: if the monitored users have very diverse spatiotemporal trajectories, an adversary will likely find a single trajectory micro-data record matching the side information points he owns.

In a pioneering work, Bettini *et al.* [14] first hint at unicity in trajectory micro-data by introducing the notion of a location-based quasi-identifier (LBQID). The formal definition[1] of LBQID is *"a spatio-temporal pattern specified by a sequence of spatio-temporal constraints each one defining an area and a time span, and by a recurrence formula"*. In other words, a LBQID is a sequence of spatiotemporal areas (which could specialize to points at high granularity) that define a mobility pattern, and some associated frequency of occurrence (which could specialize to just one occurrence). The intuition of Bettini *et al.* [14] is that it is possible to define LBQIDs that require a minimum amount of knowledge about the spatiotemporal trajectory of a user, and yet allow uniquely pinpointing his trajectory micro-data among those of a large population. In this sense, an LBQID becomes a spatiotemporal pattern that is unique to one individual in a database of trajectory micro-data.

The demonstration of such a conjecture is provided by the seminal work of De Montjoye *et al.* [38], who show that minimal LBQIDs, *i.e.*, very little positioning side information, is sufficient to carry out successful linkage attacks against trajectory micro-data. Specifically, the authors prove that knowledge of a few random points in the trajectory micro-data of a user allow pinning him down almost certainly, even within a very large population[2]. For instance, an adversary having observed the whereabouts of a target individual at two random moments (whose corresponding spatiotemporal points are present in the target database) during a whole year has a 50% probability of recognizing his target in a dataset of millions; the percentage grows to 95% if as little as four random points are known to the attacker. Figure 9 shows the exact dynamics of unicity versus the number of haphazard spatiotemporal points in the side information, in the considered scenario.

The works above analyse mobile phone trajectories, but unicity is a general characteristic of trajectory micro-data, no matter their original source. This is confirmed by subsequent works that test the effectiveness of linkage attacks against trajectory micro-data collected

[1] The definition is inspired by the literature on privacy in traditional relational databases. There, each record is associated with some data values, and a set of quasi-identifiers: for instance, in a renowned study Sweeney [114] analyses a database of all US citizens, where their individual healthcare records (*i.e.*, the actual data, including admit/discharge dates, diagnoses, cures, and charges) are stored along with three quasi-identifiers (*i.e.*, date of birth, gender, and ZIP code). The author finds that 87% of individuals have a unique combination of the three quasi-identifiers above, which clearly raises questions on the privacy of the dataset. In databases of trajectory micro-data, there is no clear distinction between quasi-identifiers and the actual data, as any portion of the spatiotemporal trajectory itself can become a quasi-identifier, depending on the positioning side information possessed by an adversary. The formal definition of LBQID captures this condition.

[2] The dataset is composed of CDRs of 1.5 million users, collected by a network operator during 15 months.

in a variety of ways. Rossi *et al.* [102] investigate unicity in GPS traces[3], and show that the high spatial accuracy (in the order of meters) of this kind of trajectory micro-data exacerbates the phenomenon. Indeed, 100% of the users in the datasets are pinpointed with just two random spatiotemporal points. Sapiezynski *et al.* [105] employ Wi-Fi trajectory micro-data[4] and demonstrate that knowing as little as 0.1% of the APs seen by a user (*i.e.*, around 20 APs on average) allows tracking down a typical user during 90% of her movements. De Montjoye *et al.* [39] confirm that unicity persists also in the case of trajectory micro-data from digital payments[5]. In this context, 90% of users are unique by assuming knowledge of four random points in their trajectory micro-data. Interestingly, the authors also show that unicity varies with gender (women being 1.2 times more prone to unicity than men) and income level (high-income users being 1.7 times more prone to unicity than low-income ones): the authors speculate that the reason is the location entropy, *i.e.*, the variety of different shops visited, which is higher in women and high-income individuals.

2.3 Record linkage via diverse sampling of trajectory micro-data

As mentioned in Section 2.1, in practical cases the background knowledge of the adversary is hardly a subset of the target trajectory micro-data. In a more credible scenario, the attacker collects side information from a source that has the same spatiotemporal granularity of the target database, but a different sampling of the actual trajectory of the target individuals. This maps to class **O.1**/**F.1b** in Table 1.

As illustrated in Figure 6, record linkage attacks must reconciliate spatiotemporal trajectories that are not one a subset of the other, which clearly makes the problem more challenging. A first set of strategies is proposed in a seminal work by Ma *et al.* [82], who devise four different estimators to measure the similarity of mobility traces in trajectory micro-data and diversely sampled side information[6]. These are *(i)* a Maximum Likelihood Estimator based on Euclidean distances, *(ii)* a Minimum Square Approach computing the negative sum of the squared absolute value of the trajectory difference, *(iii)* a Basic Approach that assumes Gaussian-noisy side information and defines the identity between two spatiotemporal points based on the standard deviation of such noise, and *(iv)* a Weighted Exponential Approach that generalizes the previous technique to any noise distribution. The methods achieve a linkage success rate of 50%, when the adversary owns as little as 10 observations of the original trajectory. The result then confirms the high privacy risks associated to trajectory micro-data, even when the side information does not precisely match the spatiotemporal points in the target database.

A similar study is carried out by Rossi and Musolesi [101] in the case of location-based social networks (LBSNs), where localised and timestamped check-ins of individual users result in trajectory micro-data[7]. The authors show that a classifier based on a modified

[3]The study leverages GPS trajectory micro-data from CabSpotting [97] CenceMe [91] and GeoLife [137]. The three datsets cover 536, 20 and 182 users, respectively, for several weeks and with diverse sampling frequencies. A subset of the original data, of variable size, is used as side information; the complement is used as target data.

[4]The data consists of Wi-Fi scans of 63 users, obtained by storing the list of APs that respond to probe messages broadcasted by the users' mobile devices. As AP locations are easily obtained from services such as Google Geolocation API, timestamped Wi-Fi scans are a form of client-centric trajectory micro-data. The authors also use GPS traces of the same users as ground truth.

[5]The work used 3 months of credit card transactions of 1.1 million users in 10,000 geo-referenced shops.

[6]Three pseudonymised GPS traces are considered in the study: 536 cabs in San Francisco from CabSpotting [97], 2,348 buses in Shanghai and 4,438 cabs in Shanghai. The side information is generated by adding noise to randomly sampled portions of the trajectory micro-data.

[7]The data is crawled from three different LBSNs, *i.e.*, Brightkite and Gowalla [33], and Foursquare [54], and

Hausdorff distance between trajectories allows linking between 30% and 60% of the users in the trajectory micro-data with 10 LBSN check-ins. Further evaluations of this approach with other datasets[3] presented by Rossi *et al.* [102] show success probabilities over 90%.

While the previous studies approximate diversely sampled side information by perturbing or splitting the trajectory micro-data, further experiments consider the more realistic case where trajectory micro-data and side information come from two sources that are actually different. That boundary is first crossed by Tockar [117], who carries out a linkage attack against trajectory micro-data of cabs in New York City, US[8]. Tockar [117] proves that the unicity of taxi trips makes them easily linked to other databases that contain information about taxi rides of specific individuals. To this end, he gathers an ad-hoc database by browsing gossip blogs and collecting where and when celebrities used yellow cabs in the NYC area in 2013. Linkage of spatiotemporal points allows the author to re-identify VIP passengers in the trajectory micro-data, and hence the origin or destination of their trips, as well as the associated tip amount. The latter are treated as the sensitive information, unveiling frequently visited locations of celebrities, as well as their (lack of) generosity in tipping. Although these pieces of information are deemed "relatively benign" by the author, the trial represents a first clear example of actual privacy breach through a linkage attack on trajectory micro-data.

Cecaj *et al.* [25, 26] employ a real-world dataset of trajectory micro-data, and use geotagged social network metadata as side information[9]. By applying a simple statistical learning approach based on matching and mismatching of spatiotemporal points in the trajectories, they can link tens of social network usernames to specific trajectory micro-data records. The result is in fact merely probabilistic, as it is not based on actual ground truth (*i.e.*, the identity – as social network username – of users in the trajectory micro-data): instead, a maximum a-posteriori estimation is used to compute the match probability. A similar evaluation approach is considered by Kondor *et al.* [74], who investigate the *matchability* of large-scale datasets of trajectory micro-data[10]. To this end, they define space and time thresholds tailored towards the characteristics of urban movements, and identify matching points (which are within the aforementioned thresholds) and alibis (which are within the threshold in time, but not in space) in trajectory pairs across the two datasets. Then, each trajectory in one dataset is linked to that with the highest number of matching points and no alibis on the other dataset. According to the results, the authors expect a successful match for around 8% of users in one-week datasets, and for about 33% in one-month datasets. Such percentages grow to 15% and 60% respectively, when focusing on very active, regular users only.

A more elaborate approach and dependable evaluation are proposed by Riederer *et al.* [100], who design a dedicated algorithm for linking trajectory micro-data and social network metadata. The algorithm starts by computing a score for each pair of users across the two

covers users in three US cities, *i.e.*, San Francisco, New York and Los Angeles, for several months. The location information is approximated by the check-in venue in the first two datasets, while it corresponds to the actual user position in the third one. In each LBSN, a subset of the data is used as side information, and the rest as the target trajectory micro-data.

[8] The data comprises pickup and drop-off times and locations, fare and tip amounts of every yellow cab ride in New York City, US, in 2013. The dataset, released by the New York City Taxi and Limousine Commission under Freedom of Information Law (FOIL), is pseudonymised.

[9] The trajectory micro-data consist of pseudonymised CDRs of 2 million mobile subscribers, while the side information are timestamped and geo-referenced posts of 700 usernames crawled from Flickr and Twitter.

[10] They use CDRs of 2.8 million mobile subscribers in Singapore, and smart card bus/train transportation data of 3.3 million users collected by the Singapore Land Transportation Authority (LTA). The two datasets are collected in the same time period.

databases, representing the likelihood of the user pair being actually the same person; it then maximizes the overall score via a bipartite matching. The algorithm is proven to be theoretically correct under the assumption that visits to a specific location during a certain period follow a Poisson distribution and are independent of other visits. Tests with real-world datasets[11] show that the algorithm outperforms approaches based on sparsity, frequency of visit and density, reaching up to 0.95 precision and 0.7 recall in the best case.

The first test at scale is that recently performed by Wang *et al.* [128]. They leverage an impressive collection of large-scale real-world datasets[12] to carry out a comparative analysis of record linkage attacks proposed in the literature, including those by Ma *et al.* [82], Rossi and Musolesi [101], Cecaj *et al.* [25, 26], and Riederer *et al.* [100]. All these strategies achieve hit precision sensibly higher than zero, hence they can successfully link users across databases. However, and quite interestingly, the results show that the performance of these strategies in presence of large-scale real-world datasets are reduced with respect to those reported in the original papers, with hit precision well below 20% even when the side information comprises tens of spatiotemporal points. The authors' explanation is that each attack only addresses a subset of the issues emerging in practical settings, which mostly stem from spatiotemporal mismatches between target and side information data, and from database sparsity. They then propose an attack technique that leverages a probabilistic representation of the spatiotemporal mismatch and uses a simple Markov model to estimate missing spatiotemporal points. The approach achieves a substantial gain in linking records, with a maximum hit ratio of 40%.

2.4 Record linkage via mobility models

A different attack surface for linkage is represented by indirect knowledge inferred from trajectory micro-data. Let us imagine an adversary who does not have access to precise spatiotemporal points of his target's trajectory micro-data, but knows instead some profiles that characterize the movements of his target. Such an attacker could then extract the same profiles for all records in the trajectory micro-data, and try to link a specific record to his side information. Figure 7 illustrates this concept.

Mobility models are the first type of profile that has been considered in the literature, as per class **O.1**/**F.2a** in Table 1. The strong *regularity* that is known to characterize human movements [60] allows constructing simple models of individual movements that approximate well the actual mobility patterns, and that thus represent a valuable side information. The early work by De Mulder *et al.* [40] is especially influential in this sense. The authors assume that an adversary owns a Markovian model of his target's mobility; the model describes memoryless transition probabilities among visited locations. The authors then propose two matching strategies: the first builds a Markovian model from each record in the target trajectory micro-data, and then compares the Markovian models directly; the second calculates the probability that the specific sequence of locations in each record of the trajectory micro-data is generated by the side information Markovian model. Evaluations

[11] Three different pairs of trajectory micro-data databases are employed, *i.e.*, 862 users in two databases crawled from Foursquare and Twitter, 1717 users in two databases crawled from Instagram and Twitter, and 452 users in CDR and credit card record databases. In each scenario, all databases are pseudonymised, but the ground-truth mapping of users between the two databases is known.

[12] Three types of datasets are considered in the study: mobile network CDRs of more than 2 million subscribers, which is the target trajectory micro-data; GPS data of 56,000 Weibo social network users; check-in locations of 10,000 Weibo and 45,000 Dianping application users. All datasets are collected during the same week, and the side information datasets only contain users who are also present in the target CDR database. User pseudonyms are consistent across the datasets, which allows validate the results of record linkage attacks on ground truth.

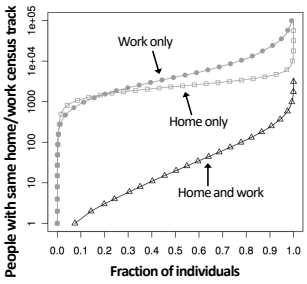

Figure 10: Fraction of population (x axis) whose home (squares), work (circles) or both home and work (triangles) locations (recorded at the census tract level) are shared with at most y other individuals. While the set of people sharing the same home or work places are typically large (in the order of several hundreds to tens of thousands), more than 5% of workers do not share their combination of home and work locations with any other individual in the US. Adapted from Golle and Partridge [59].

with real-world data[13] show that the linkage is successful in 80% of cases. Although the scale of the experiment is small, the result still demonstrates that a risk exists.

Markovian representations of mobility have been then considered as the side information by several follow-up studies. Shokri *et al.* [109] assume an adversary who aims at linking transition probability matrices to whole trajectories in a dataset of trajectory micro-data[14] via a maximum likelihood approach. Gambs *et al.* [53] consider a wide range of techniques to pair Markovian models with trajectories in the target trajectory micro-data. Tests with heterogeneous datasets[15] result in a success ratio of linkage attacks between 10% and 50%. Interestingly, the authors show that the percentages grow proportionally with the sampling rate of the trajectory micro-data. Murakami *et al.* [90] investigate the case where the side information possessed by the attacker is limited and possibly collected from a non-identical set of users. These conditions result in transition probability matrices of the Markovian models that are sparse and erroneous. In this scenario, previous techniques for linkage, such as the maximum likelihood used by Shokri *et al.* [109], do not perform well. The authors then propose an attack that *(i)* reduces the problem dimension by means of group sparsity regularization (*i.e.*, clustering) of the locations visited by users, and *(ii)* estimates the complete transition probability matrices via a dedicated factorization of the tensor composed by all side-information (sparse) transition matrices. Tests with real-world datasets[16] prove the effectiveness of the solution, which attains around 70% AUC (Area Under the Curve) of the ROC (Receiver Operating Characteristic) of the true positive rate against the false positive rate.

2.5 Record linkage via important locations

Another class of trajectory micro-data profile used as side information are locations that are frequently visited by the users, which maps to class **O.1**/**F.2b** in Table 1. The most intuitive example is that of home and work locations. Such an apparently basic knowledge poses in

[13]The study employs a processed version of the Reality Mining dataset [47]. Individual spatiotemporal trajectories of 100 volunteers are recovered from the entry and exit time of each user at GSM cells. Side information is constructed from two months of mobility, whereas another two months are used as target trajectory micro-data.

[14]The trajectory micro-data comprises of 20 mobility traces from the CabSpotting dataset [97], where the sampling frequency is fixed at 5 minutes over 8 hours. Users move within the San Francisco bay area, which is divided into 40 regions forming a 5×8 grid. Side-information Markovian models are directly extracted from noisy subsets of the trajectory micro-data.

[15]The study employs GPS logs from five different scenarios, *i.e.*, Arum [72], GeoLife [137], Nokia [73], San Francisco cabs [97], and Borlange [51]. All datasets describe individual spatiotemporal trajectories, covering from a few users to almost two hundred individuals. Side information Markov models are extracted from a subset of each dataset, while the rest is considered as the trajectory micro-data for linkage.

[16]The authors employ data from 80 GeoLife users [137], 250 Gowalla users [33], and 400 Foursquare users [131]. All datasets cover tens of months of trajectory micro-data. The side information is computed from 1.5% to 9% of the individual data, and the remaining data is used as the target trajectory micro-data.

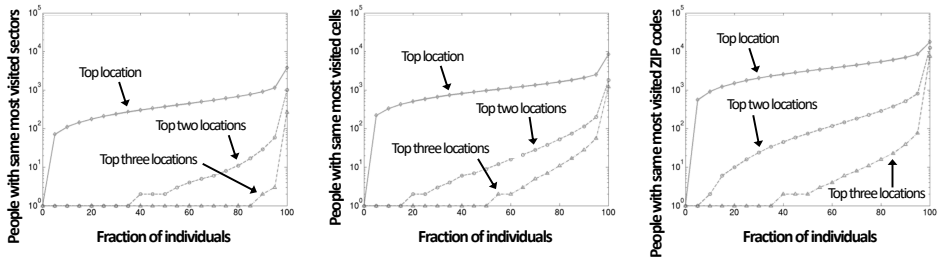

Figure 11: Fraction of trajectories (x-axis) that share the top one, two or three most frequently visited locations with at most y other trajectories in the target database. Plots refer to different spatial granularity levels with decreasing resolution: US census sector (left), network cell (middle), and US ZIP code (right). While no trajectory is unique when considering that only the top visited location is known to the attacker, all other situations allow the adversary to perform successful record linkage on 5% (top two locations known at a ZIP code level) to 85% (top three locations known at a sector level) of the trajectories of 20 million users. Adapted from Zang and Bolot [134].

fact a very severe risk to privacy: Golle and Partridge [59] show that the home and work locations of over 100 million individuals in the US, collected by the Longitudinal Employer-Household Dynamics (LEHD) program, suffer from unicity. Namely, more than 5% of the population shows unique home-work location pairs at tract granularity. Figure 10 provides a complete view of the result.

