IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 July 2024, accepted 6 September 2024, date of publication 16 September 2024, date of current version 30 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3462088

RESEARCH ARTICLE

TrajectGuard: A Comprehensive Privacy-Risk Framework for Multiple-Aspects Trajectories

FERNANDA OLIVEIRA GOMES[1,2], ROBERTO PELLUNGRINI[3], ANNA MONREALE[1], CHIARA RENSO[4], AND JEAN EVERSON MARTINA[2]

[1]Department of Computer Science, University of Pisa, 56126 Pisa, Italy
[2]Graduate Program on Computer Science, Department of Informatics and Statistics, Federal University of Santa Catarina (UFSC), Florianópolis 88040-370, Brazil
[3]Classe di Scienze - Scuola Normale Superiore, 56126 Pisa, Italy
[4]Institute of Information Science and Technologies (ISTI)-National Research Council (CNR), 56124 Pisa, Italy

Corresponding author: Chiara Renso (chiara.renso@isti.cnr.it)

This work was supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brazil (CAPES)—Finance Code 001; in part by SoBigData.it through European Union (EU)—NextGenerationEU—National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza—PNRR)—Project: "SoBigData.it—Strengthening the Italian RI for Social Mining and Big Data Analytics"—Prot. IR0000013—Avviso n. 3264 del 28/12/2021; and in part by the PNRR-M4C2-Investimento 1.3, Partenariato Esteso PE00000013-"FAIR-Future Artificial Intelligence Research"-Spoke 1 "Human-Centered AI" funded by European Commission under the NextGeneration EU Programme.

ABSTRACT With the rise of the Internet of Things (IoT), social networks, and mobile devices, vast amounts of mobility data are continuously generated. These data encompass diverse location information from various sources, including smart vehicles, sensors, wearables, and social media platforms. By leveraging these data, we explore the semantic enrichment of trajectory components related to moving objects and locations, bringing the so-called multiple-aspects trajectories and relative privacy issues. Privacy risk analysis is crucial for the earlier detection of privacy problems, particularly when dealing with semantically enriched trajectories. In this study, we introduced the **TrajectGuard** privacy risk assessment framework. **TrajectGuard**, an extension of PRUDEnce, achieved significant results by formulating and assessing the privacy risk of multiple-aspects trajectories under several proposed attacks. The framework introduced a nuanced risk evaluation using *AspectGuard* and conducted fair privacy assessments on anonymized datasets using *AnonimoGuard*. Its adaptability and versatility make **TrajectGuard** a valuable tool for preserving data privacy with multiple-aspects.

INDEX TERMS Multiple-aspects trajectories, privacy, privacy risk, privacy risk assessment, re-identification, trajectory, human mobility.

I. INTRODUCTION

The widespread adoption of mobile devices equipped with location-tracking technologies such as GPS, Wi-Fi, and 3G/4G has led to an increased collection of trajectory data, providing insights into users' spatio-temporal evolution. In its raw form, a trajectory is a sequence of spatiotemporal points that reveal the object's position at specific times. However, trajectory data can be semantically enriched, as introduced by Spaccapietra et al. [1].

The rise of the Internet of Things (IoT), combined with the popularity of social networks and mobile devices, has generated vast amounts of data every second. These data, originating from diverse sources such as smart vehicles, houses, sensors, wearables, appliances, social networks (e.g., Facebook, Instagram, Twitter, LinkedIn), and location-based services, can be leveraged to enhance trajectory components. Enriched data includes information about moving objects (e.g., heartbeat, mood, and blood pressure) or spatiotemporal points (e.g., temperature, air quality, and noise pollution). Integrating these diverse data with trajectories creates a sophisticated trajectory model known as a *multiple-aspects trajectory* [2], [3]. Although multiple-aspects trajectories

The associate editor coordinating the review of this manuscript and approving it for publication was Amjad Mehmood.

provide valuable insights into human mobility, benefiting from various fields such as security, urban planning, public transport management, and epidemic prevention, their utilization raises significant privacy concerns during collection and publication. A recent interest in European common data spaces[1] and mobility data sharing[2] creates a new landscape in which data from different sources are combined. When location data are involved, they tend to pose new privacy concerns that must be investigated.