Based on this observation, in his seminal work, Krumm [76] leverages home location side information to run linkage attacks on trajectory micro-data[17]. The author proposes four simple heuristic algorithms to infer the latitude and longitude of home locations of each pseudonymised user whose trajectories are stored in a given trajectory micro-data record. He then performs a reverse lookup for the home location in the side information: if a unique matching entry is present, the attack allows linking the trajectory micro-data record to an identity. In the considered scenario, the chance of success for an adversary is at least 5%.

The study is extended to the case where the attacker has side information about both home and work locations by Freudiger *et al.* [51]. Their linkage attack infers home and work locations in trajectory micro-data by: *(i)* clustering spatial points that map to frequently visited locations, via a variant of k-means; *(ii)* tagging the most popular location overnight as home, and that during working hours as work. The unicity of home and work locations allows pinpointing users in real-world datasets[18] with probabilities from 10% to 90%.

Zang and Bolot [134] consider a more generic notion of important locations, as the top n locations that are most frequently visited by a user, where n is a small number. They show that such side information is sufficient to distinguish a large fraction of users among millions others[19]. In their tests, 50% of the individuals can be singled out by considering

[17]The analysis is based on two-week (or longer) GPS trajectory micro-data of 172 volunteers. The side information is easily obtained from an online "white pages" service, which provides an association between the name and home address of individuals. The ground truth are the actual home and work locations of the volunteers, who communicated them as part of the experiment.

[18]The study uses data from two-year GPS mobility of 24 cars in Borlange, Sweden and one-year GPS mobility of 143 users in Lausanne, Switzerland [73]. Home and work side information is artificially obtained by selecting a limited subset of points from each user trace, via different subsampling strategies, and then applying the same heuristic described above.

[19]The study is performed on Call Detail Records (CDRs) of a nationwide US mobile network operator, collected over a month and describing the spatiotemporal trajectories of approximately 20 million subscribers. Side-information on important locations is directly extracted from the same data used as target trajectory micro-data.

the top 3 mobile network cells they are observed at, and linkage is shown to be successful also at different spatial granularity levels. Figure 11 portrays complete results for $n \in [1, 3]$.

Alternative versions of the trajectory micro-data profile of Zang and Bolot [134] are proposed in subsequent works by Unnikrishnan and Naini [123], Naini *et al.* [93], Goga *et al.* [58], and Rossi and Musolesi [101]. In the first study, Unnikrishnan and Naini [123] investigate the case where the adversary's side information is in the form of histograms of the time spent by users at different locations. Once comparable histograms are derived from the target trajectory micro-data, linkage is formulated as a matching problem on a bipartite graph where vertices represent records in the two datasets, and a maximum likelihood technique is used to solve it. Their evaluation[20] results in a success probability of more than 50% of the subject when using one day per week, and of 70% when considering two days. A generalization of the study is proposed by Naini *et al.* [93], using non-indentical sets of individuals in the side information and target trajectory micro-data.

The approach is similar in the work by Goga *et al.* [58], where the frequency histograms, called location profiles, are computed at ZIP-code geographic granularity, and weighted so that locations that are less common across all profiles but more representative of specific profiles are valued more. The similarity between histograms is computed using a Cosine distance, as other similarity metrics are shown to yield little difference in results. Interestingly, the authors employ target and side information trajectory micro-data from actual different databases[21], and show that their attack strategy is highly effective, linking records with 40% to 80% of true positives and 1% of false positives. Finally, Rossi and Musolesi [101] assume a slightly different form of side information, *i.e.*, (time-dependent) distributions of visit frequencies at different locations. They leverage a (time-dependent) multinomial naïve Bayes model to match the adversary knowledge to equivalent distributions extracted from trajectory micro-data. Experimental tests[7] show that the attack has an accuracy between 40% and 90%, depending on the dataset, when the adversary knowledge is built from 10 spatiotemporal points from each record in the original trajectory micro-data.

2.6 Record linkage via mobility features

A couple of works examine situations where the side information consists of specific features that characterize the movements of individuals, hence falling in class **O.1**/**F.2c** in Table 1. Rossi *et al.* [102] assume that the adversary possesses knowledge of the travelled distance, average speed, and heading direction collected during a number n of different (and typically small) time window. This is equivalent to replacing spatiotemporal points with distance, speed or heading points in the linkage attack. As all these features are easily inferred from trajectory micro-data records, they can then be mapped to the side information. The authors demonstrate that knowledge of these features poses an equivalent privacy risk to that of regular spatiotemporal points. Tests with real-world trajectory micro-data[3] demonstrate that unicity can reach values up to 95%, although there is significant variability across datasets and feature types. Performance are summarized in Figure 12.

[20] The authors employ trajectory micro-data from Wi-Fi access to geo-referenced APs by over a thousand students in the campus of EPFL, Lausanne, Switzerland. The data covers Mondays and Tuesdays in two weeks. The histogram side information is inferred from days in the first week of the trajectory micro-data, while days in the second week are used in their original format as the target dataset for linkage.

[21] The databases are crawled from Flickr, Twitter, and Yelp. Ground truth on the identity of users appearing across databases is built by looking for accounts in the three platforms that are associated to a same e-mail address, for all addresses mined from a very large e-mail dataset. The final datasets include check-in data of 232,000 Twitter users (used as the target trajectory micro-data), 22,000 Flickr users and 28,000 Yelp users (used as side information).

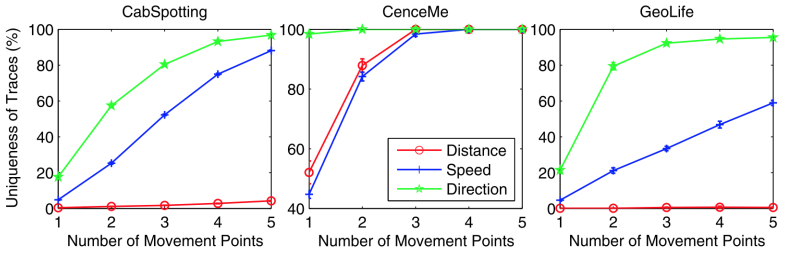

Figure 12: Average percent unicity of trajectories in terms of their mean travelled distance (red circles), speed (blue crosses), or direction of movement (green stars) over time windows of 30 seconds. Unicity is portrayed versus the number of distance, speed or direction feature points in the adversary's side information. The three plots refer to different datasets: CabSpotting [97] (left), CenceMe [91] (middle), and GeoLife [137] (right). Unicity typically rises well above 50% with just a few known feature points, and travel direction appears to be the most distinguishing feature. Reproduced from Rossi *et al.* [102].

Another case where features are retained as side information is the peculiar scenario envisioned by Zan *et al.* [133], who focus on vehicular trajectory micro-data. Their side information comprises features of the driving style of each individual. An attack is proposed that builds on machine learning models to classify vehicles in the trajectory micro-data based on type (car, truck, motorcycle). The performance evaluation shows that this classification is already sufficient to significantly reduce diversity in real-world fine-grained vehicular movement traces[22], hence making the unicity problem at least more severe.

2.7 Record linkage via social graphs

In a distinctive study, Srivatsa and Hicks [112] consider social graphs as a profile side information, *i.e.*, class **O.1**/**F.2d** in Table 1. The assumption of the authors is that physical encounters are more frequent among friends: hence, a contact graph derived from trajectory micro-data is tightly correlated to a social graph such as that obtained by crawling friendships in social networks. Under this condition, it is then possible for an adversary to link names in the social graph with records in the trajectory micro-data, by extracting a contact graph from the latter. Their proposed attack aims at linking nodes in the two graphs: it takes as an input the pseudonymised contact graph, where a small number of nodes, called landmarks, are de-anonymized either by leveraging centrality features or by exploiting leaked information. Starting from landmarks, the method completes the mapping between graphs using distance vectors, spanning tree matching or local subgraph features. The output mapping allows assigning social network identifiers to the spatiotemporal trajectories in trajectory micro-data. Evaluations with real-world datasets[23] show that the attack can achieve a high accuracy above 80%; however, it is important to note that the result holds when considering a relatively small user base (less than 125 individuals), and that the scalability of the attack to much larger datasets is unclear. Improvements to the

[22] Vehicular trajectory micro-data are collected by NGSIM on the US101 highway [2], during 10 minutes. Side information comprises profiles of speed, acceleration, lane changing, and headway distance of each vehicle.

[23] The analysis leverages trajectory micro-data in the form of individual trajectories from geo-referenced associations of mobile devices with Wi-Fi APs in the university campus of St Andrews, UK [15]. The side information is derived from Facebook friendship relationships of the same set of student volunteers in the trajectory micro-data. Ground-truth information, mapping Facebook identities to mobility traces, is also provided by the experiment participants.

approach have been more recently proposed by Ji *et al.* [69, 70], which however do not address the questions on scalability.

2.8 Record linkage via auxiliary side information

Auxiliary data that is unrelated to the movements of the target individual may also be exploited during attacks on trajectory micro-data. The nature of the auxiliary information is potentially very diverse.

For instance, one could consider as auxiliary data the knowledge of whether the target person is actually present in the database under attack; in fact, this piece of information is typically taken for granted in the works reviewed in the previous Sections, hence all previously discussed attacks implicitly leverage auxiliary data in a sense. As another example, in the case where attacks occur against anonymized datasets, knowledge of the privacy-preserving transformation used to protect the trajectory micro-data can also be regarded as auxiliary information; this type of auxiliary data is instead never part of the adversary's side information in the attacks surveyed in the present Section, which are only run on pseudonimyzed datasets.

A couple of works in the literature show how auxiliary side information impacts the unicity of trajectory micro-data in practical case studies, and fall in class **O.1**/**F.3** in Table 1. Zang and Bolot [134] show that unicity in real-world large-scale trajectory micro-data [19] is sensibly increased when the auxiliary data consists of minimal social information about the target user. Namely, they assume that the adversary also knows whether his target individual is especially social or not, *i.e.*, he calls more than 20 unique persons in a month or otherwise. This notion is also included as a flag field in each record of the target trajectory micro-data database. Knowledge of this auxiliary one-bit piece of information permits unicity increase by around 50% on average in the million-strong dataset considered in the study. De Montjoye *et al.* [39] examine instead the case where auxiliary purchase cost data is associated to each spatiotemporal point in trajectory micro-data from credit card transactions[5]. If the adversary happens to be able to associate a purchase cost to each of the target's trajectory point he knows, his chances of success grow by 22% on average.

The auxiliary data leveraged by Goga *et al.* [58] consists instead of language and timing information, which is used to complement the important locations that represent the adversary's baseline side information. Tests with different real-world datasets[21] reveal that such an auxiliary information allows for mild improvements of re-identification rates: in particular, knowing the language of individuals does not help pinpointing users in a more accurate way than just relying on important location profiles.

2.9 Attribute linkage via subset of trajectory micro-data

We now move to a different category of attacks, whose objective is not record linkage but attribute linkage. These map to class **O.2**/**F.1a** in Table 1. In this case, the adversary exploits a weakness of databases that is referred to as *homogeneity*, and is subtler than unicity. Hence, as mentioned in Section 2.1.1, attribute linkage attacks are also called homogeneity attacks in the literature.

In order to clarify the privacy risk associated with homogeneity, let us consider a trajectory micro-data database where unicity is completely absent, and each record contains in fact trajectory micro-data that cannot be told apart from those of a large number of other records. In this case, any side information always matches many records, and a linkage with the correct record has low chances of success. However, recall that the ultimate goal

of an adversary is the inference of the sensitive information within a record, and not of the record itself. And, the fact that the trajectory micro-data in each record are not unique does not necessarily imply the same for the sensitive part of the database. Homogeneity is precisely the lack of diversity in the sensitive information across the set of indistinguishable trajectory micro-data records. In other words, it can be understood as an extension of unicity to the level of combined trajectory data and sensitive attributes, rather than at the trajectory level only.

As depicted in Figure 4, attribute linkage attacks exploit homogeneity. The figure refers to the case where sensitive attributes are specific attribute fields in each record, separated from the spatiotemporal points; yet, attribute linkage can be also cast to the other perspective outlined in Section 2.1.1, where sensitive information is embedded in the spatiotemporal trajectory of the user.

To date, there are no investigations of attribute linkage attacks against trajectory micro-data databases that also contain separate fields with sensitive information. Instead, a recent work by Sui *et al.* [113] explores the case where the sensitive knowledge is embedded in the trajectory micro-data. More precisely, this study assumes that the two most frequently locations visited by an individual are sensitive implicit attributes in trajectory micro-data. Considering that the side information of an adversary consists of three random spatiotemporal points, the authors show that real-world trajectory micro-data[24] is affected by significant homogeneity. Namely, 35% of records in their database are not unique, but 40% of such non-unique records are homogeneous. An interesting corollary observation from the results in this study is that homogeneity scales exponentially with the number of indistinguishable records in the considered scenario: therefore, the problem of homogeneity may thus be inherently mitigated in fairly large sets of records where a large number of records share similar, non-unique trajectory micro-data.

2.10 Probabilistic attacks via subset of trajectory micro-data

Probabilistic attacks are the third type of threat against trajectory micro-data that has been considered in the literature, and correspond to class **O.3**/**F.1a** in Table 1. In this case, the adversary is successful if, upon accessing the database, he increases his knowledge of the target individuals' trajectories by any non-negligible amount. In a sense, probabilistic attacks can be understood as a generalization of attribute linkage: while attribute linkage leverages the homegeneity of specific attributes (*i.e.*, those deemed to be sensitive), a probabilistic attack can take advantage of the homogeneity of any field in the database. Therefore, probabilistic attacks have a significantly broader scope than record or linkage attacks. An illustration of this concept is in Figure 5, where a successful probabilistic attack is performed on a trajectory dataset where unicity and homogeneity are removed.

The concept of probabilistic attacks against trajectory micro-data is evoked in a recent work by Gramaglia *et al.* [62], under the simple assumption of a spatiotemporal subset format obtained from intra-record subsampled data[25]. In their work, the authors only introduce the threat from a conceptual standpoint, and do not run actual attacks on real-world

[24] The evaluation employs association logs of 150,000 users to 2,670 APs of the Wi-Fi network of Tsinghua University, China. The authors assume the side information available to the attacker to be formed by a subset of the target trajectory micro-data.

[25] The study leverages CDR datasets provided by mobile network operators and describing spatiotemporal trajectories of subscribers in the cities of Abidjan (29,191 individuals), Dakar (71,146 individuals) and Shenzhen (50,000 individuals), and in the countries of Ivory Coast (82,728 individuals) and Senegal (286,926 individuals) during a period of two weeks in each scenario.

data. In conclusion, we currently lack comprehensive, data-driven assessments of the effectiveness of probabilistic attacks in the context of trajectory micro-data, which limits our current understanding of the actual risks entailed by this kind of adversarial strategy.

3 Anonymization of trajectory micro-data

Having reviewed the different classes of attacks on trajectory micro-data, in the second part of the survey we turn our attention to the different techniques adopted in order to protect databases of trajectory micro-data from the threaths in Section 2. We first propose a taxonomy of the anonymization solutions in the literature, in Section 3.1; this allows us to structure the subsequent detailed discussion of relevant works, in Sections 3.2 to 3.4.

3.1 A taxonomy of anonymization techniques

Anonymization techniques are primarily characterized by the *privacy principle* they seek to implement in the data. A privacy principle expresses some conditions on the maximum knowledge that can be gained by the attacker upon accessing the target database. Two major privacy principles have been considered in the literature, as far as the anonymization of trajectory micro-data is concerned.

- *Indistinguishability* commends that each record in a database must not be distinguishable from a large enough group of other records in the same database, called *anonymity set*. In the context of trajectory micro-data, the principle implicitly assumes that the adversary's knowledge is limited to some portion of the movement of the target individual. It ensures that such an adversary will not be able to pinpoint a single trajectory in the target database; instead, the attacker will retrieve the whole batch of indistinguishable trajectories of all users in the anonymity set. Ultimately, when applied at the record level, the principle effectively removes unicity in trajectory micro-data. Extensions to the baseline principle can also tackle homogeneity, by applying the same principle at the attribute level, and granting that a sufficiently large set of trajectories share not only mobility information but also sensitive attributes. OVerall, indistinguishability is a sound countermeasure against record and attribute linkage attacks.

- *Uninformativeness* enforces that the difference between the knowledge of the adversary before and after accessing a database must be small. It is apparent that uninformativeness is a much more general principle that does not make any assumption on the adversary knowledge. This principle is suitable to address probabilistic attacks, hence providing much stronger privacy guarantees than indistinguishability.

In addition to the principles above, a substantial amount of works adopt less rigorous privacy notions, and we group those under the following loose privacy principle.

- *Mitigation* aims at reducing circumstantial privacy risks associated with the data, without pursuing a well-defined privacy principle.

The privacy principles above constitute the first dimension of our taxonomy, reflected in the first column in Table 2. Note, however, that privacy principles are abstract definitions. In order to be applied in practical cases, they need to be specialized into *privacy criteria* that define the exact requirements that a database needs to meet in order to comply with

Table 2: Classification of the techniques proposed in the literature to anonymize trajectory micro-data databases. Our taxonomy is outlined by the first two columns, that tell apart anonymization solutions based on the privacy principle they adopt, and on the privacy criterion used to implement the principle. The last three columns indicate which class of attack each solution aims at countering, based on the adversary's objective (**O**) outlined in Section 2.1.1. Solutions based on the three privacy principles in the leftmost column are presented in Section 3.2 (mitigation), Section 3.3 (indistinguishability), and Section 3.4 (uninformativeness), respectively.

		Attacker objective (O)		
Privacy principle	**Privacy criterion**	Record linkage (*Re-identification*) **O.1**	Attribute linkage (*Homogeneity*) **O.2**	Probabilitic (*Inference*) **O.3**
Mitigation	Obfuscation	Srivatsa *et al.* [112]		
	Cloaking	Hoh *et al.* [68] Murakami *et al.* [90] Ma *et al.* [82] Rossi *et al.* [102]		
	Segmentation	Song *et al.* [111]		
	Swapping	Salas *et al.* [104]		
Indistinguishability	k-anonymity via spatiotemporal generalization	Yavonoy *et al.* [132] De Montjoye *et al.* [38] Zang *et al.* [134] Gramaglia and Fiore [61]		
	k-anonymity via suppression	Terrovitis and Mamoulis [116]		
	k-anonymity via generalization and suppression	Nergiz *et al.* [94] Monreale *et al.* [88] Gramaglia and Fiore [61]		
	k-anonymity via micro-aggregation and suppression	Domingo-Ferrer and Trujilo-Rasua [43] Torres and Trujilo-Rasua [118] Naini *et al.* [93]		
	Generalized k-anonymity with spatial uncertainty	Abul *et al.* [4] Abul *et al.* [5] Kopanaki *et al.* [75]		
	l-diversity, t-closeness	Tu *et al.* [122, 121]	Tu *et al.* [122, 121]	
Uninformativeness	(ϵ,δ)-differential privacy			Shao *et al.* [107]
	ϵ-differential privacy			Chen *et al.* [32] Chen *et al.* [31] Bonomi and Li [20] Qardaji *et al.* [98] Zhang *et al.* [136] He *et al.* [66] Mir *et al.* [87] Roy *et al.* [103] Gursoy *et al.* [64]
	Plausible deniability			Bindschaedler and Shokri [16] Bindschaedler *et al.* [17]
	$k^{\tau,\epsilon}$-anonymity	Gramaglia *et al.* [62]		Gramaglia *et al.* [62]

the principle. Privacy criteria are categorized as per the second column of Table 2. Next, we provide brief primers on each privacy principle, possibly introducing the key criterion used to implement it in the literature. Then, we discuss in detail mitigation techniques in Section 3.2, approaches based on indistinguishability in Section 3.3, and those aiming at uninformativeness in Section 3.4.

3.1.1 Mitigation of privacy risks

A large body of works does not target a rigorously defined privacy principle, rather aims at mitigating privacy risks in trajectory micro-data. Approaches in this category propose perturbation of the data records, reduction of the spatial or temporal resolution of the data, or arbitrary trimming of the trajectories. However, such strategies do not offer any formal guarantee in terms of privacy. A number of variants of these simple models exist, and we survey them in Section 3.2.

3.1.2 Indistinguishability via k-anonymity.

Indistinguishability is mainly implemented via *k-anonymity*, a privacy criterion first introduced by Sweeney [115] for relational micro-data that has also found wide application with trajectories. The idea behind k-anonymity is that any subset of the spatiotemporal points of each user in a trajectory micro-data database shall not be distinguishable from the spatiotemporal points of at least $k-1$ other users in the same database. Formally:

Definition 1. Let $\mathcal{D}$ be a database of trajectory micro-data and $LBQID$ the location-based quasi-identifier associated with it, and let $\mathcal{D}[LBQID]$ be the set of records returned by a query for $LBQID$ on $\mathcal{D}$. Then, $\mathcal{D}$ satisfies k-anonymity if and only if the records in $\mathcal{D}[LBQID]$ are at least k.

The k-anonymity criterion can be implemented with many and varied techniques, from generalization to microaggregation, and has been also augmented in several ways. These variants map to different rows under the indistinguishability privacy principle in Table 2, and we will discuss them in detail in Section 3.3.

3.1.3 Uninformativeness via differential privacy

Uninformativeness is typically achieved through *differential privacy*, whose original definition by Dwork [46] imposes that the result of a query on a differentially private database must yield only a small variation depending on whether a specific record is present or not in the database. Formally, ϵ-differential privacy, which is the standard form of differential privacy, is defined as follows:

Definition 2. A randomized algorithm $\mathcal{A}$ offers ϵ-differential privacy if for all datasets $\mathcal{D}'$ and $\mathcal{D}''$ differing in one element ($|\mathcal{D}' - \mathcal{D}''| = 1$), and for all subsets S of the output of $\mathcal{A}$, it holds $Pr[\mathcal{A}(D') \in S] \leq e^{\epsilon} \times Pr[\mathcal{A}(D'') \in S]$.

The concept of "small" difference between query results is embodied by the so-called budget parameter ϵ, which regulates the amount of diversity in the query result allowed when removing a single individual from the database. Thus, differential privacy realizes uninformativeness by ensuring that the additional knowledge gained by the adversary when he accesses the database is bounded to ϵ.

An important remark is that differential privacy, as defined above, is a condition on the algorithm used to extract information from the database, and not on the database itself. Therefore, differential privacy is not immediately related to PPDP of trajectory micro-data, rather to privacy-preserving data mining (PPDM). PPDM is a completely different problem from PPDP, as it assumes that the exact operations that will be run on the database are known a-priori, and can be included in the anonymization process. However, as we will see, this criterion can be adapted to the case where databases of trajectory micro-data are to be published for generic future mining purposes.

Moreover, differential privacy is not the only criterion that implements uninformativeness. Other models, based on extensions of k-anonymity, have also been proposed. We will review all anonymization techniques aiming at satisfying the uninformativeness principle for PPDP of trajectory micro-data in Section 3.4.

3.1.4 Literature classification

We classify the works in the literature according to the privacy criterion they implement, as listed in the rows of Table 2. However, we also complete our taxonomy with one additional dimension, orthogonal to privacy principles and criteria, *i.e.*, the type of attack on trajectory micro-data that each anonymization technique is intended to tackle. The columns of Table 2 tell apart different attacker objectives (**O**), categorized according to the discussion in Section 2.1.1. The labels within parenthesis report alternative names for these attacks that are frequently used in the literature on anonymization: specifically, record and attribute linkage are often referred to as re-identification and homogeneity attacks, respectively; inference is the terminology typically employed to indicate probabilistic attacks.

The taxonomy in Table 2 allows us to catalogue works that propose anonymization techniques for trajectory micro-data based on the combination of their underlying privacy principle/criterion and the type of attack they are effective against. We should note that some of the works discussed in this survey do not make their assumed attacker model explicit. However, the proposed anonymization model description implies the type of attack they could be used against. For example, in the cases of Torres and Trujillo-Rasua [118] and Kopanaki *et al.* [75], the proposed k-anonymity models naturally protect against record linkage attacks, even if not specified in the papers.

Although fairly sparse, the table highlights how the vast majority of the literature is focused on mitigating or preventing record linkage on published databases of trajectory micro-data. Also, some expected correlations emerge: indistinguishability is mostly suitable to counter record linkage, while uninformativeness tends to be used to develop solutions against probabilistic attacks. Interestingly, mitigation techniques can only cope with the simplest class of attack, *i.e.*, record linkage, due to their heuristic nature. Variations of these baseline matches of criterion and attack are rare, and we will detail them in our following discussion.

3.2 Solutions providing mitigation of privacy risks

We start our review of solutions for the anonymization of trajectory micro-data by presenting techniques that do not implement any well-defined privacy principle, rather mitigate privacy risks without theoretical or provable guarantees. In the following, we tell apart such heuristic solutions based on the type of transformation they perform on the data.

3.2.1 Obfuscation

A very simple solution consists in distorting location data by adding noise to it. The *value distortion* technique is originally introduced for privacy preserving data mining (PPDM) of location data by Agrawal and Srikant [7], and later formalized as *obfuscation* in LBS environments by Duckham and Kulik [45]. Srivatsa and Hicks [112] add different models of random noise to their social-graph representations of trajectories (see Section 2.7), and show[23] that they can reduce the success of record linkage attacks in a substantial way only if the level of noise is high.

3.2.2 Cloaking

Another baseline strategy is to reduce the granularity of the trajectory data in space or time, which is often referred to as *cloaking* as per the seminal work by Gruteser and Grunwald [63] in LBS systems. Hoh *et al.* [68] show that increasing the sampling interval from one to four minutes (*i.e.*, only retaining every fourth sample) in their trajectory micro-data[26] reduces home identification rates from 85% to 40%, although the risk is far from being removed. Murakami *et al.* [90] adopt a slightly different approach, and selectively remove a given fraction of points from the original trajectories, either randomly or so as to minimize the opportunities for linkage by an attacker (this second option assumes knowledge of the adversary's side information). The authors report that, by deleting up to 5 points from all trajectories in their reference databases[16], the performance of record linkage attacks are halved, yet remain high in absolute terms.

Ma *et al.* [82] tamper instead with the spatial dimension of trajectory micro-data, and show that, in the case of their datasets[6], reducing the geographical accuracy of the spatiotemporal points does not have a clear positive effect on unicity: the chances that a record linkage attack is successful stay above 50% when the adversary knows as little as 8 points of its target's mobility. In fact, in situations where the adversary knowledge is also inaccurate, a lower granularity may even lead to increased record linkage: the authors ascribe this effect to the fact that a coarser cell structure mitigates mistakes in the attacker's side information. Similar conclusions are drawn by Rossi *et al.* [102], who reduce the accuracy of GPS data by truncating the longitude and latitude values to increasingly fewer decimal places: this effectively allows them to consider geographical resolutions that range from around 1 m^2 to over 10 km^2. However, even at the lowest spatial granularity, 5% to 60% of users are still unique in the considered datasets[3].

3.2.3 Segmentation

A third straightforward technique is that suggested by Song *et al.* [111], *i.e.*, segmenting each trajectory and using a different pseudo-identifier for each segment. The rationale is that the unicity of a trajectory increases with its length, hence slicing each original trajectory into many output trajectories typically makes the latter less unique and easier to protect via anonymization. However, the authors show this simple technique cannot reduce unicity in a significant way: 80% of truncated trajectories in their dataset[27] are still unique even when they only span 6 consecutive hours. Moreover, this approach risks to dramatically reduce

[26] Tests are conducted on real-world GPS traces of vehicles in the Detroit area, tracked during a week with a frequency of 1 sample per minute.

[27] Experiments are conducted on CDR-based trajectories of 1.14 million users, tracked for one week, with a sampling interval of 15 minutes.

the utility of the trajectory micro-data, preventing many analyses that require complete movement information about each user.

3.2.4 Swapping

A recent work by Salas *et al.* [104] proposes a model where portions of the trajectories are iteratively swapped among users, so that the output trajectories are in fact composed of segments belonging to multiple actual users. The technique, named SwapMob, operates opportunistically on pairs of trajectories that come close enough to be swapped. Tests with real-life data[28] demonstrate that SwapMob effectively dissociates the segments of trajectories from the subject that generated them, significantly reducing the space for record linkage. Yet, an adversary knowing 10 spatiotemporal points is still able to link 42% of the users, and learn more than 50% of the original trajectories in 5% of cases. Also, it holds again the consideration that the output trajectory micro-data does not retain any utility for studies that require the possibility of following users for long, continued time intervals.

3.3 Solutions providing indistinguishability

Indistinguishability is the first proper privacy criterion that we consider in our survey. As already mentioned, k-anonymity is the de-facto standard privacy criterion for indistinguishability in trajectory micro-data. k-anonymity is attained by transforming the spatiotemporal points of the trajectories in the database, so that all points in every spatiotemporal trajectory are found in least $k-1$ other trajectories. Different types of transformations can be applied to the spatiotemporal points, telling apart the diverse methods to implement k-anonymity that are outlined by the first set of rows in Table 2, and that we will review in the rest of this Section.

Table 3 provides a summary of the main features of solutions proposed in the literature to implement k-anonymity in trajectory micro-data. It offers a quick outlook of the assumptions, approach and performance of each technique, and is thus a useful reference to start comparing different strategies for the k-anonymization of trajectory databases. Before delving into the details of these techniques, two important remarks about k-anonymity are in order, which apply beyond the context of trajectory micro-data.

First, the privacy level granted by k-anonymity is very much dependent on the value of k: in presence of a k-anonymous database, the probability of re-identification under a random guess by the adversary is $1/k$, hence k is inversely proportional to the chance of success of a record linkage. Yet, there is no clear consensus on which k is safe enough, and the values adopted in the literature tend to be application-dependent. Also, it should be noted that attaining higher k values typically reduces the utility of the trajectory micro-data, as it requires distorting the spatiotemporal points when applying the transformation. Again, this creates a trade-off that is not simply solved, and is highly use-case-dependent.

Second, k-anonymity has well-known and severe limitations. Basically, this privacy criterion offers strong protection against record linkage attacks only; however, it does not remove privacy risks associated to attribute linkage or any form of probabilistic attack. This has been repeatedly shown, considering, *e.g.*, attacks aiming at localizing users, or at disclosing their presence, meetings and sensitive places [109, 84, 62]. The fact that k-anonymity has been at times misunderstood or oversold as a comprehensive solution for

[28] Experiments are run on GPS trajectories of $10,357$ taxies in Beijing, China, during one week in February 2008. The database contains over 15 million spatiotemporal points, with an average sampling interval of 177 seconds and 623 meters.

Table 3: Comparative roster of the main features of the techniques proposed to achieve k-anonymization in trajectory micro-data databases. The columns indicate: *(i)–(ii)* the reference and acronym of the solution; *(iii)–(iv)* the type of trajectories they operate with and the LBQID they assume; *(v)–(vi)* the approach they adopt, including the distance metric between trajectories; *(vii)–(viii)* typical performance figures, in terms of removed spatiotemporal points, and resulting data quality.

Reference	**Name**	**Trajectory**	**LBQID**	**Approach**	**Pairwise trajectory distance metric**	**Suppressed points**	**Data quality ($k = 2$)**
Terrovitis and Mamoulis [116]	–	Spatial (discrete)	Subset of points	Suppression	Euclidean distance	30-50%	–
Yarovoy *et al.* [132]	–	Spatial	Subset of points	Generalization	Hilbert distance	–	7-62% query distortion
Nergiz *et al.* [94]	–	Spatiotemporal	Any	Generalization & suppression	Log cost metric (LCM)	3-4%	50-90% clustering accuracy
Monreale *et al.* [88]	KAM	Spatial	Any	Generalization & suppression	Longest common subsequence	–	0.5-0.7 clustering precision
Gramaglia *et al.* [61]	GLOVE	Spatiotemporal	Any	Generalization & suppression	Fingerprint stretch effort	–	1 km, 1 hour
						5%	0.5 km, 40 min
Naini *et al.* [93]	–	Spatiotemporal	Location histogram	Microaggregation	Normalized information loss	–	–
Torres *et al.* [118]	–	Spatiotemporal	Any	Microaggregation	Fréchet/Manhattan coupling distance	29%	0.2-0.95 S/T range query distortion
Domingo *et al.* [43]	–	Spatiotemporal	Any	Microaggregation & suppression	Synchronized trajectory distance	80%	2.4 km, 100 min
Abul *et al.* [4]	NWA	Spatial	Any	Microaggregation & suppression	Euclidean distance	–	Several km
Abul *et al.* [5]	W4M	Spatiotemporal	Any	Microaggregation & suppression	EDR / LSTD	5-20%	Several km, several hours
Kopanaki *et al.* [75]	WCOP	Spatiotemporal	Any	Microaggregation & suppression	EDR / LSTD	–	Several km, several hours

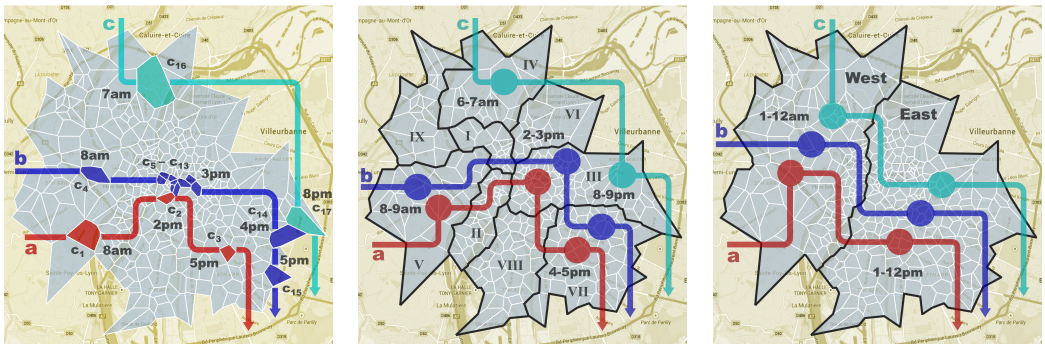

Figure 13: Example of k-anonymity via spatiotemporal generalization. (a) Original database: user locations are represented at cell level, and the temporal information has hourly precision. The trajectory micro-data of three users, a, b, and c are highlighted. (b) Spatiotemporal generalization: positions are recorded at the city administrative zoning level, and the time granularity is reduced to two hours. The trajectory micro-data of users a and b is now indistinguishable, and those two users are 2-anonymized. (c) Increased generalization: locations are limited to the Eastern or Western half of the city, and time has 12-hour granularity. All three users are indistinguishable and 3-anonymized. Reproduced in part from Gramaglia and Fiore [61], with permission.

PPDP of micro-data has led to diffused criticism on its use over the past years. Still, it remains a sensible privacy criterion within its scope of application [52]. Moreover, as we will see in the following, k-anonymity represents a basis on top of which more complex privacy-preserving solutions can be developed, so as to counter complex attacks that go beyond record linkage.

3.3.1 k-anonymity via spatiotemporal generalization

Spatiotemporal generalization is the baseline method used to achieve k-anonymity in trajectory micro-data. It reduces the spatial accuracy and temporal granularity of the spatiotemporal points in the trajectories contained in the target database, so that all points of each trajectory are indistinguishable from the points of $k-1$ other trajectories in the same database.