A substantial risk of privacy violations arises for the individuals involved, as the data encompass highly sensitive and personal information. This vulnerability could lead to dangerous privacy breaches. A considerable threat emanates from re-identification attacks, which focus on identifying individuals or locations within trajectory datasets and presenting a significant threat to privacy. Studies on raw trajectory datasets underscore this risk, revealing that merely four spatiotemporal points can re-identify 95% of the individuals in a low-granularity trajectory dataset [4]. Notably, the top three locations in a path are sufficient to identify over 80% of the individuals [5]. Such disclosures raise privacy concerns, as location data can lead to intrusive inferences about habits, social behavior, and even religious and sexual preferences [6], heightening the risk of activities such as stalking. The implementation of privacy-preserving mechanisms are crucial for data publishing.

An important stage in any privacy-preserving process involves privacy risk assessment, a process aimed at comprehending which individuals in the data are susceptible to privacy violations and quantifying the associated risk. The *Lei Geral de Proteção de Dados* (LGPD) in Brazil and the *General Data Protection Regulation* (GDPR) in Europe have established principles and requirements for processing personal data. They impose responsibilities on data providers to handle data in a manner that ensures data protection. Hence, data providers must conduct a quantitative assessment of privacy risks oversee. Numerous methodologies have been proposed to evaluate the privacy risks of individuals across various types of data [7], [8], [9], [10], [11]. Recent research, exemplified by studies such as [12], [13], and [14], has focused on privacy-preserving risk assessment of trajectories. However, the state of the art reveals a gap in the trajectory privacy research. More specifically, additional analyses are required to address privacy risks associated with multiple-aspect trajectories. Quantifying privacy risk is essential for informed decision-making when selecting suitable privacy-preserving anonymization techniques. No analysis of the privacy risk associated with multiple-aspects trajectories or data with several dimensions is available.

Another crucial point is that methods devised to ensure privacy in raw and semantic trajectories may not directly apply to multiple-aspect trajectories, given their heterogeneous and multi-dimensional nature. Although we have some privacy-preserving methods with another type of aspect other than semantic location, such as [15], [16], [17], the existing literature lacks investigations demonstrating the viability of adapting current privacy-preserving methods against re-identification to this new paradigm, encompassing both privacy-risk assessment and anonymization strategies.

Privacy risk estimation becomes notably complex when dealing with multiple-aspect trajectories. Privacy risk estimation encounters several challenges when dealing with multiple-aspects. The varying granularity and semantics of different components within a multiple-aspect trajectory necessitate redesigning existing methodologies. This is essential for both efficiency and effectively addressing the diverse nature of the data.

For instance, consider a multiple-aspect trajectory that may consist of a moving object's location, time, temperature, and age range. Extending the existing methodologies designed for raw trajectories, containing only space and time, and handling multiple-aspects is not straightforward. The diverse nature of these aspects requires redesigning methodologies to account for their differences efficiently.

Our study addressed these challenges by providing tailored solutions for estimating privacy risks for multiple-aspect trajectories. This ensures that our approach effectively captures the nuances of each aspect while efficiently assessing privacy risks.

The main contribution of this study is the **TrajectGuard** framework. **TrajectGuard** is an extension of the PRUDEnce framework [12] that specifically focuses on evaluating privacy risks associated with multiple-aspect trajectories. It also introduces two new features, *AspectsGuard* and *AnonimoGuard*, which enhance the framework's privacy risk assessment capabilities. The framework provides definitions and mathematical formulations for assessing privacy risks associated with multiple-aspect trajectories. It addresses potential threats that could compromise individuals' privacy and evaluates the computation of risk and its distribution within the three experimental datasets.

AspectsGuard and *AnonimoGuard* are integral components of the **TrajectGuard** framework, which extends the capabilities of the PRUDEnce framework. *AspectsGuard* was introduced to evaluate the impact of single aspects or their combinations on privacy risk by leveraging minimal sample uniqueness on background knowledge combinations. This assessment extends the analysis beyond single-aspect contributions to investigate the collective risk arising from the interaction between multiple-aspects.

AnonimoGuard offers a privacy risk assessment for anonymized datasets by retaining initial background knowledge and comparing it with the anonymized version. This assessment aims to provide a fair and reasonable privacy risk evaluation for anonymized datasets.