An illustrative example of spatiotemporal generalization of trajectory micro-data is shown in Figure 13. Despite its simplicity, the example highlights how achieving k-anonymity has a cost in terms of spatiotemporal accuracy of the database: ensuring that all trajectories are k-anonymized may force reducing the spatial and temporal accuracy of records up to the point where the trajectory micro-data becomes useless. Several works have quantified such a trade-off between privacy and utility, when enforcing k-anonymity on trajectory micro-data. Zang and Bolot [134] investigate how unicity decreases as the spatial granularity of a trajectory micro-data database is lowered; this is equivalent to assessing which reduction of geographical accuracy is needed to attain 2-anonymity, *i.e.*, the minimum level of k-anonymity that removes unicity and grants indistinguishability. The study unveils how unicity is very hard to eliminate from spatiotemporal trajectories: if the three most frequently visited locations are known to the attacker, the only way to grant 2-anonymity is to just publish user movements among US States, even for very large databases[19].

A more thorough evaluation is carried out by De Montjoye et al. [38], who establish an empirical relation between unicity and the spatiotemporal resolution of the trajectory micro-data dataset[2]. They find that unicity decreases as a power law of both spatial and tem-

poral granularity: this implies that the cost in terms of data accuracy loss increases very quickly as a larger fraction of trajectories in a same dataset becomes 2-anonymized, and 2-anonymizing the last percentile of trajectories may require a generalization that is orders of magnitude stronger than that needed to 2-anonymize the first percentile. Even worse, the power law exponent decreases linearly in the number of spatiotemporal points known to the attacker, and a few additional points make k-anonymization much more expensive.

The reasons behind the high cost of k-anonymization in trajectory micro-data are studied by Gramaglia and Fiore [61]. They show that spatiotemporal trajectories in large databases[29] typically have a substantial fraction of points that are easily generalized into those of a single other trajectory in the same database; however, they also have a small but non-negligible set of points that are very unique, and hard to hide into another trajectory. Since k-anonymity is only achieved when all points are made indistinguishable, the cost of generalization is driven by such hard-to-hide points. This study also investigates cases when $k > 2$, demonstrating how higher k values induce a superlinear growth in the utility loss.

Based on these observations, the authors propose GLOVE, an algorithm that achieves k-anonymity of trajectory micro-data via spatiotemporal generalization, at a sensibly lower cost than the approaches by Zang and Bolot [134] and De Montjoye et al. [38]. The key idea is to operate generalization on each spatiotemporal point individually, instead of applying the same reduction of granularity to all points of all trajectories, as done previously. Based on this intuition, the authors define a pairwise trajectory similarity metric named *fingerprint stretch effort*, which quantifies the loss of spatial and temporal granularity needed to hide each sample of one trajectory into the closest sample of the other trajectory. Then, a simple greedy clustering based on fingerprint stretch efforts lets GLOVE 2-anonymize a complete trajectory micro-data database with tens of thousands of records[29] while retaining median resolutions of 1 km in space and 1 hour in time. Interestingly, performance tends to improve, *i.e.*, the data loss is reduced, as the database size grows.

All the above studies consider the k-anonymization of full-length spatiotemporal trajectories, *i.e.*, they assume that all spatiotemporal points of each trajectory must be indistinguishable from the points of other $k-1$ trajectories in the same database. Yarovoy et al. [132] relax this challenging constraint, and study k-anonymity in a setting where a known subset of the points of each trajectory is used as LBQID: therefore, only that subset needs to be k-anonymized, for each trajectory. This significantly reduces the cost of generalization, since the LBQID contains a number of points much smaller than that in the complete trajectory; however, it also introduces the new problem of selecting the so-called anonymization group, *i.e.*, the set of $k-1$ records within which the LBQID of each trajectory must be made indistinguishable. Indeed, a careless choice may lead to a successful record linkage by an adversary with knowledge of the LBQIDs of multiple users.

The authors then propose algorithms that select anonymization groups so as to ensure proper k-anonymity in this scenario. A first solution identifies sets of k records based on a Hilbert distance measure, and ensures that every trajectory in a group is generalized with respect to all LBQIDs of all other trajectories in the same group. A second solution operates on a per-record basis rather than on a per-group basis: for each trajectory i in the database, it finds suitable trajectories j to enforce symmetric k-anonymization of the LBQID, *i.e.*, it generalizes the LBQID points of i into those of j and vice-versa. Tests with databases featuring fixed temporal periodicity[30] show that, when LBQIDs include between

[29] The authors employ two trajectory micro-data databases from two-week nationwide CDRs of 82,000 and 320,000 mobile network subscribers located in Ivory Coast and Senegal, respectively, released in the context of the D4D Challenge [18].

[30] The authors employ a real-life dataset includes GPS trajectory micro-data of cars in Milan, Italy. The data is

5% and 50% of the total spatiotemporal points of each trajectory, the proposed schemes achieve k-anonymity, with k from 2 to 32; however, they also induce spatial distortions that cause 7%–62% of location-based queries to fail in the anonymized database.

3.3.2 k-anonymity via suppression

A different technique to achieve k-anonymity is *suppression*, which removes spatiotemporal points from the original trajectories. Terrovitis and Mamoulis [116] propose an algorithm that iteratively removes points from trajectories, simplifying the movement description until k-anonymity is satisfied. At each iteration, all points that break k-anonymity are identified, and the one entailing minimum Euclidean distortion is selected for removal. However, the simplicity of the solution entails strong assumptions on the trajectory micro-data format and attacker model in order to produce reasonable results: *(i)* trajectories are purely spatial, *i.e.*, do not have a temporal dimension; *(ii)* space is discretized in a finite number of locations; and, *(iii)* adversaries are in a small number, and their exact knowledge is available and can be used as an input to the anonymization process. The latter point implies that the k-anonymization is limited to a very specific set of LBQIDs, *i.e.*, sequences of points. These aspects are reflected in the performance evaluation, carried out with synthetic data[31], where 2 to 7 adversaries have side information (known to the anonymization algorithm) of all points in 1 out of 100 total locations.

3.3.3 k-anonymity via generalization and suppression

Generalization and suppression can in fact be used jointly. In the light of the analysis by Gramaglia and Fiore [61], suppression can be highly beneficial to k-anonymization: indeed, discarding the small fraction of unique points may take away a substantial portion of the diversity among trajectories, whose generalization then retains a higher accuracy level.

The first example of approach based on suppression is that by Nergiz *et al.* [94]. The solution is close in spirit to GLOVE, as it also relies on per-spatiotemporal point generalization. However, *(i)* it enforces that no two points of one trajectory can be generalized with a single point of the other trajectory, which leads to suppression in presence of trajectories with a non-matching number of points; *(ii)* it is based on a different pairwise trajectory similarity metric, named *log cost metric*, which scales logarithmically the loss of spatial and temporal accuracy and accounts for suppressed points. Evaluations with real-world and synthetic trajectory data[32] show that the proposed solution can achieve 2-anonymization by suppressing 3-4% of data, while 2-anonymization has a much higher cost typically around 25% of removed points. Although the authors do not report on exact error figures of the anonymized trajectories in space and time, they show that the results of one specific analysis, *i.e.*, clustering, are preserved with precision and recall in the range 50-90%.

A different solution is proposed by Monreale *et al.* [88], which however only operates on spatial trajectories that do not have time labels. Their strategy involves a first phase in which space is discretized via a Voronoi tessellation: the seeds are obtained by clustering all

pre-processed to include one sample every 5 minutes, resulting in more than 45,000 trajectories and 2009 timestamps. A synthetic dataset is also used; it is created using Brinkhoff's generator [22] and includes 150,000 trajectories with 400 timestamps over the road-network of Oldenbur, Germany.

[31] The authors use 2,000 to 15,000 trajectories returned by the Brinkhoff's moving object generator [22] in Oldenburg, Germany.

[32] The real dataset includes 1,000 GPS trajectories of taxis in Milan, Italy, with a total of 98,544 samples, collected as part of the GeoPKDD project [56]. The synthetic dataset comprises 1,000 trajectories and 70,118 samples, which are obtained using Brinkhoff's moving object generator [22].

spatiotemporal points in the dataset in a way that a minimum number of trajectories is ensured to flow between any two adjacent Voronoi cells. Trajectories are then all generalized in space according to the voronoi tessellation. Two algorithms are proposed for the second phase, which actually implements k-anonymity. KAM_CUT is intended for dense datasets: it first creates an efficient tree structure of trajectories, where common subtrajectories are the parent nodes to child nodes representing more complete (but diverse) subtrajectories of the same users; it then traverses the tree by suppressing branches shared by less than k trajectories. KAM_REC extends the above for sparse datasets: to this end, it tries to reinsert the pruned sub-trajectories back in the tree, by finding their longest subsequence of points that maps to some popular sub-trajectory either still in the tree, or shared by at least k other trimmed sub-trajectories. Experiments are run on measurement data[33], and show that KAM_CUT and KAM_REC attain precision and recall that typically are in the range 0.5-0.7, for $k \in [2, 30]$. In these tests, suppression removes between 10% ($k = 2$) and 80% ($k = 16$) of the trajectories.

Also Gramaglia and Fiore [61] extend GLOVE so as to include suppression. This is realized by removing points that induce an over-threshold generalization cost during the calculation of the fingerprint stretch effort. Tests with real-world data[29] show that suppressing 5% of points reduces the loss of accuracy in space and time by approximately 30-50%.

3.3.4 k-anonymity via microaggregation and suppression

Microaggregation is a family of two-step perturbative Statistical Disclosure Control (SDC) methods that can be used to implement k-anonymity in trajectory micro-data. In the first step (partition), the set of original trajectories is clustered based on similarity, so that each cluster has cardinality at least k. In the second step (aggregation), the trajectories in a cluster are replaced by a cluster prototype, computed through an operator over the spatiotemporal points in the cluster. Overall, this effectively k-anonymizes the dataset, by making all k or more trajectories in a same cluster identical to the prototype.

A seminal work partially based on microaggregation of spatiotemporal trajectories is that by Domingo-Ferrer and Trujilo-Rasua [43]. The authors introduce a new pairwise trajectory similarity metric, which we refer to as *synchronized trajectory distance*. The distance is computed in two steps: first, trajectories are synchronized, *i.e.*, linearly interpolated and sampled with an identical periodicity; second, the total Euclidean distance between contemporary points is computed. In the case where the two trajectories span different time intervals (*i.e.*, the times of their first and last points do not match), all non-overlapping points are suppressed, and the distance metric is divided by the percentage of suppressed points as a similarity penalty. An interesting property of this metric is that it satisfies the triangle inequality, which allows speeding up calculations of all-pair distances.

The SwapLocation algorithm employs a legacy clustering technique based on the synchronized trajectory distance. Then, for each trajectory in a cluster, it swaps all of its spatiotemporal points with points of other trajectories in the same cluster. The exchange of points must respect configurable thresholds in space and time, which can be seen as the maximum allowed distortion of each point in a trajectory; also, a point is suppressed if no switch is possible under the imposed thresholds. We remark that the swap operation is not fully

[33] The dataset consists of 5,707 GPS trajectories of cars moving around Milan, collected by the automotive service provider Octotelematics within the GeoPKKD project [56]. Note that the trajectory micro-data is pre-processed by splitting trajectories when two consecutive points are too far in space and time, resulting in more than 45,000 fairly space- and time-continuous trajectories in the final dataset.

consistent with the standard microaggregation strategy, however the rest of the solution is coherent with such a model.

The authors use both synthetic and real-life datasets[34] to assess the performance of SwapLocation. The solution imposes significant suppression in the synthetic dataset and with $k = 10$: a 1-km spatial threshold in the swap operation leads to removing 50% of trajectories and 80% of points. Under a 3-km threshold, suppression is reduced to 5% for trajectories, but it remains almost unchanged for points; moreover, the average distortion is at 1 km approximately. In the case of real-world data, and $k = 2$, 29% of points are suppressed and the mean spatial distortion is at 2.4 km. All results refer to a 100-minute temporal granularity of the anonymized dataset.

A main limitation of the synchronized trajectory distance is its reduced capability to manage pairs of trajectories that do not perfectly overlap, which leads to substantial suppression of points. Torres and Trujillo-Rasua [118] propose a novel pairwise metric that is based on the *Fréchet/Manhattan coupling distance* [48] and overcomes such an issue. The metric is based on the notion of coupling, *i.e.*, a sequence of matching point pairs (one per trajectory) that respects the time ordering of points and ensures that all points are considered. The Fréchet/Manhattan coupling distance is then the coupling that minimizes the sum of Euclidean distances between matched points. Notably, the metric does not formally account for temporal distances between points, and its computation has a limited complexity $\mathcal{O}(pq)$, where p and q are the number of points in the two input trajectories.

The Fréchet/Manhattan coupling distance is then leveraged in a randomized clustering algorithm, and prototype trajectories are obtained for each cluster by means of an obfuscation process: linear interpolation and downsampling are first used to homogenize trajectories in time, and spatial averaging is then adopted to compute prototype locations at each instant. In order to evaluate their solution, the authors employ synthetic data[35] and perform spatiotemporal range queries [119], which aim at inferring if a specific trajectory has some or all points inside a target region during a given time interval. Distortions of query results are between 0.2 and 0.95 for $k = 2$, which are lower than those induced by the approaches proposed by Nergiz *et al.* [94] and Domingo-Ferrer and Trujilo-Rasua [43].

A special case of microaggregation is considered by Naini *et al.* [93], who adapt the approach to the case where histograms of popular locations are to be k-anonymized, rather than the complete trajectories. The authors then use a normalized information loss metric that is suitable for the probability distributions they target, and run a legacy clustering algorithm based on that metric. Evaluations with real-world datasets[36] show that reidentification is impaired by k-anonymity with $k > 10$, however that comes at a substantial information loss above 65%.

3.3.5 Generalized k-anonymity with spatial uncertainty

Abul *et al.* [4] propose a generalization to the k-anonymity criterion, by assuming that published trajectories must be indistinguishable within an *uncertainty threshold* δ. In other words, the strong indistinguishability requirement of k-anonymity that each trajectory

[34]The synthetic dataset is obtained via the Brinkhoff's moving object generator [22], and consists of 1,000 trajectories which visit a total of 45,505 locations in Oldenburg, Germany. The real-life dataset consists of 4,582 GPS cab traces collected in San Francisco, USA, with a mean of 94 points each.

[35]The data were generated with Brinkhoff's moving object generator [22], and consist of 5,000 trajectories containing 492,105 locations in Oldenburg, Germany, with 98,421 locations per trajectory on average.

[36]The authors use three different datasets: the two-week CDR of 50,000 Orange customers in Ivory Coast released within the context of the D4D Challenge [18], the Web browsing history of 472 users from the Web History Repository [67], and 154 users with an average of 15.4 weeks of data each from the GeoLife experiment [137].

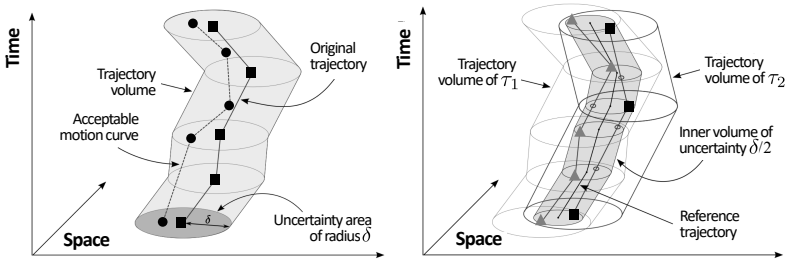

Figure 14: Example of $(2, \delta)$-anonymity. Left: an original trajectory (solid line with square markers) is assigned an uncertainty radius δ around each of its spatiotemporal points, which yields a trajectory volume resulting from joining all uncertainty areas. An acceptable distorted motion curve for the original trajectory (*e.g.*, the one denoted with dashed line and circle markers) lies entirely within such a volume. Right: the original spatiotemporal points (black squares and grey triangles, respectively) of two trajectories τ_1 and τ_2 are included into volumes (black for τ_1, grey for τ_2) characterized by uncertainty δ. The inner cylindrical volume is drawn with an uncertainty $\delta/2$ around a reference trajectory corresponding to the arithmetic mean of the points of the two original trajectories ((interpolated if necessary): a minimum translation of points within such inner volume ensures $(2, \delta)$-anonymity. In the example, no translation is needed since only two trajectories are concerned. Adapted from Abul *et al.* [4].

must be identical to at least $k - 1$ others (see Definition 1 in Section 3.1.2) is relaxed, allowing a spatial distance up to δ among each point of an anonymized trajectory and the concurrent ones of the $k - 1$ other trajectories. The intuition is that uncertainty among points within a geographical distance δ still provides sufficient protection, while it requires less distortion. The generalized criterion is named (k, δ)-*anonymity*, and its concept is illustrated in Figure 14. The criterion offers inherently reduced privacy guarantees with respect to pure k-anonymity: as shown by Trujillo-Rasua and Domingo-Ferrer [120], the two criteria match only under perfect spatial identity of the anonymized trajectories, *i.e.*, $\delta = 0$.

In the original work by Abul *et al.* [4], an algorithm named Never Walk Alone (NWA) is proposed to implement (k, δ)-anonymity in uniformly sampled trajectory micro-data while minimizing distortion. NWA is organized in three phases. In the first phase, trajectories are trimmed so that they start and end at a limited set of time instants, and are then divided into groups with identical time spans. In the second phase, trajectories in a same group are separated into clusters of cardinality k based on Euclidean distance, and possibly suppressed if the operation cannot be performed while keeping distances below an adjustable threshold. In the third and final phase, NWA performs, for each time instant and each cluster, a minimum translation of spatial points of all trajectories, so that they are within a distance $\delta/2$ from their arithmetic mean, as exemplified in Figure 14. It is apparent that NWA performs a partial microaggregation, and can be classified along solutions in Section 3.3.4 when $\delta = 0$.

Experiments with both real-world and synthetic trajectory datasets[37] show that NWA attains (k, δ)-anonymity by inducing spatial displacements from several kilometers to several tens of kilometers, as k grows from 2 to 100, and δ increases from a few hundreds to several thousand meters.

[37] Real-life trajectory micro-data consists of 273 trajectories of trucks [50]. The second one has been generated using Brinkhoff's network-based synthetic generator of moving objects [22], and contains 100,000 trajectories during one day in the city of Oldenburg, Germany.

NWA only operates on trajectories with identical, periodic sampling. To overcome this important limitation, Abul *et al.* [5] also propose an improved method that can achieve (k, δ)-anonymity with proper spatiotemporal trajectories, named Wait For Me (W4M). W4M largely builds on NWA, but instead of relying on Euclidean distance it uses pairwise trajectory distance measures that account for different temporal samplings of the input trajectories. The considered metrics are: the *Edit Distance on Real sequences (EDR)* [30], which targets quality preservation but has complexity $\mathcal{O}(pq)$, where p and q are the number of points in the two input trajectories; and the linear spatiotemporal distance (LSTD), which is designed for efficiency, having linear complexity $\mathcal{O}(p+q)$. Adopting EDR and LSTD also has the advantage that the first phase of NWA can be skipped, and all trajectories can be processed at once in the clustering phase. Tests with heterogeneous datasets[38] prove that W4M achieves substantially lower distortion of spatiotemporal range queries [119] than NWA, and induces spatial and temporal translations in the order of several kilometers and hours, respectively. Also, the approach suppresses 5% of points and creates 5% new points to attain $k = 2$; these figures grow up to 20% when $k > 30$.

Further extensions to NWA and W4M are proposed by Kopanaki *et al.* [75], who introduce a suite of algorithms under the denomination of Who-Cares-about-Others'-Privacy (WCOP). The variants of WCOP allow accounting for *(i)* personalized values of k and δ that vary for each trajectory, *(ii)* a temporal segmentation of the trajectories such that each time segment of the dataset can be anonymized in isolation, and *(iii)* bounded distortion of the output trajectories.

3.3.6 Indistinguishability beyond k-anonymity: l-diversity and t-closeness

All anonymization solutions presented above implement indistinguishability at the record level, hence are suitable countermeasures in presence of record linkage attacks. However, they do not address the indistinguishability of attributes, leaving the door open to the attribute linkage (or homogeneity) attacks discussed in Section 2.9 and Section 2.1.1. This holds for both cases where sensitive attributes are separated from or embedded in the trajectory data, as per the discussion in Section 2.1.1. As an example, let us consider the case where the adversary is interested in finding some sensitive location embedded in the trajectory of the target user: by accessing a k-anonymized database with a few known spatiotemporal samples, the attacker would retrieve k movement patterns; if, however, they all include the position of a gentlemen club, a privacy breach occurs.

In fact, it is well known that k-anonymity is not a sufficient criterion in the case of attribute linkage, which asks instead for more complex privacy definitions. Specifically, popular criteria designed to counter attribute linkage attacks are *l-diversity* [83] and *t-closeness* [79]. The former assumes that a precise set of attributes (either separated from or embedded in the trajectory data, as per the discussion in Section 2.1.1) is identified as sensitive: then, any trajectory must be indistinguishable from a set of others whose sensitive attributes are sufficiently different from those of the original trajectory. Formally:

Definition 3. Let $\mathcal{D}$ be a database of trajectory micro-data, which includes a set of sensitive attributes; also, let $LBQID$ be the location-based quasi-identifier associated with $\mathcal{D}$, and $\mathcal{D}[LBQID]$ the set of records returned by a query for $LBQID$ on $\mathcal{D}$. Then, $\mathcal{D}$ is said to satisfy l-diversity if and only if the records in $\mathcal{D}[LBQID]$ contain at least l "well-represented"

[38] The authors employ a real-world dataset of 45,000 GPS trajectories of cars in Milan, Italy, collected during one week by the GeoPKDD project [56], and synthetic data generated with Brinkhoff's network-based simulator of moving objects [22] for 100,000 trajectories over one day in Oldenburg, Germany.

values for the sensitive attributes. Multiple notions of the concept of "well-represented" are possible, the simplest one being that at least l distinct values for the sensitive fields be present in $\mathcal{D}[LBQID]$.

A step further, t-closeness imposes a statistical constraint on the sensitive attributes, rather than the numerical one defined by l-diversity. The reason is that in practical cases the semantics of the attribute values are critical: for instance, a query returning a set of records with l different but correlated attribute values (*e.g.*, l variants of the same rare illness) satisfy l-diversity but still reveals sensitive information about the target user (*e.g.*, the fact that she suffers from the rare illness). To avoid these situations, t-closeness ensures that there is no substantial statistical difference between the attribute values in every set of indistinguishable users and those in the whole user population. Formally:

Definition 4. Let $\mathcal{D}$ be a database of trajectory micro-data, which includes a set of sensitive attributes; also, let $LBQID$ be the location-based quasi-identifier associated with $\mathcal{D}$, and $\mathcal{D}[LBQID]$ the set of records returned by a query for $LBQID$ on $\mathcal{D}$. Then, $\mathcal{D}$ is said to satisfy t-closeness if and only if the records in $\mathcal{D}[LBQID]$ contain sensitive attributes whose distribution has a distance lower than t to the distribution of the attributes in the whole $\mathcal{D}$.

While the problem has long been addressed in the context of LBS, for instance in the early study by Riboni *et al.* [99], the only works to date that tackles the anonymization of trajectory micro-data in a way to achieve both l-diversity and t-closeness are those by Tu *et al.* [122, 121]. The authors focus on semantic attacks where the sensitive information is embedded in the spatiotemporal points, and corresponds to the points of interest (PoI) present in a target geographical region. Therefore, they propose an algorithm that builds on GLOVE by Gramaglia and Fiore [61]; as such, it leverages both generalization and suppression of samples, however these operations are augmented to ensure that each generalized sample fulfills the $t - closeness$ (and, implicitly, l-diversity) criterion. Specifically, the difference between the PoI distributions within each sample and in the whole database, measured in terms of Kullback–Leibler (KL) divergence, must be below a threshold t. The performance evaluation with measurement data[39] shows that the proposed solution can reduce KL divergence by a factor three while sacrificing an additional 30% of the spatial and temporal resolution over the baseline k-anonymity granted by GLOVE.

3.4 Solutions providing uninformativeness

The second important privacy principle explored in the literature is that of uninformativeness. This principle aims to cope with probabilistic attacks and has received substantial attention in recent times. As anticipated in Section 3.1.3, the de-facto standard criterion to implement uniformativeness is differential privacy, a popular privacy criterion first introduced by Dwork *et al.* [46] for PPDM. Implementing differential privacy is especially elegant and simple in presence of algorithms that execute numeric or categorical queries. In the former case, the output is a vector of scalars, and differential privacy is obtained by the *Laplacian mechanism*, which adds Laplacian noise to such an output as first proposed by Dwork *et al.* [46]. In the second case, the output is a probability distribution over a discrete, finite set of outcomes, and differential privacy is obtained by randomizing the probability according to an *exponential mechanism*, as first explained by McSherry and Talwar [86].

[39] Two datasets are used for the evaluation. The first dataset is from a Chinese cellular network in Shanghai, and covers one week of data generated by 5,9 million users with an average number of 261 points each. The second dataset is from 15,500 users of a mobile application in Shanghai, and features a mean of 496 points per user.

In both situations above, the level of noise or randomization is calibrated according to ϵ, as well as to the maximum difference among all possible outputs when a single record is removed. In addition, under such query models, differentially private algorithms enjoy composition properties that describe how multiple queries consume the budget ϵ; this allows calibrating noise to the type and frequency of queries permitted on the database.

In the context of trajectory data, differential privacy has been successfully used to guarantee location privacy, *i.e.*, warranting that queries on single spatiotemporal points satisfy the uninformativeness principle. Criteria like *geo-indistinguishability*, first introduced by Andrés *et al.* [10], or based on the location-privacy metrics proposed by Shokri *et al.* [110] adapt differential privacy to the specific case of location data. A number of works have implemented and possibly enhanced the criteria above, including those by Assam *et al.* [11], Chatzikokolakis *et al.* [27], Bordenabe *et al.* [21], Xiao and Xiong [129], or Ngo and Kim [96]. However, as explained in Section 1.3, solutions that anonymize queries on instantaneous locations are relevant for LBS, but not for trajectory PPDP.

Closer to our context of data publishing, a fairly large body of works have concentrated on PPDP of aggregate statistics from trajectory micro-data. A commonly studied class of aggregates is that of spatial densities, especially in the form of *quadtrees*, *i.e.*, hierarchical spatial structures that allow for efficient querying: solutions such as those proposed by Cormode *et al.* [36], Qardaji *et al.* [98], Li *et al.* [78] or by Zhang *et al.* [136] allow generating differentially private density databases from the actual trajectories, which can then be publicly released and safely queried. Extensions, such as those by Acs and Castelluccia [6] or Alaggan *et al.* [9], consider spatiotemporal densities from trajectory micro-data, developing solutions that account for the temporal dimension of the aggregate statistics in addition to the spatial one. Other classes of trajectory data aggregates that can be transformed to meet differential privacy guarantees include weighted spatial graphs that describe transit counts between locations, such as those considered by Brunet *et al.* [23], or histograms, such as those assumed by Hay *et al.* [65]. Further investigations, *e.g.*, by Kellaris *et al.* [71] or Cao *et al.* [24], adapt the techniques above to the case of streaming data, where privacy-preserving spatial density information needs to be continuously published. Nevertheless, these works do not allow releasing spatiotemporal trajectories, but only their density or count statistics; hence, they are beyond the focus of our review on PPDP of trajectory micro-data.

When applied to our target milieu, *i.e.*, publishing actual trajectory data, differential privacy recommends that the output of an algorithm run on the released database is not affected by any single original trajectory beyond the privacy budget ϵ. Unfortunately, due the very high dimensionality of each trajectory, there is no current method to achieve such a goal by directly adding noise to the trajectory micro-data with existing mechanisms such as Laplacian or exponential. Therefore, two alternative approaches have been explored: *(i)* considering softened notions of differential privacy; or, *(ii)* generating synthetic trajectories that mimic the properties of true individual user movements yet ensure that the differential privacy criterion is fully met.

Below, we review solutions that adopt the first strategy in Section 3.4.1, and present works that instead take the second approach in Section 3.4.2. We also present a couple of works that adopt other criteria than differential privacy to realize uninformativeness, in Sections 3.4.3 and 3.4.4.

3.4.1 (ϵ, δ)-differentially private trajectory micro-data

A weaker notion of differential privacy that has been successfully adopted for the PPDP of trajectories is (ϵ, δ)-*differential privacy*. This is a relaxation of the basic notion of differential

privacy provided in Section 3.1.3 (which we recall to be also referred to as ϵ-differential privacy), where privacy breaches are allowed to occur with a (small) probability δ. Formally:

Definition 5. A randomized algorithm $\mathcal{A}$ offers (ϵ, δ)-differential privacy if for all datasets $\mathcal{D}'$ and $\mathcal{D}''$ differing in one element ($|\mathcal{D}' - \mathcal{D}''| = 1$), and all subsets S of the output of $\mathcal{A}$, then $Pr[\mathcal{A}(D') \in S] \leq e^{\epsilon} \times Pr[\mathcal{A}(D'') \in S] + \delta$.

Shao *et al.* [107] propose techniques that achieve (ϵ, δ)-differential privacy by combining trajectory sampling and interpolation, either in this order (a-priori) or in the reverse one (a-posteriori). The sampling phase realizes a $(0, \delta)$ form of differential privacy, by preserving one original point in every $1/\delta$: these points are publicly disclosed, and represent the privacy breach. The interpolation (a classic cubic Bézier) instead completes the gaps in between the retained points; by the composition properties of differential privacy, such a deterministic operation preserves the privacy properties of the sampling. Then, under the important assumption that the initial and final points of each trajectory are publicly known and their disclosure does not represent a privacy breach, both strategies attain $(0, \delta)$-differential privacy. Experiments[40] show that the a-posteriori method tends to have better results in terms of average error when querying the privacy-preserving database.

3.4.2 Differentially private synthetic trajectory micro-data

Proper differential privacy can be ensured by a different process where *(i)* some representation of the original trajectory micro-data is randomized so as to meet differential privacy requirements, and *(ii)* synthetic trajectories are derived from such representations. Then, databases of synthetic trajectories can be distributed with provable privacy guarantees.

Representing trajectory micro-data as trees. The first work to adopt the methodology above is that by Chen *et al.* [32]. They model the original database as a *prefix tree*, *i.e.*, a hierarchical structure where trajectories are grouped based on matching location subsequences whose length grows with tree depth[41]. A privacy-preserving version of the prefix tree is then obtained by considering multiple levels of spatial generalization based on a predetermined location taxonomy, and iterating on the following operations at each prefix tree layer. First, nodes are created for all locations at the highest level of generalization, as children of each leaf from the previous iteration; second, Laplacian noise is added to the count of trajectories associated to each generalized node at the current prefix tree layer; third, nodes with a noisy count below a tunable threshold are not expanded further, while nodes with noisy counts above threshold generate children nodes for all locations at the following level of generalization. The process is repeated from the second step above. Iterations conclude once a user-defined tree height is reached, with Laplacian noises set so that the total privacy budget ϵ is equally divided across all tree levels and nodes within each level. An example is provided in Figure 15, plots (a)-(c).

Then, differentially private synthetic trajectories can be derived from the sanitized prefix tree. To this end, the tree is pruned so that only nodes at the lowest level of generalization (*i.e.*, retaining the maximum spatial granularity) are preserved. Then, the noisy counts associated to such nodes are made consistent across levels, ensuring that the count of each node is not less than the sum of counts of its children nodes. Finally, the synthetic trajectories are generated by visiting the resulting prefix tree. An example is provided in Figure 15, plots (d)-(e).

[40] The study uses one-hour GPS data of ships in the Singapore Straits.

[41] Although we present it in the context of trajectory micro-data, the approach is general, and can operate on any type of sequential data.

We remark that the solution is introduced for trajectories that are defined on a discrete space, but it can be extended to include discrete time information as well. Tests in a real-world case study[42] show that the private synthetic trajectories can be mined to count passengers at stations, as well as to identify frequent sequential patterns of public transport usage, with a relatively low error.

Chen *et al.* [31] propose a variant to the strategy above, where the main difference is that the prefix tree is replaced by an n-*gram* representation. This probabilistic model describes trajectories as transition probabilities based on a past history of $(n-1)$ locations, *i.e.*, correspond to a Markovian model of order $(n-1)$. Figure 16 illustrates the concept of n-grams. The rest of the procedure is equivalent, by deriving a private prefix tree from the n-grams rather than from the original trajectories, and skipping intermediate generalization. Specifically, properly calibrated Laplacian noise is added to the counts of all 1-grams, expanding them to 2-grams only if their noisy count is above a threshold. Then, the procedure is repeated for always longer n-grams, descending into the prefix tree structure. The final differentially private variable-length n-grams can then be publicly released, or used to generate synthetic trajectories. Experiments with real-life datasets[43] show that also in this case count queries and frequent pattern mining run on the synthetic data return reliable results.

Various refinements of the techniques above are proposed, *e.g.*, by Bonomi and Li [20] and Qardaji *et al.* [98]. Note however, that these works aim at developing differentially private tree synopses of trajectory databases, but do not leverage them for the generation of synthetic trajectory micro-data. Instead, Zhang *et al.* [136] propose to extend PrivTree – a method they originally introduced for privacy-preserving release of spatial density data – to the case of synthetic sequential data generation. They adopt a *prediction suffix tree* model of trajectory micro-data that is similar in spirit to the prefix tree considered by Chen *et al.* [32]; however, adapting PrivTree to work on prediction suffix trees allows the authors to remove two limitations of previous techniques. First, it automatically adapts the tree height to the data, which is thus not a fixed parameter anymore; second, the decision of expanding a tree node is not based on a simple count, but adopts a more advanced strategy that also accounts for the entropy of the eventual children nodes. A comparative evaluation against the solution proposed by Chen *et al.* [31] proves that synthetic trajectories generated by PrivTree from real-life sequential data[44] allow for a 10% or higher improvement in *(i)* top-k frequent string mining, and *(ii)* approximating the distribution of sequence lengths.

He *et al.* [66] demonstrate that the approaches above work well with coarse trajectories defined on small location domains, but fail to scale to realistic database where fine-grained trajectories unfold over moderately large geographical span. The reason is that the representations used by Chen *et al.* [31] grow in size as a power law of the number of discrete locations, with an exponent equal to the depth of the prefix tree. Therefore, the authors propose to generate multiple prefix trees, each referring to a different spatial resolution; each transition in a trajectory contributes to one specific tree, based on the travelled distance (*i.e.*, low-resolution trees for long distances, and high-resolution trees for short distances). This results in multiple trees with a very small branching factor each, and in a significant reduction of the overall number of counts maintained. Then, the usual procedure of adding

[42] The authors employ information collected by the Société de Transport de Montréal (STM) about the transit history of passengers in the underground and bus networks of Montreal, Canada. The data contains over 1.5 million trajectories, with an average of around 5 locations each, out of a universe of 90 an 121 locations.

[43] The authors use the same STM dataset described in footnote 42, as well as 989,000 sequences of URL categories browsed by users on msnbc.com [42] with a mean length of 5.7 categories in a total set of 17.

[44] The study uses 80,362 learners' sequences of activities (among 8 possible states) on a MOOC platform, as well as 989,818 sequences of URL categories browsed by visitors at msnbc.com during a 24-hour period.