TrajectGuard, with its components *AspectsGuard* and *AnonimoGuard*, extends the capabilities of the PRUDEnce framework to support researchers, practitioners, and companies in evaluating the privacy risk of multiple-aspects

[1] https://digital-strategy.ec.europa.eu/en/policies/data-spaces
[2] https://digital-strategy.ec.europa.eu/en/policies/mobility-data

trajectory datasets enriched with various heterogeneous semantic dimensions. This enables users to identify the riskiest data by comparing the risk associated with each individual, thus facilitating informed decision-making and enhancing privacy protection measures.

The remainder of this paper is organized as follows. Section II provides the data definitions of trajectories, the state-of-the-art PRUDEnce framework [12], and a special uniqueness detection algorithm. Section III presents the state-of-the-art privacy risk assessment frameworks. Section IV introduces the **TrajectGuard** framework. Section V presents the experimental details, results, and final discussion. Finally, in Section VI, we discuss our conclusions and future work.

II. BASIC CONCEPTS

In this section, we introduce the necessary background to understand our work.

A. TRAJECTORY

A trajectory denotes the spatio-temporal evolution of a moving object. The representation of trajectories varies based on the level of semantic information linked to the pure movement data. Trajectories can range from *raw trajectories*, which only include spatiotemporal coordinates and no semantics, to *semantic trajectories*, and further to *multiple-aspects trajectories*, introducing a higher level of semantic complexity.

B. RAW TRAJECTORIES

A trajectory, or raw trajectory, represents the pure spatiotemporal part of the movement and is described as a discrete sequence of points. It is formally defined in Definition 1. Each point is a tuple containing spatial coordinates, and a timestamp [18], as detailed in Definition 2, which we call trajectories features in this work. Furthermore, a segment of the trajectory is named a sub-trajectory, with the option to consider a sub-trajectory itself as a trajectory, as described in Definition 3 [18].

Definition 1 (Trajectory): A trajectory T is a sequence of spatio-temporal points, denoted as $T = \{p_0(x_0, y_0, t_0), \ldots, p_n(x_n, y_n, t_n)\}$. Here, x_i and y_i (for $i = 0, 1, \ldots, n$) are spatial coordinates in the set of real numbers $\mathbb{R}$, and t_i represents the time in the set of positive real numbers $\mathbb{R}^+$. The parameter n indicates the size of the trajectory, with $t_0 < t_1 < \ldots < t_n$, ensuring that the trajectory follows a chronological order.

Definition 2 (Point): A point p is a tuple (x, y, t), where x and y are spatial coordinates representing a location, and t is the timestamp representing the time at which the visit to that location occurred.

Definition 3 (Sub-Trajectory): A sub-trajectory s of a trajectory T is a sequence of points $s = \{p_{i_1}, p_{i_2}, \ldots, p_{i_k}\}$, where $0 \leq i_1 < i_2 < \ldots < i_k \leq n$ and $1 \leq k < n$. Each point p_{i_j} (for $j = 1, 2, \ldots, k$) belongs to the trajectory T. This means that s is an ordered subsequence of T containing at least one point and fewer than all points of T.

In this work, we use the terms *point* or *visit* to refer to a single element of a trajectory, while with the term *location* l, we refer to the point's spatial information. A sub-sequence of locations (Definition 4) is an ordered list of locations.

Definition 4 (Sub-Sequence): Let $\mathcal{L} = \{l_1, l_2, \ldots, l_w\}$ denote a set of locations. A sequence $S = \langle s_1, s_2, \ldots, s_m \rangle$, where $s_i \in \mathcal{L}$, is an ordered list of locations, and a location can occur multiple times in the sequence.

A sequence $T = \langle t_1, t_2, \ldots, t_z \rangle$ is a sub-sequence of S (denoted $T \preccurlyeq S$) if there exist integers $1 \leq i_1 < i_2 < \cdots < i_z \leq m$ such that:

$$t_j = s_{i_j} \quad \text{for} \quad j = 1, 2, \ldots, z.$$

This means that each element t_j in the sub-sequence T is equal to the element s_{i_j} in the original sequence S at position i_j, ensuring that T follows the order of S.

C. SEMANTIC TRAJECTORIES

A semantic trajectory is constructed based on stops and moves, allowing moving objects to enrich their trajectories with semantic information relevant to their application domain [19]. A semantic trajectory is essentially a raw trajectory enhanced by semantic information and is often accompanied by one or more complementary segmentations [20].

Definition 5 (Semantic Trajectory): A semantic trajectory ST is a finite sequence $I_\Delta, I_\Theta, \ldots, I_n$, where each I_k represents either a stop or a move.

Definition 6 (Stop): A stop is a sub-trajectory that starts at time t_i and ends at time t_j. For a stop, the object must remain in a given semantic location for a minimum period $\Delta t = t_j - t_i$, where $t_j > t_i$. Additionally, each stop must be distinct, such that:

$$\text{stop}_1 \cap \text{stop}_2 \cap \ldots \cap \text{stop}_n = \emptyset$$

Definition 7 (Move): A move is a spatio-temporal segment, move_x, defined between stop_a and stop_b. Here, t_i (the end time of stop_a) is the start time of the move, and t_j (the start time of stop_b) is the end time of the move. Thus, the duration of the move is $\Delta t = t_j - t_i$.

The concept of semantic trajectory based on stops and moves was first introduced by Spaccapietra et al. [1] and Alvares et al. [21]. Stops are crucial trajectory elements. They are characterized by start and end times and occur when a moving object remains at a location for a minimum duration, as defined in Definition 6. Moves are sub-trajectories of sampling points delineating the displacement between two consecutive stops, as presented in Definition 7. Another notable characteristic of semantic trajectories is their capacity to incorporate contextual information, such as the mode of transportation used or the names of places visited [22].

D. MULTIPLE-ASPECTS TRAJECTORIES

The foundation of our proposal lies in the concept of multiple-aspects trajectories. We aim to explore the implications of this trajectory data on privacy, and devise appropriate methods to evaluate the privacy risks associated with such data.

This intricate and heterogeneous trajectory, characterized by numerous semantic dimensions, is called a multiple-aspect trajectory. This subsection introduces the definitions of the multiple-aspect trajectories initially proposed in [2].

The aspect, described in Definition 8, is a real-world feature that contextualizes the trajectory. It can be a number, range, text, geometry, or any complex object type.

Definition 8 (Aspect): An aspect A = desc represents a relevant real-world feature described by desc.

This aspect can be classified according to durability. There are three types of aspects:

- The **volatile aspect (VA)** frequently varies during object movement. For example, places visited (or stops or POI), heart rate, social media posts, and weather conditions.
- The **long term aspect (LA)** does not change during the entire trajectory. Examples of this Aspect are the place (e.g., city, state, county) where the trajectory occurs, occupation, marital status, and age.
- The **permanent aspect (PA)** holds throughout the life of an object; it is always related to the object, such as birthday or birthplace.

In Section IV, we introduce attacks related to these aspects and evaluate how these different aspects may affect privacy risk.

Each aspect of the trajectory may encompass contextual details, such as a restaurant, thereby expanding the dimensions of the data. Furthermore, these aspects, which represent contextual information, exhibit heterogeneity. As the number of aspects within the trajectory increases, so does its intricacy, enabling more in-depth analyses and inferences.

Now that we have presented the concept of aspects and their types, first introduced by [2], we can formally define a multiple-aspects trajectory, as shown in Definition 9.

Definition 9 (Multiple-Aspects Trajectory): A Multiple-Aspects Trajectory $MAT_u = \langle PA_u, MAT_{LA_u}, u\rangle$ is composed of a set of permanent aspects $PA_u = \{pa_1, pa_2, \ldots, pa_h\}$ of a moving object u with length h. The moving object u has a trajectory and a set of long-term aspects $MAT_{LA_u} = \{LAT_{1_u}, \ldots, LAT_{j_u}\}$, where each $LAT_{i_u} = \langle T_{i_u}, LA_{i_u}\rangle$ and T_{i_u} is the i-th trajectory. Each $LA_{i_u} = \{la_1^{i_u}, la_2^{i_u}, \ldots, la_r^{i_u}\}$ is a set of long-term aspects of length r related to the i-th trajectory, where $r \geq 0$.