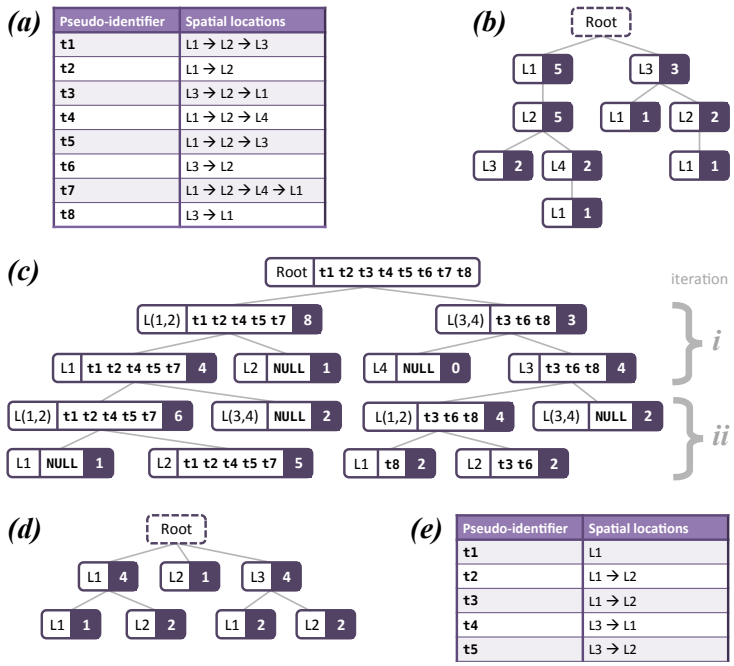

Figure 15: Toy example of the approach by Chen *et al.* [32] to generate differentially private synthetic trajectories. (a) Original spatial trajectory database. (b) Prefix tree structure summarizing the original database. (c) Differentially private prefix tree with height 2, under two levels of generalization: high-level generalized locations $L(1,2)$ and $L(3,4)$, and the low-level full-granularity locations $L1, L2, L3, L4$. Two iterations i and ii are needed to attain the desired tree height of 2. Within each iteration, nodes for all possible higher-level generalizations – $L(1,2)$ and $L(3,4)$ in the example – are created from each leaf of the previous iteration. Such nodes are then expanded to lower-level full-granularity nodes only if their Laplacian noisy count is above a threshold (set to 3 here). The model can accommodate more than two levels of generalization, which creates additional layers within each iteration i and ii, one for each generalization level. (d) Differentially private prefix tree upon pruning of all nodes for generalized locations, and with noisy counts made consistent among parent and child nodes. (e) Example of the synthetic trajectories extracted from the final prefix tree structure. Adapted from Chen *et al.* [32].

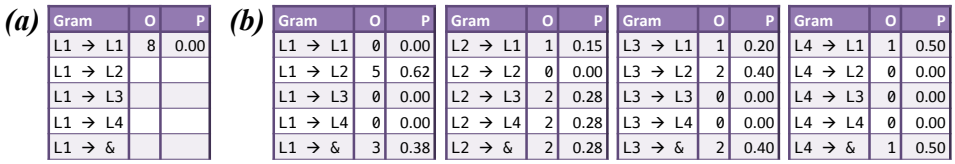

Figure 16: Examples of n-grams used as trajectory representation by Chen *et al.* [31], for the database of trajectory micro-data in Figure 15. Columns report the sequence (Gram), number of occurrences (O), and probability (P) for (a) 1-grams, and (b) 2-grams. Adapted from Chen *et al.* [31].

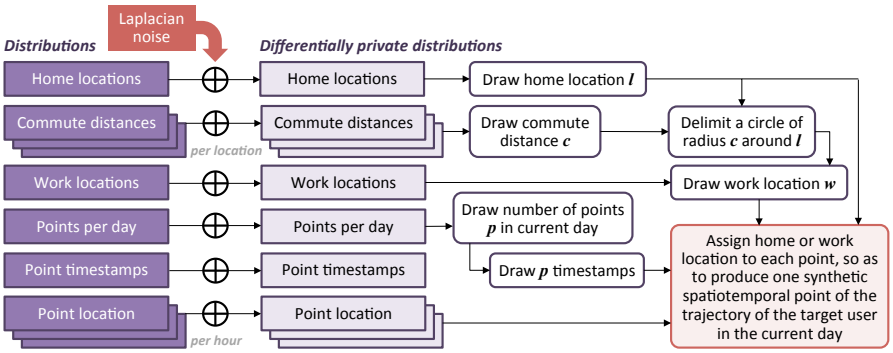

Figure 17: Block diagram of the operation of the DP-WHERE solution by Mir *et al.* [87]. DP-WHERE creates differentially private noisy version of distributions of home and work locations, commute distances (based on the home location), points per day, time at which points are recorded, and locations at which points are recorded (on a hourly basis). Then, values are drawn from such distributions, and combined so as to associate home and work locations to each synthetic user, and generate spatiotemporal points that are then associated with a user. Adapted from Mir *et al.* [87].

Laplace noise to counts, pruning the prefix trees, and extracting the synthetic trajectories is followed. In this last step, the authors also adopt an original sampling technique that allow preserving the correct directionality in the output trajectories. The proposed solution, named Differentially Private Trajectories (DPT), is evaluated with both real and synthetic datasets[45] that are queried for distributions of diameters and trips, and for frequent sequential patterns. Results show that DPT largely outperforms the n-grams-based approach by Chen *et al.* [31] in the considered case studies.

Representing trajectory micro-data as probability distributions. A quite different strategy from those above is proposed by Mir *et al.* [87]. They introduce DP-WHERE, a differentially private synthetic trajectory generator that does not rely on a tree model of the original trajectory micro-data. Instead, DP-WHERE performs the following steps: *(i)* derives a number of distributions that describe different statistical features of the movements in the original trajectory database, such as the spatial distribution of home and work locations, or the number of spatiotemporal points in a trajectory; *(ii)* adds Laplacian noise to such distributions; *(iii)* extracts realizations from the noisy distributions to generate synthetic trajectories. A more detailed view of the considered distributions and their combination is in Figure 17. The synthetic movement data[46] produced by DP-WHERE is proven to preserve population density distributions over time, as well as daily ranges of commutes in the reference area.

Roy *et al.* [103] follow a similar approach in their proposed Sanitization Model. First, they remove outlying records from the original dataset by applying the statistical interquartile range rule to all attributes. Second, they run legacy independence and homogeneity tests on attributes, and group attributes with high associativity in non-disjoint sets. Third, they

[45] Experiments are run on over 4 million GPS trajectories of 8,600 cabs in Beijing, China, during May 2009. Space is discretized into over 138,000 100×100-m^2 cells, leading to an average of 20 points per trajectory. Further tests employ Brinkhoff's network-based generator for moving objects [22] The data consists of 15 million trajectories of 50,000 synthetic individuals in the region of Oldenburg, Germany, with a spatial resolution of 50×50-m^2.

[46] Experiments are carried out on 10,000 synthetic users generated from 1-billion CDR of 250,000 subscribers in the region of New York, NJ, USA, during three months in 2011. The spatial resolution of the synthetic data is of 7 miles, *i.e.*, around 11 km.

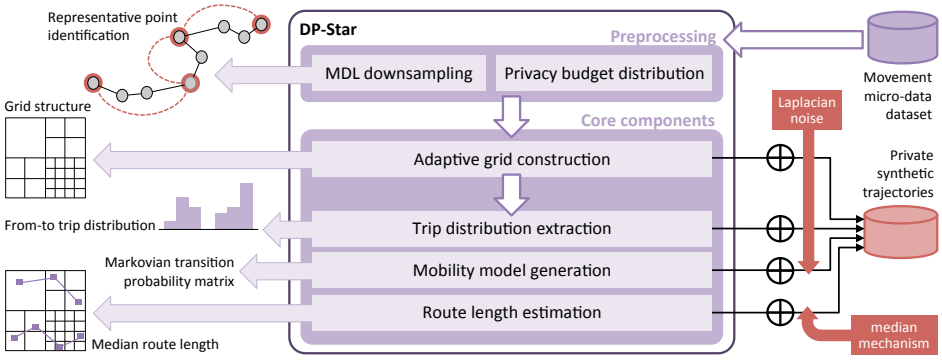

Figure 18: Block diagram of the operation of the DP-Star solution by Gursoy *et al.* [64]. The input trajectory micro-data is downsampled so as to only retain utility-preserving movement information. Then, multiple representations are extracted from the downsampled trajectories. The representations are made differentially private my means of Laplacian and median mechanisms. Finally, privacy-preserving synthetic trajectories are generated from the noisy representations. Adapted from Gursoy *et al.* [64].

derive synthetic distributions for each attribute group, and add Laplacian noise to them based on the available privacy budget. Fourth, synthetic records are generated by drawing samples from these distributions and aggregating them. An interesting aspect of the work is the strategy it adopts to assess the quality of the synthetic trajectory micro-data obtained via the Sanitization Model. The authors consider a database published during a data visualization contest[47], and replicate the competition submissions using both the original and differentially private trajectories. They find that the vast majority of the results are nearly identical, although it should be noted that the reference data is limited to trajectories where only the start and end locations and times are known.

The most recent proposal in probability-distribution-based approaches is DP-Star by Gursoy *et al.* [64], whose operation is summarized in Figure 18. DP-Star first runs a preprocessing phase, during which raw trajectories are downsampled via Minimum Description Length (MDL), and reduced minimum sequences of representative points; also during preprocessing, the privacy budget ϵ is automatically split among the different core components by solving an optimization problem. The, DP-Star generates discretized representations of: *(i)* space, as a non-uniform grid whose cell granularity is adapted to the geographical density of trajectory points; *(ii)* trips between start and end locations, as a probability distribution; *(iii)* internal trip structures, as a Markovian model of transition probability among locations; *(iv)* route lengths, as the median distance covered by trajectories starting at each location. Representations in *(i)–(iii)* are perturbed with Laplacian noise, while the noisy median route lengths are obtained with the *median mechanism* proposed by Cormode *et al.* [36].

Differentially private synthetic trajectory micro-data is then extracted by combining the representations above, by selecting start and end grid cells, determining a route length based on the start cell, defining a sequence of cells via the Markovian model, and finally converting cells to actual points. We remark that DP-Star generates spatial trajectories that

[47] The study leverages data provided by the Hubway bike sharing initiation and the Metropolitan Area Planning Council (MAPC) of Boston within their Hubway Data Visualization Challenge. The database consists of historical data about over one million bike trips in the grater Boston area. For each trip, the data include the start and end spatiotemporal points, as well as non-positioning information about the gender and subscription type of the rider.

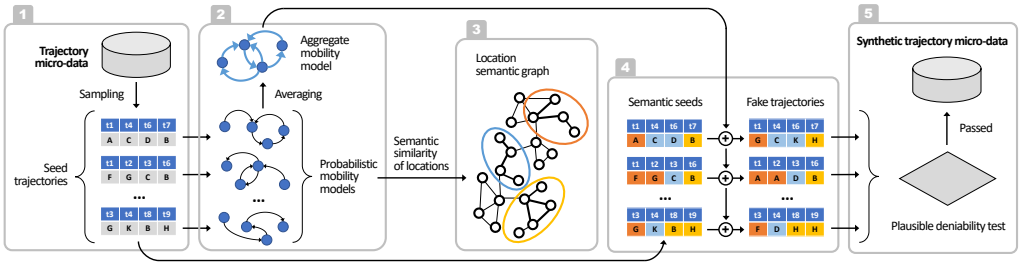

Figure 19: Descriptive overview of the technique adopted to implement plausible deniability in synthetic trajectory micro-data. Steps 1 to 4 correspond to the transformation phase *(i)*, and step 5 maps to the verification phase *(ii)*. Adapted from Bindschaedler and Shokri [16].

do not include temporal information, unlike, *e.g.*, DP-WHERE. Evaluations with substantial real-world data[48] shows that DP-Star retains significantly higher utility than the n-gram-based solution by Chen *et al.* [31] and DPT by He *et al.* [66], under several types of queries on trajectory micro-data.

3.4.3 Plausible deniability

Bindschaedler and Shokri [16] propose an alternative definition of privacy for synthetic trajectories that also aims at realizing the uninformativeness principle. Their criterion is based on *plausible deniability*: only a subset of *seed* original trajectories is leveraged to generate the synthetic output trajectory micro-data, and the inclusion of a particular real trajectory among the seeds must be plausibly deniable. Such a condition is achieved by requiring that each synthetic trajectory in the released database could have been generated by a sufficiently large number of users in the original database, including those that were not selected as seeds. Formally:

Definition 6. A synthetic trajectory f generated from a seed trajectory $s \in \mathcal{S} \subset \mathcal{D}$ satisfies (k, δ)-plausible deniability if there are at least $k \geq 1$ alternative trajectories $a \in \mathcal{D}$ such that the similarity σ of f and s is not much higher than the same similarity measured between f and any a, or $|\sigma(s, f) - \sigma(a, f)| \leq \delta$.

In the definition above, k denotes the number of trajectories in the original database that is large enough to ensure that the deniability of the presence of s in the released data is actually plausible. Such original trajectories must yield a similarity with the synthetic trajectory, measured by a metric σ, within a threshold δ from that between the synthetic and seed trajectories.

In order to ensure plausibly deniability of synthetic trajectory micro-data, the authors adopt a strategy of: *(i)* transforming each seed input trajectory into a semantic space, and then probabilistically transform it back to the original geographic space; *(ii)* verifying if the proposed re-transformed trace satisfies plausible deniability when confronted to the original database, and only adding it to the output database if the answer is positive. Figure 19 offers an overview of the proposed scheme.

[48] Three different datasets are considered: 14,650 GPS trajectories from the GeoLife project [137], with an average of over 900 points each; 30,000 taxi traces collected in Porto, Portugal, with 43 points each on average; 50,000 synthetic trajectories created using Brinkhoff's generator [22].

Phase *(i)* allows generating a synthetic copy of an original trajectory so that spatiotemporal features are preserved, and its main challenge is defining the transformation. To this end, the authors proceed as follows. First, each seed input trajectory is summarized into a mobility model that captures statistical information on the visiting probability to every location and the transition probabilities among such locations. A pairwise semantic similarity of mobility models is then computed as their maximum geographical closeness under all possible mappings of visited locations: the rationale is that two similar trajectories follow equivalent spatiotemporal patterns (*e.g.*, the same home $\rightarrow$ work $\rightarrow$ other $\rightarrow$ home repeated sequences) at different locations (*i.e.*, the exact home, work and other locations are distinct for the two users, and possibly far apart), hence a suitable mapping of locations (*e.g.*, considering the two home locations to be the same, and similarly for work and other) can reveal their resemblance. Finally, all similarity information is aggregated in a location semantic graph: the nodes are the locations, and edge weights are the average semantic similarity between two locations over all pairs of mobility models. Intuitively, high weights characterize pairs of locations that are visited in similar ways by many users in the original database. The scheme thus performs a clustering on the graph, so that semantically similar (but geographically distinct) locations are grouped together in a same class.

The transformation in *(i)* above consists in replacing locations in the original seed trajectory with their classes, which thus represent the semantic space. The re-transformation occurs according to an aggregate mobility model (*i.e.*, an average of all individual mobility models), which is run across the semantic space under the constraint that its locations are a subset of the locations of the target semantic trace. The constraint ensures that each synthetic trajectory shares the same semantic trace with its original version.

The probabilistic nature of the aggregate mobility model allows generating multiple synthetic trajectories for a same original seed user, which is leveraged in phase *(ii)* of the solution. Once one synthetic trajectory is generated, it undergoes a privacy test based on plausible deniability: specifically, it is verified that it does not leak more information about the real mobility of the user than it does for other original trajectories in the input database. If the test fails, then a different synthetic trajectory is generated for the user, and the process is iterated. In a follow-up work, Bindschaedler *et al.* [17] show that a randomized form of the solution above achieves in fact (ϵ, δ)-differential privacy, under a rigid set of ϵ and δ values.

Experiments with measurement data[49] show that the synthetic trajectory micro-data retains, under $k = 1$, significant utility in terms of visit frequency distributions, top-n location coverage, user time allocation, spatiotemporal and semantic mobility features.

3.4.4 $k^{\tau,\epsilon}$-anonymity

Gramaglia *et al.* [62] introduce $k^{\tau,\epsilon}$-anonymity, a privacy criterion that stems from the original k-anonymity but has fundamentally different semantics. The idea behind $k^{\tau,\epsilon}$-anonymity is ensuring that any subset of spatiotemporal points known to the adversary matches k trajectories, and that such k trajectories are sufficiently diverse from each other in the rest of their points. Formally:

[49] The authors generate synthetic trajectories from the interpolated GPS data of 30 seed users participating in the Nokia Lausanne Data Collection Campaign [73]. The data is preprocessed so that all trajectories have a fixed sampling interval of 20 minutes, and a duration of one day; also, rarely visited locations are clustered together, so that the total number of locations is reduced by 60%. A different day of mobility of the same 30 users is leveraged as the alternative database during the test phase that ensures plausible deniability.

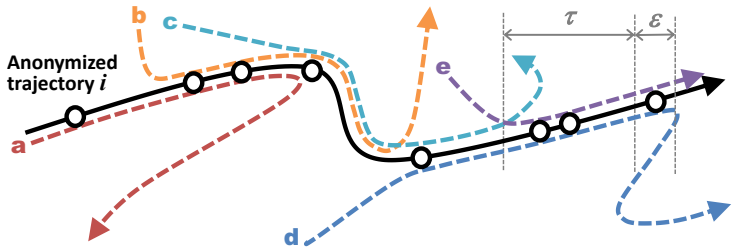

Figure 20: Illustrative example of $k^{\tau,\epsilon}$-anonymity of user i, for $k = 2$, using the movements of five other subscribers a, b, c, d, e. An adversary knowing a sub-trajectory of i during any time interval of duration τ always finds at least one (*i.e.*, $k - 1$) other user with a movement pattern that is identical to that of i during that interval, but different elsewhere. With this knowledge, the adversary cannot tell apart i from the other subscriber, and thus cannot attribute full trajectories to one user or the other. As this holds no matter where the knowledge interval is shifted to, the attacker can never retrieve the complete movement patterns of i: this achieves the uninformativeness principle. Still, the adversary can increase its knowledge in some cases. Let us consider the interval τ indicated in the figure: the trajectories of i, d and e are identical for some time after τ, which allows associating to i the movements during ϵ: the attacker learns one additional spatiotemporal point of i. Adapted from Gramaglia *et al.* [62].