Each trajectory $T_{i_u} = (p_1, \ldots, p_o)$ is a sequence of points $p_1, \ldots, p_o$, such that $p_z = \langle l_z, t_z, VA_z\rangle$, where VA_z is a non-empty set of volatile aspects. The volatile aspects $VA_z = \{va_1, va_2, \ldots, va_o\}$ are of size o.

Each MAT_u is part of a multiple-aspects trajectory dataset D. This D is part of a mobility database denoted as $\mathcal{D}$, as described in Definitions 10 and 11.

Definition 10 (Mobility Database): Let $\mathcal{D}$ be a mobility database and $D \subseteq \mathcal{D}$ a mobility dataset extracted from $\mathcal{D}$. A mobility database $\mathcal{D}$ is an organized collection of mobility data.

Definition 11 (Mobility Dataset): A mobility dataset D is a subset of $\mathcal{D}$, where $D = \{T_1, T_2, \ldots, T_n\}$ and $T_u \equiv D_u$ is the trajectory data structure of a moving object u $(1 \leq u \leq n)$. In the case of multiple-aspect trajectories, we represent it as $D = \{MAT_1, MAT_2, \ldots, MAT_n\}$.

E. PRUDEnce FRAMEWORK

The privacy risk assessment framework PRUDEnce, developed by [12], is an essential tool used to evaluate risks related to privacy, particularly when working with trajectory data. Its main purpose is to help data providers (DPs)—entities that collect, manage, and share personal data—make well-informed decisions that balance privacy with data utility. PRUDEnce was created to ensure compliance with the General Data Protection Regulation (GDPR) set forth by the European Union, specifically addressing Article 25, which mandates that privacy protection be integrated into system design from the outset. Moreover, the framework assists organizations in conducting data protection impact assessments (DPIAs). Though originally designed to comply with GDPR, PRUDEnce's adaptable structure allows it to be applied across various legal and regulatory environments worldwide.

One of PRUDEnce's core applications is assessing privacy risks associated with data sharing, particularly when raw personal data is transferred between a data provider (DP) and a service developer (SD). In these scenarios, background knowledge plays a vital role. Background knowledge refers to any external information that attackers may already have about individuals, which they could use to undermine privacy protections. Understanding and evaluating this background knowledge is crucial for determining how effective privacy attacks might be, and thus, it is a central element of the PRUDEnce framework.

Background knowledge refers to the specific information an adversary may have about a particular user, denoted as u. The PRUDEnce framework assesses privacy risks by considering various levels of this knowledge, ranging from minimal to extensive. This approach allows for more accurate risk evaluation and supports informed decision-making. It is important to define the attack models and how much background knowledge the adversary has to balance keeping the data useful and protecting privacy.

A background knowledge category includes different information an attacker might possess about an individual. In the context of mobility data, for example, these categories could include data points such as geographic locations, timestamps, how often a person visits specific areas, or the likelihood of returning to a given location. The background knowledge configuration, denoted as k, refers to the quantity of information that the adversary holds.

For example, if an attacker knows $k = 3$ specific points along a user's movement path, this configuration reflects a particular set of known data points. Each piece of information the attacker has is an instance of background knowledge.

For example, in a real-world case involving a ride-sharing company, an attacker might know the geographic location and timestamp of a user's previous rides. The geographic location and timestamp will be the background knowledge categories. The attacker knows, for instance, that a user has taken three rides ($k = 3$) between specific locations at particular times. This is a concrete instance of background knowledge that could be used to compromise privacy. The background knowledge configurations would include all possible combinations of three data points from the user's ride history, each consisting of a location and a timestamp.

The formal definition of these concepts, as presented in [12], is shown below:

Definition 12 (Background Knowledge: Categories, Configurations, and Instances): A background knowledge category, $\mathcal{B}$, consists of several dimensions of data an attacker could possess. A background knowledge configuration, represented as $B_k = \{b_1, b_2, \ldots, b_w\}$, belongs to the set of background knowledge categories $\mathcal{B} = \{B_1, B_2, \ldots, B_n\}$, where k indicates the number of known data points. Each element b within B_k is an instance of background knowledge.

Given a database $\mathcal{D}$, let D represent a subset of data drawn from $\mathcal{D}$, aggregated over various dimensions, with D_u being the subset of records belonging to a specific individual u.

Definition 13 (Re-Identification Probability): To evaluate the risk of re-identification based on a specific background knowledge instance, we define a function $matching(d, b)$, which determines if a record $d \in D$ corresponds to a specific background knowledge instance $b \in B_k$. We then define $M(D, b) = \{d \in D \mid matching(d, b) = \text{True}\}$, which represents the set of records that match the instance b. The probability of re-identification for a user u based on the background knowledge instance b in dataset D is given by:

$$PR_D(d = u \mid b) = \frac{1}{|M(D, b)|}$$

This probability reflects the likelihood that a record $d \in D$ corresponds to the individual u, given the adversary's knowledge instance $b \in B_k$.

The matching function, $matching(d, b)$, is used to check if a record d in the dataset D aligns with the known background knowledge instance b. The likelihood of re-identification depends on how much background knowledge the attacker holds. PRUDEnce determines re-identification risk by calculating the highest probability of re-identification across all background knowledge instances in a given configuration, as formalized in Definition 14.

Definition 14 (Re-Identification or Privacy Risk): The re-identification risk for a user u concerning a background knowledge configuration B_k is defined as the highest probability of re-identification for all instances within B_k. The risk is $Risk(u, D) = \max PR_D(d = u \mid b)$, where $b \in B_k$. The minimum risk, representing a random guess in the dataset D, is expressed as $\frac{|D_u|}{|D|}$, and $Risk(u, D) = 0$ when the individual u is not present in the dataset.

Individuals may be exposed to varying privacy risks depending on the configuration of their background knowledge used during an attack. Each attack scenario is tailored to a particular type of background knowledge, and different configurations $\{B_1, \ldots, B_m\}$ are analyzed. For every configuration B_k, the re-identification probabilities for all instances b are calculated. The highest of these probabilities determines the privacy risk associated with the individual for that specific configuration.

F. SUDA ALGORITHM OVERVIEW

This subsection presents the state-of-the-art Special Unique Detection Algorithm, SUDA, proposed in [23]. SUDA is an algorithm designed to identify unique aspect sets and Minimal Sample Uniques (MSUs). Identifying aspect sets at the record level sets the stage for evaluating these sets to determine MSUs. MSUs play a crucial role in understanding the risk associated with individual records, considering the size and number of these unique sets.

TABLE 1. Data set example.

ID	Age Range	Labor Status	Residence	Kids
1	30–35 years	Employed	Urban	2
2	30–35 years	Employed	Urban	2
3	30–35 years	Employed	Rural	2
4	30–35 years	Employed	Rural	2
5	25–30 years	Unemployed	Urban	1
6	20–25 years	Unemployed	Urban	3
7	20–25 years	Unemployed	Urban	2
8	40–45 years	Employed	Rural	1

Table 1 illustrates an example of an MSU found in record 8: set {Employed, 1}. This set is an MSU because neither of its subsets, {Employed} or {1}, is unique in the sample, and the record itself is unique. This example underscores the significance of identifying MSUs to understand the re-identification risk associated with individual records.

Two critical considerations influence the risk associated with records:

1) The smaller the size of the MSU within a record, the greater is the risk of the record.
2) The larger the number of MSUs possessed by a record, the greater the risk of that record [24].

The next section will present the related work regarding risk assessment frameworks.

III. RELATED WORK

In this section, we explore various privacy risk assessment methods. We thoroughly compared these methods with our own and highlighted their distinctions and advancements.