Definition 7. Let $\mathcal{D}$ be a database of trajectory micro-data and $LBQID$ the location-based quasi-identifier associated with it comprising τ spatiotemporal points. Also, let $\mathcal{D}[LBQID]$ be the set of records returned by a query for $LBQID$ on $\mathcal{D}$. Then, $\mathcal{D}$ is said to satisfy $k^{\tau,\epsilon}$-anonymity if and only if the records in $\mathcal{D}[LBQID]$ are at least k, and only share an additional set of ϵ points other than the τ in the $LBQID$.

According to the definition above, given a subset of any τ points in a target trajectory, the subset is k-anonymous, *i.e.*, is indistinguishable from points in $k - 1$ other trajectories. However, apart from the τ points, the target trajectory only shares a small number of additional ϵ points with the $k-1$ trajectories used to anonymize the subset under consideration. The correct implementation of $k^{\tau,\epsilon}$-anonymity grants that the attacker cannot distinguish among k separated alternative trajectories that stem from his known points, and cannot infer substantial additional positioning information about his target user by accessing the database. Hence, $k^{\tau,\epsilon}$-anonymity realizes the uninformativeness principle. As an interesting remark, when τ maps to the whole target trajectory, $k^{\tau,\epsilon}$-anonymity reduces to the original k-anonymity; therefore, solutions providing the former can be naively reduced to offer the latter privacy criterion.

The solution proposed by Gramaglia *et al.* [62], named `kte-hide`, satisfies the privacy criterion under the assumption that the points of the LBQID are adjacent in time; in this case, τ and ϵ can be mapped to time intervals rather than numbers of points. The algorithm then alternates different and partially overlapping anonymity sets over subsequent intervals of duration $\tau+\epsilon$, realizing the structure illustrated in the toy example of Figure 20. The performance evaluation of `kte-hide` is based on real-world datasets[50], and shows that the method is capable of attaining $2^{\tau,\epsilon}$-anonymity under seamless adversary knowledge rang-

[50]The authors use nationwide and citywide CDR-based trajectories released in the context of the D4D Challenge [18], as well as CDR data collected by the University of Minnesota [135]. The resulting datasets follow 30,000–300,000 users for 1 day to 2 weeks in Abidjan, Dakar, Ivory Coast, Senegal and Shenzhen.

ing between 10 minutes and 4 hours, and $\epsilon = \tau$. The solution retains a median accuracy of the anonymized trajectory data of 1-2 km in space and less than 1 hour in time.

4 Discussion and perspectives

Based on the comprehensive survey of attacks against released databases of trajectory micro-data, and of countermeasures against such threats, our main takeaway message is that PPDP of trajectories is still a largely open problem. There is substantial space for improvement at all levels, and we outline below some directions for future research.

4.1 Realistic and credible risk assessments

It is important that the privacy risks associated with the publication of trajectory micro-data are assessed in practical settings. The vast majority of the works in the literature highlight very high re-identification (*i.e.*, successful record linkage) rates, announcing dramatic hazard for the privacy of the monitored individuals. However, these results have to be interpreted with a grain of salt. Many assume that the adversary knows some spatiotemporal points of its target user, which happen to be exactly in the target database (*i.e.*, a spatiotemporal subset format of the side information, according to our classification): this is very unlikely to happen in real life, as the adversary would have to anticipate when the user's location will be sampled by the positioning system. Similarly, it is simplistic to assume that the adversary is aware of its target's locations sampled with similar temporal frequency and spatial accuracy than those in the target dataset; or, equivalently, it is naive to expect that an attacker can build mobility profiles that are as detailed as those it can infer from the target trajectory micro-data. Indeed, there is a legitimate question on whether an attacker having such a substantial knowledge would be actually interested in making a large effort to retrieve "more of the same" data.

Note that we are not downplaying the privacy issues in trajectory micro-data– which we believe are many and extremely relevant. However, we advocate for more realistic risk assessments that are representative of the actual conditions an attacker could operate in. Practical attacks require identifying and retrieving useful side information, and performing a reliable match with the target data; moreover, in most cases the attacker has to deal with uncertainty about the presence of its target user in the target database, as well as about an eventual match (since it does not possess any ground truth information guaranteeing that the match is correct). In absence of these practical considerations, studies may lead to over-pessimistic claims on privacy risks, which are instead mitigated when attacks are run in the wild.

The recent work by Wang *et al.* [128], who show that figures on attack success rates in the literature are largely exaggerated when considering closer-to-reality settings, is a first evidence in this sense. However, it is not a definitive one, as the authors still retain many assumptions that simplify the attacker's work. More realistic and credible assessments of the actual risks associated with record linkage of trajectory micro-data are required.

4.2 Risks beyond record linkage

Record linkage absorbs almost the whole literature on attacks against trajectory micro-data, as it is well illustrated in Table 1. However, these are, at least in theory, the simplest forms

of menace against trajectory databases. Therefore, our considerations above on risk assessment are exacerbated in the case of attacks that are more complex than record linkage.

The privacy risks of, *e.g.*, attribute linkage (just to consider the next level of threat) are basically unexplored. Homogeneity, *i.e.*, the weakness that paves the way for attribute linkage, is a clearly understood concept, for which toy examples are easily constructed, and for which practical cases have been demonstrated in the context of relational databases. However, whether homogeneity actually exists, and, if so, to which level, remains a fully open question when it comes to the sensitive attributes one could link to trajectory micro-data. To date, we can only imagine that the risk may exist, but we do not even have a rough picture of its practical viability. The situation is similar for probabilistic attacks. Therefore, and even more than in the case of record linkage, realistic risk assessments of attribute linkage or probabilistic attacks represent an opportunity for future investigation.

4.3 Silver bullet anonymization

Anonymizing trajectory micro-data is extremely complex, and this is apparent from the number of solutions proposed over the past few years. We have understood that mitigation techniques simply do not work: reducing the spatial or temporal resolution of the data does not help, and also shortened or intertwined trajectories retain re-identifiability risks. Unfortunately, also more complex approaches are far from perfect.

On the one hand, techniques that grant k-anonymity are today fairly mature, preserve individual trajectories, and can retain a decent level of precision in the anonymized data (see Table 3). However, they typically scale poorly with k. More importantly, they only offer a protection against record linkage, and leave the data prone to more complex attacks, disregarding for the moment the question if these are actually feasible or not (see above).

On the other hand, differential privacy and its extensions for location data are very hard to apply to trajectory micro-data. As of today, all solutions implementing such privacy principle construct some model from the original data, apply noise so as to make the model differentially private, and then generate synthetic trajectories from the noisy model. It is clear that the anonymized dataset only retain global properties, and prevents analyses that require following actual individuals. Moreover, the global properties that can be explored through data mining are the same that are preserved by the noisy model: *i.e.*, there is no guarantee that features that are lost during the modelling phase will be reflected in the output database. Again, this poses potential limits to the nature of queries one can safely run on the anonymized data. Finally, most solutions for differential privacy also do not scale well, and are only demonstrated with simplistic databases of trajectories that are either very short, only defined over space, or spanning a small set of total locations.

As a result, the quest for a silver bullet anonymization solution for trajectory micro-data is still open, and it may pass through new privacy principles that go beyond k-anonymity or differentialy privacy.

4.4 Reproducible research and comparative evaluations

A striking aspect of most works in the literature on trajectory micro-data anonymization is that they provide very little in terms of comparison with previous solutions. This is clearly an issue that hinders our capability of untangling the body of literature and name a clear winner in the contest for the current state-of-the-art. We identify three main reasons for such an undesirable situation.

First, there is a lack of reference dataset of trajectory micro-data. Synthetic datasets (*e.g.*, those presented in [132, 13, 124] struggle to rise to such a status, due to their artificial nature. Publicly available datasets collected in the real world (*e.g.*, those in the CRAWDAD repository) are fairly old and limited in size. Some larger real-life databases have been released, *e.g.*, as part of challenges by mobile network operators [18], yet they are protected by non-disclosure agreements that prevent their open distribution. Many works thus rely on proprietary data that is not made accessible to the research community, again due to agreements with the data providers, which are typically companies, for legal reasons related to the publication of real mobility traces. Such a scenario makes it hard to develop a reference set of trajectory databases like in other communities, hence limits the possibility of verifying the performance of different solutions on the same ground. Overall, we argue that there is a significant need for some large academic initiative to collect and release such open trajectory micro-data.

Second, the approaches adopted to evaluate different anonymization techniques vary wildly across studies. Works in the literature use a plethora of different quality measures, error metrics, queries and data mining analyses, which are however very diverse. Researchers have a tendency to always design new metrics (possibly well suited to their proposed solution), making it impossible to confront the performance evaluations carried out in two different papers. Also in this case, we need a reference set of metrics or tasks for quality assessment of the anonymized data, to be adopted throughout all studies and allowing a direct comparison of the performance figures. Clearly, such a set shall be large enough to cover a vast range of data usages, and avoid favouring a solution over another.

Third, very few researchers release the source code of their solutions. This is a despicable but common practice that curbs not only the reproducibility and comparability, but also the mere verifiability of the results. We argue that, as a community, we should move to a fully verifiable model where all papers proposing anonymization techniques shall be accompanied by their source code, possibly written in a commonly agreed language.

Overcoming the three problems above would make comparison straightforward and unavoidable, and improve the scientific rigour of the process towards solving the problem of anonymization of trajectory micro-data.

Acknowledgements

The authors would like to thank Emilie Sirvent-Hien and Marc-Olivier Killijian for the constructive discussions. This research was supported by BPIFrance through the Programme d'Investissement d'Avenir, project n.P128356-2659748 (ADAGE).

References

[1] Mobile location-based services – 9th edition. Technical report, Berg Insight AB, 2015.

[2] *Next generation simulation (ngsim)*, 2017. Accessed: 2018-12-28.

[3] *Orange Flux Vision*, 2017. Accessed: 2018-12-28.

[4] Osman Abul, Francesco Bonchi, and Mirco Nanni. Never walk alone: Uncertainty for anonymity in moving objects databases. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 376–385. Ieee, 2008.

[5] Osman Abul, Francesco Bonchi, and Mirco Nanni. Anonymization of moving objects databases by clustering and perturbation. *Information Systems*, 35(8):884–910, 2010.

[6] Gergely Acs and Claude Castelluccia. A case study: privacy preserving release of spatio-temporal density in paris. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1679–1688. ACM, 2014.

[7] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. *SIGMOD Rec.*, 29(2):439–450, May 2000.

[8] Dalal Al-Azizy, David Millard, Iraklis Symeonidis, Kieron O'Hara, and Nigel Shadbolt. A literature survey and classifications on data deanonymisation. In Costas Lambrinoudakis and Alban Gabillon, editors, *Risks and Security of Internet and Systems: 10th International Conference, CRiSIS 2015, Mytilene, Lesbos Island, Greece, July 20-22, 2015, Revised Selected Papers*, pages 36–51. Springer International Publishing, 2016.

[9] Mohammad Alaggan, Sébastien Gambs, Stan Matwin, and Mohammed Tuhin. Sanitization of call detail records via differentially-private bloom filters. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 223–230. Springer, 2015.

[10] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 901–914. ACM, 2013.

[11] Roland Assam, Marwan Hassani, and Thomas Seidl. Differential private trajectory protection of moving objects. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on GeoStreaming*, IWGS '12, pages 68–77, New York, NY, USA, 2012. ACM.

[12] Claudio Bettini. Privacy protection in location-based services: a survey. In *Handbook of Mobile Data Privacy*, pages 73–96. Springer, 2018.

[13] Claudio Bettini, Sergio Mascetti, X Sean Wang, Dario Freni, and Sushil Jajodia. Anonymity and historical-anonymity in location-based services. In *Privacy in location-based applications*, pages 1–30. Springer, Berlin Heidelberg, 2009.

[14] Claudio Bettini, X. Wang, and Sushil Jajodia. Protecting privacy against Location-Based personal identification. In Willem Jonker and Milan Petkovic, editors, *Secure Data Management*, volume 3674 of *Lecture Notes in Computer Science*, chapter 13, pages 185–199. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2005.

[15] Greg Bigwood, Tristan Henderson, Devan Rehunathan, Martin Bateman, and Saleem Bhatti. CRAWDAD dataset st_andrews/sassy (v. 2011-06-03), June 2011.

[16] V. Bindschaedler and R. Shokri. Synthesizing plausible privacy-preserving location traces. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 546–563, May 2016.

[17] Vincent Bindschaedler, Reza Shokri, and Carl A. Gunter. Plausible deniability for privacy-preserving data synthesis. *Proc. VLDB Endow.*, 10(5):481–492, January 2017.

[18] Vincent D. Blondel, Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. Data for development: the D4D challenge on mobile phone data. *CoRR*, abs/1210.0137, 2012.

[19] Francesco Bonchi, Laks VS Lakshmanan, and Hui Wendy Wang. Trajectory anonymity in publishing personal mobility data. *ACM Sigkdd Explorations Newsletter*, 13(1):30–42, 2011.

[20] Luca Bonomi and Li Xiong. A two-phase algorithm for mining sequential patterns with differential privacy. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 269–278, New York, NY, USA, 2013. ACM.

[21] Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Optimal geo-indistinguishable mechanisms for location privacy. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 251–262, New York, NY, USA, 2014. ACM.

[22] Thomas Brinkhoff. Generating traffic data. *IEEE Bulletin of the Technical Committee on Data Engineering*, 26, 2003.

[23] Solenn Brunet, Sébastien Canard, Sébastien Gambs, and Baptiste Olivier. Novel differentially private mechanisms for graphs. *IACR Cryptology ePrint Archive*, 2016:745, 2016.

[24] Yang Cao and Masatoshi Yoshikawa. Differentially private real-time data release over infinite trajectory streams. In *2015 16th IEEE International Conference on Mobile Data Management*, volume 2, pages 68–73. IEEE, 2015.

[25] Alket Cecaj, Marco Mamei, and Nicola Bicocchi. Re-identification of anonymized CDR datasets using social network data. In *2014 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 237–242, March 2014.

[26] Alket Cecaj, Marco Mamei, and Franco Zambonelli. Re-identification and information fusion between anonymized CDR and social network data. *Journal of Ambient Intelligence and Humanized Computing*, 7(1):83–96, February 2016.

[27] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Marco Stronati. A predictive differentially-private mechanism for mobility traces. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 21–41. Springer, 2014.

[28] Brian X. Chen. iPhone Tracks Your Every Move, and There's a Map for That. *WIRED*, 2011.

[29] Brian X. Chen. Why and How Apple Is Collecting Your iPhone Location Data. *WIRED*, 2011.

[30] Lei Chen, M. Tamer Özsu, and Vincent Oria. Robust and fast similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, pages 491–502, New York, NY, USA, 2005. ACM.

[31] Rui Chen, Gergely Acs, and Claude Castelluccia. Differentially private sequential data publication via variable-length n-grams. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, CCS '12, pages 638–649, New York, NY, USA, 2012. ACM.

[32] Rui Chen, Benjamin C.M. Fung, Bipin C. Desai, and Nériah M. Sossou. Differentially private transit data publication: A case study on the montreal transportation system. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 213–221, New York, NY, USA, 2012. ACM.

[33] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1082–1090, New York, NY, USA, 2011. ACM.

[34] Chi-Yin Chow and Mohamed F. Mokbel. Trajectory privacy in location-based services and data publication. *SIGKDD Explor. Newsl.*, 13(1):19–29, August 2011.

[35] Delphine Christin. Privacy in mobile participatory sensing: Current trends and future challenges. *Journal of Systems and Software*, 116(Supplement C):57 – 68, 2016.

[36] Graham Cormode, Cecilia Procopiuc, Divesh Srivastava, Entong Shen, and Ting Yu. Differentially private spatial decompositions. In *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering*, ICDE '12, pages 20–31, Washington, DC, USA, 2012. IEEE Computer Society.

[37] Enrique Dans. Telefonica and the fourth platform. *Medium*, 2017.

[38] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3, March 2013.

[39] Yves-Alexandre de Montjoye, Laura Radaelli, Vivek K. Singh, and Alex Pentland. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539, January 2015.

[40] Yoni De Mulder, George Danezis, Lejla Batina, and Bart Preneel. Identification via location-profiling in GSM networks. In *Proceedings of the 7th ACM workshop on Privacy in the electronic society*, WPES '08, pages 23–32, New York, NY, USA, 2008. ACM.

[41] Michael Decker. Location privacy-an overview. In *2008 7th International Conference on Mobile*

Business, pages 221–230, July 2008.

[42] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017.

[43] Josep Domingo-Ferrer and Rolando Trujillo-Rasua. Microaggregation-and permutation-based anonymization of movement data. *Information Sciences*, 208:55–80, 2012.

[44] Anthony Dornan and Philip Laidler. Fast-tracking operator plans to win in the $5-billion location insight market. *Telco 2.0 Research Paper*, 2015.

[45] Matt Duckham and Lars Kulik. A formal model of obfuscation and negotiation for location privacy. In *Proceedings of the Third International Conference on Pervasive Computing*, PERVASIVE'05, pages 152–170, Berlin, Heidelberg, 2005. Springer-Verlag.

[46] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag.

[47] Nathan Eagle and Alex (Sandy) Pentland. Reality mining: Sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, March 2006.

[48] Thomas Eiter and Heikki Mannila. Computing discrete fréchet distance. Technical report, Technische Universität Wien, 1994.

[49] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), May 2016.

[50] Elias Frentzos, Kostas Gratsias, Nikos Pelekis, and Yannis Theodoridis. Nearest neighbor search on moving object trajectories. In *Proceedings of the 9th International Conference on Advances in Spatial and Temporal Databases*, SSTD'05, pages 328–345, Berlin, Heidelberg, 2005. Springer-Verlag.

[51] Julien Freudiger, Reza Shokri, and Jean-Pierre Hubaux. Evaluating the Privacy Risk of Location-based Services. In *Proceedings of the 15th International Conference on Financial Cryptography and Data Security*, FC'11, pages 31–46, Berlin, Heidelberg, 2012. Springer-Verlag.

[52] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42(4):14:1–14:53, June 2010.

[53] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. De-anonymization attack on geolocated data. *Journal of Computer and System Sciences*, 80(8):1597–1614, April 2014.

[54] Huiji Gao, Jiliang Tang, and Huan Liu. gscorr: Modeling geo-social correlations for new check-ins on location-based social networks. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1582–1586, New York, NY, USA, 2012. ACM.

[55] Simson L. Garfinkel. De-identification of personal information. Technical Report NISTIR8053, National Institute of Standards and Technology, October 2015.

[56] Fosca Giannotti, Dino Pedreschi, and Franco Turini. Mobility, data mining and privacy the experience of the geopkdd project. In Francesco Bonchi, Elena Ferrari, Wei Jiang, and Bradley Malin, editors, *Privacy, Security, and Trust in KDD*, pages 25–32. Springer-Verlag, Berlin, Heidelberg, 2009.

[57] Aris Gkoulalas-Divanis and Claudio Bettini. *Handbook of Mobile Data Privacy*. Springer International Publishing, Cham, 2018.

[58] Oana Goga, Howard Lei, Sree Hari Krishnan Parthasarathi, Gerald Friedland, Robin Sommer, and Renata Teixeira. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 447–458, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering

Committee.

[59] Philippe Golle and Kurt Partridge. On the anonymity of home/work location pairs. In *International Conference on Pervasive Computing*, pages 390–397. Springer, 2009.

[60] Marta C. González, César A. Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[61] Marco Gramaglia and Marco Fiore. Hiding mobile traffic fingerprints with glove. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*, page 26. ACM, 2015.

[62] Marco Gramaglia, Marco Fiore, Alberto Tarable, and Albert Banchs. Preserving mobile subscriber privacy in open datasets of spatiotemporal trajectories. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pages 1–9, May 2017.

[63] Marco Gruteser and Dirk Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the 1st International Conference on Mobile Systems, Applications and Services*, MobiSys '03, pages 31–42, New York, NY, USA, 2003. ACM.

[64] Mehmet E. Gursoy, Ling Liu, Stacey Truex, and Lei Yu. Differentially private and utility preserving publication of trajectory data. *IEEE Transactions on Mobile Computing*, pages 1–1, 2018.

[65] Michael Hay, Ashwin Machanavajjhala, Gerome Miklau, Yan Chen, and Dan Zhang. Principled evaluation of differentially private algorithms using dpbench. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, pages 139–154, New York, NY, USA, 2016. ACM.

[66] Xi He, Graham Cormode, Ashwin Machanavajjhala, Cecilia M. Procopiuc, and Divesh Srivastava. Dpt: Differentially private trajectory synthesis using hierarchical reference systems. *Proceedings VLDB Endow.*, 8(11):1154–1165, July 2015.

[67] Eelco Herder, Ricardo Kawase, and Georgios Papadakis. Experiences in building the public web history repository. In *Proc. of Datatel Workshop at the Alpine Rendez-Vous*, 2011.

[68] Baik Hoh, Marco Gruteser, Hui Xiong, and Ansaf Alrabady. Enhancing security and privacy in traffic-monitoring systems. *IEEE Pervasive Computing*, 5(4):38–46, 2006.

[69] Shouling Ji, Weiqing Li, Mudhakar Srivatsa, Jing Selena He, and Raheem Beyah. Structure based data de-anonymization of social networks and mobility traces. In *International Conference on Information Security*, pages 237–254. Springer, 2014.

[70] Shouling Ji, Weiqing Li, Mudhakar Srivatsa, Jing Selena He, and Raheem Beyah. General graph data de-anonymization: From mobility traces to social networks. *ACM Transactions on Information and System Security (TISSEC)*, 18(4):12, 2016.

[71] Georgios Kellaris, Stavros Papadopoulos, Xiaokui Xiao, and Dimitris Papadias. Differentially private event sequences over infinite streams. *Proceedings of the VLDB Endowment*, 7(12):1155–1166, 2014.

[72] Marc-Olivier Killijian, Matthieu Roy, and Gilles Trédan. Beyond San Francisco Cabs: building a*-lity mining dataset for social traces analysis. In *Workshop on the Analysis of Mobile Phone Networks*, page 6p, Cambridge, MA, USA, 2010.

[73] N. Kiukkonen, Blom J., O. Dousse, Daniel Gatica-Perez, and J. K. Laurila. Towards rich mobile phone datasets: Lausanne data collection campaign. In *Proc. ACM Int. Conf. on Pervasive Services (ICPS,',',') , Berlin.*, 7 2010.

[74] Dániel Kondor, Behrooz Hashemian, Yves-Alexandre de Montjoye, and Carlo Ratti. Towards matching user mobility traces in large-scale datasets. *ArXiv e-prints, 1709.05772*, September 2017.

[75] Despina Kopanaki, Vasilis Theodossopoulos, Nikos Pelekis, Ioannis Kopanakis, and Yannis Theodoridis. Who cares about others' privacy: Personalized anonymization of moving object trajectories. In *EDBT*, pages 425–436, 2016.

[76] John Krumm. Inference attacks on location tracks. In *International Conference on Pervasive Computing*, pages 127–143. Springer, 2007.

[77] Siddique Latif, Muhammad Usman, Sanaullah Manzoor, Waleed Iqbal, Junaid Qadir, Gareth Tyson, Ignacio Castro, Adeel Razi, Maged N Kamel Boulos, Adrian Weller, et al. Leveraging data science to combat covid-19: A comprehensive review. 2020.

[78] Chao Li, Michael Hay, Gerome Miklau, and Yue Wang. A data- and workload-aware algorithm for range queries under differential privacy. *Proceedings VLDB Endow.*, 7(5):341–352, January 2014.

[79] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *IEEE 23rd International Conference on Data Engineering, ICDE 2007.*, pages 106–115. IEEE, 2007.

[80] Changchang Liu, Supriyo Chakraborty, and Prateek Mittal. Dependence makes you vulnerable: Differential privacy under dependent tuples. In *Proc. Network and Distributed System Security Symposium (NDSS'16)*, 2016.

[81] Natasha Lomas. How "anonymous" wifi data can still be a privacy risk. *TechCrunch*, October 2017.

[82] Chris Y. T. Ma, David K. Y. Yau, Nung K. Yip, and Nageswara S. V. Rao. Privacy vulnerability of published anonymous mobility traces. *IEEE/ACM Transactions on Networking*, 21(3):720–733, June 2013.

[83] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering. ICDE'06.*, pages 24–24. IEEE, 2006.

[84] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), March 2007.

[85] Jonathan Mayer, Patrick Mutchler, and John C. Mitchell. Evaluating the privacy properties of telephone metadata. *Proceedings of the National Academy of Sciences*, 113(20):5536–5541, May 2016.

[86] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '07, pages 94–103, Washington, DC, USA, 2007. IEEE Computer Society.

[87] Darakhshan J. Mir, Sibren Isaacman, Ramón Cáceres, Margaret Martonosi, and Rebecca N. Wright. Dp-where: Differentially private modeling of human mobility. In *IEEE International Conference on Big Data, 2013*, pages 580–588. IEEE, 2013.

[88] Anna Monreale, Gennady L Andrienko, Natalia V Andrienko, Fosca Giannotti, Dino Pedreschi, Salvatore Rinzivillo, and Stefan Wrobel. Movement data anonymity through generalization. *Trans. Data Privacy*, 3(2):91–121, 2010.

[89] Muhammad Haris Mughees, Hamed Haddadi, and Pan Hui. Privacy leakage in mobile computing: Tools, methods, and characteristics. *CoRR*, abs/1410.4978, 2014.

[90] Takao Murakami, Atsunori Kanemura, and Hideitsu Hino. Group sparsity tensor factorization for re-identification of open mobility traces. *IEEE Transactions on Information Forensics and Security*, 12(3):689–704, March 2017.

[91] Mirco Musolesi, Kristof Fodor, Mattia Piraccini, Antonio Corradi, and Andrew Campbell. CRAWDAD dataset dartmouth/cenceme (v. 2008-08-13), August 2008.

[92] Diala Naboulsi, Marco Fiore, Stephane Ribot, and Razvan Stanica. Large-scale mobile traffic analysis: A survey. *IEEE Communications Surveys Tutorials*, 18(1):124–161, Firstquarter 2016.

[93] Farid M. Naini, Jayakrishnan Unnikrishnan, Patrick Thiran, and Martin Vetterli. Where You Are Is Who You Are: User Identification by Matching Statistics. *IEEE Transactions on Informa-*

tion Forensics and Security, 11(2):358–372, February 2016.

[94] Mehmet Ercan Nergiz, Maurizio Atzori, Yücel Saygin, and Baris Güç. Towards trajectory anonymization: A generalization-based approach. *Trans. Data Privacy*, 2(1):47–75, April 2009.

[95] Philip Neustrom. Want to see something crazy? open this link on your phone with wifi turned off. *Medium*, 2017.

[96] Hoa Ngo and Jong Kim. Location privacy via differential private perturbation of cloaking area. In *2015 IEEE 28th Computer Security Foundations Symposium*, pages 63–74, July 2015.

[97] Michal Piorkowski, Natasa Sarafijanovic-Djukic, and Matthias Grossglauser. CRAWDAD dataset epfl/mobility (v. 2009-02-24), February 2009.

[98] Wahbeh Qardaji, Weining Yang, and Ninghui Li. Differentially private grids for geospatial data. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pages 757–768, April 2013.

[99] Daniele Riboni, Linda Pareschi, Claudio Bettini, and Sushil Jajodia. Preserving anonymity of recurrent location-based queries. In *2009 16th International Symposium on Temporal Representation and Reasoning*, pages 62–69, July 2009.

[100] Christopher Riederer, Yunsung Kim, Augustin Chaintreau, Nitish Korula, and Silvio Lattanzi. Linking Users Across Domains with Location Data: Theory and Validation. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 707–719, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[101] Luca Rossi and Mirco Musolesi. It's the way you check-in: Identifying users in location-based social networks. In *Proceedings of the Second ACM Conference on Online Social Networks*, COSN '14, pages 215–226, New York, NY, USA, 2014. ACM.

[102] Luca Rossi, James Walker, and Mirco Musolesi. Spatio-temporal techniques for user identification by means of GPS mobility data. *EPJ Data Science*, 4(1), August 2015.

[103] Harichandan Roy, Murat Kantarcioglu, and Latanya Sweeney. Practical differentially private modeling of human movement data. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 170–178. Springer, 2016.

[104] Julián Salas, David Megías, and Vicenç Torra. Swapmob: Swapping trajectories for mobility anonymization. In Josep Domingo-Ferrer and Francisco Montes, editors, *Privacy in Statistical Databases*, pages 331–346, Cham, 2018. Springer International Publishing.

[105] Piotr Sapiezynski, Arkadiusz Stopczynski, Radu Gatej, and Sune Lehmann. Tracking human mobility using wifi signals. *PLOS ONE*, 10(7):1–11, 07 2015.

[106] Julie Segal. The next big source of alpha: Your phone? *Institutional Investor LLC*, 2017.

[107] Dongxu Shao, Kaifeng Jiang, Thomas Kister, Stéphane Bressan, and Kian-Lee Tan. Publishing trajectory with differential privacy: A priori vs. a posteriori sampling mechanisms. In *Proceedings of the 24th International Conference on Database and Expert Systems Applications - Volume 8055*, DEXA 2013, pages 357–365, Berlin, Heidelberg, 2013. Springer-Verlag.

[108] Sharad, Kumar and Danezis, George. De-anonymizing d4d datasets. In *6th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2013)*, Bloomington, Indiana, USA, 2013.

[109] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. Quantifying location privacy. In *IEEE Symposium on Security and Privacy (SP), 2011*, pages 247–262. IEEE, May 2011.

[110] Reza Shokri, George Theodorakopoulos, Carmela Troncoso, Jean-Pierre Hubaux, and Jean-Yves Le Boudec. Protecting location privacy: Optimal strategy against localization attacks. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, CCS '12, pages 617–627, New York, NY, USA, 2012. ACM.

[111] Yi Song, Daniel Dahlmeier, and Stephane Bressan. Not so unique in the crowd: A simple

and effective algorithm for anonymizing location data. In *International Workshop on Privacy Preserving IR (PIR 2014)*, pages 19–24, July 2014.

[112] Mudhakar Srivatsa and Mike Hicks. Deanonymizing mobility traces: Using social network as a side-channel. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, CCS '12, pages 628–637, New York, NY, USA, 2012. ACM.

[113] Kaixin Sui, Youjian Zhao, Dapeng Liu, Minghua Ma, Lei Xu, Li Zimu, and Dan Pei. Your trajectory privacy can be breached even if you walk in groups. In *2016 IEEE/ACM 24th International Symposium on Quality of Service (IWQoS)*, pages 1–6, June 2016.

[114] Latanya Sweeney. Simple demographics often identify people uniquely. 2000.

[115] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[116] Manolis Terrovitis and Nikos Mamoulis. Privacy preservation in the publication of trajectories. In *Proceedings of the The Ninth International Conference on Mobile Data Management*, MDM '08, pages 65–72, Washington, DC, USA, 2008. IEEE Computer Society.

[117] Anthony Tockar. Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset. *Neustar*, 2014.

[118] Christof Ferreira Torres and Rolando Trujillo-Rasua. The fréchet/manhattan distance and the trajectory anonymisation problem. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 19–34. Springer, 2016.

[119] Goce Trajcevski, Ouri Wolfson, Klaus Hinrichs, and Sam Chamberlain. Managing Uncertainty in Moving Objects Databases. *ACM Transactions Database Systems*, 29(3):463–507, September 2004.

[120] Rolando Trujillo-Rasua and Josep Domingo-Ferrer. On the privacy offered by (k, δ)-anonymity. *Information Systems*, 38(4):491 – 494, 2013. Special section on BPM 2011 conference.

[121] Zhen Tu, Kai Zhao, Fengli Xu, Yong Li, Li Su, and Dapeng Jin. Protecting Trajectory from Semantic Attack Considering k-Anonymity, l-diversity and t-closeness. *IEEE Transactions on Network and Service Management*, pages 1–1, 2018.

[122] Zhen Tu, Kai Zhao, Fengli Xu, Yong Li, Li Su, and Depeng Jin. Beyond k-anonymity: protect your trajectory from semantic attack. In *14th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), 2017*, pages 1–9. IEEE, 2017.

[123] Jayakrishnan Unnikrishnan and Farid M. Naini. De-anonymizing private data by matching statistics. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1616–1623, October 2013.

[124] Sandesh Uppoor, Oscar Trullols-Cruces, Marco Fiore, and Josè M. Barcelo-Ordinas. Generation and analysis of a large-scale urban vehicular mobility dataset. *IEEE Transactions on Mobile Computing*, 13(5):1061–1075, May 2014.

[125] Jennifer Valentino-DeVries, Natasha Singer, Michael H. Keller, and Aaron Krolik. Your apps know where you were last night, and they're not keeping it secret. *The New York Times*, December 2018.

[126] Narseo Vallina-Rodriguez and Srikanth Sundaresan. 7 in 10 smartphone apps share your data with third-party services. *The Conversation*, July 2017.

[127] Paul Vines, Franziska Roesner, and Tadayoshi Kohno. Exploring adint: Using ad targeting for surveillance on a budget – or – how alice can buy ads to track bob. In *Proceedings of the 16th ACM Workshop on Privacy in the Electronic Society*, WPES'17. ACM, 2017.

[128] Huandong Wang, Chen Gao, Yong Li, Gang Wang, Depeng Jin, and Jingbo Sun. De-anonymization of mobility trajectories: Dissecting the gaps between theory and practice. In *Proceedings of the Network and Distributed Systems Security Symposium*, NDSS'17, 2018.

[129] Yonghui Xiao and Li Xiong. Protecting locations with differential privacy under temporal cor-

relations. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1298–1309. ACM, 2015.

[130] Fengli Xu, Zhen Tu, Yong Li, Pengyu Zhang, Xiaoming Fu, and Depeng Jin. Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1241–1250, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[131] Dingqi Yang, Daqing Zhang, Vincent W. Zheng, and Zhiyong Yu. Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1):129–142, Jan 2015.

[132] Roman Yarovoy, Francesco Bonchi, Laks V. S. Lakshmanan, and Wendy Hui Wang. Anonymizing moving objects: How to hide a mob in a crowd? In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '09, pages 72–83, New York, NY, USA, 2009. ACM.

[133] Bin Zan, Zhanbo Sun, Macro Gruteser, and Xuegang Ban. Linking Anonymous Location Traces Through Driving Characteristics. In *Proceedings of the Third ACM Conference on Data and Application Security and Privacy*, CODASPY '13, pages 293–300, New York, NY, USA, 2013. ACM.

[134] Hui Zang and Jean Bolot. Anonymization of location data does not work: a large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking*, MobiCom '11, pages 145–156, New York, NY, USA, September 2011. ACM.

[135] Desheng Zhang, Jun Huang, Ye Li, Fan Zhang, Chengzhong Xu, and Tian He. Exploring human mobility with multi-source data at extremely large metropolitan scales. In *Proceedings of the 20th annual international conference on Mobile computing and networking*, pages 201–212. ACM, 2014.

[136] Jun Zhang, Xiaokui Xiao, and Xing Xie. Privtree: A differentially private algorithm for hierarchical decompositions. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, pages 155–170, New York, NY, USA, 2016. ACM.

[137] Yu Zheng, Xing Xie, and Wei-Ying Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data(base) Engineering Bulletin*, June 2010.