Regarding quantitative privacy risk assessment, Song et al. [25] proposed a modification-based anonymization approach and evaluated the privacy risk based on the uniqueness of the trajectory data. Achara et al. [26] investigated the privacy implications of a list of apps installed by users on smartphones, emphasizing the re-identifiability issue. Basu et al. [27] considered the cost of privacy attacks and provided a more practical assessment of the privacy risks of data release. The evaluation of this model involved the use of k-anonymized data, adding a real-world dimension to the analysis of privacy risks. In Armando et al. [28], the proposed framework integrates runtime risk assessment into information disclosure access control by utilizing disclosure risk for decision-making. Other studies in the literature explore re-identification risk as a privacy measure within the realms of network and social media data [29], [30]. In Cecaj et al., they combined network data with mobile phone data to achieve the re-identification of individuals [31]. Khalfoun et al. [32] presented in this paper selects the optimal Numerous Location Privacy Protection Mechanisms and configuration without exposing raw geo-located traces. In Pellungrini et al. [33], they propose a study of privacy risk for social network data, and in Mariani et al., a study of privacy risk for psychometric profile [34]. In the study by Silva et al. [35], the Personal Data Analyser is introduced, employing automated data monitoring with Regular Expressions, NLP, and machine learning to boost privacy and reduce risks. In Guo et al. [36], they propose a risk-sensing approach to vehicle location privacy based on the continuous adaptive risk and trust assessment strategy.

In this work, we adopted the PRUDEnce framework introduced by Pratesi et al. [12]. This framework offers a detailed methodology for computing privacy risk in a data-driven manner. Essentially, PRUDEnce revolves around the foundational principle of k-anonymity, wherein privacy risk assessment is intricately linked to the dimensions of the k-sets associated with each individual in the dataset. The practical effectiveness of PRUDEnce was demonstrated using real mobility data and by exploring the presence, trajectory, and road segment data formats. Our decision to use PRUDEnce was based on its flexible extension and suitability for trajectory data.

However, the computational intensity of PRUDEnce has encouraged the exploration of machine learning approaches that aim to predict privacy risks, circumventing the need for computationally exhaustive processes. Pellungrini et al. [13] presented a swift and adaptable method for estimating privacy risk in human mobility data. In EXPERT framework developed by Naretto et al. [14]. This framework not only refines PRUDEnce by introducing a machine learning methodology proficient in directly forecasting privacy risk from sequential data but also enhances the interpretability of these predictions. Another study proposed by Naretto et al. [37] presents an optimization of EXPERT, the EXPHLOT. They use distinct time series classifications, such as ROCKET and INCEPTIONTIME, to improve risk prediction while reducing computation time. These innovations collectively contribute to a more streamlined and interpretable privacy risk assessment, effectively addressing the computational challenges inherent in traditional PRUDEnce computations.

A. CORRELATION WITH WORKS

Table 2 compares various related studies in the privacy risk assessment field. Each row represents a different work, whereas the columns indicate the specific characteristics or features of the study. Each feature and its correlation with the work are explained as follows.

- **Quantitative**: These works involve numerical data or measurable factors for assessing privacy risks, as in all works in the table.
- **Raw Trajectory**: Direct application to raw trajectory data to address privacy risks. Works such as [12], [13], [14], [25], [32], [36], and [37] fall into this category. Unlike most approaches, our framework was designed to evaluate trajectories with aspects.
- **Individual Risk**: Focuses on assessing or addressing privacy risks at an individual level, as indicated in all works.
- **Computation Improvements**: Includes works that mention enhancing the efficiency or reducing the computational requirements of privacy risk assessment methodologies, as in [13], [14], and [37]. Both approaches require an initial conventional risk analysis to generate training data for predicting the risk of new data.
- **Re-evaluation**: This refers to works that mention or imply a continuous or iterative process of evaluating privacy risks, as in [12]. Our work also introduces a novel method for assessing risk after applying privacy-preserving mechanisms. The key distinction of our method lies in its more realistic approach to evaluating risk in anonymized datasets.
- **Machine Learning**: Utilizes machine learning techniques to assess or predict privacy risks. Machine learning has been used to estimate privacy risk levels, as in [13], [14], [32], [35], [36], and [37].
- **Quality**: Works that aim to improve or ensure the quality of data or privacy protection measures as in [12], [13], and [32].

Machine learning algorithms and quality evaluations are beyond the scope of this work. However, our innovation lies in incorporating aspect-level risk assessment, named *AspectGuard* and introducing *AnonimoGuard*, a new method for anonymity risk assessment, and developing data models for risk assessment tailored to multiple-aspects trajectories.

IV. TRAJECTGUARD

In this section, we introduce one of the main contributions of this work, the **TrajectGuard** framework.[3] This framework extends *PRUDEnce* by introducing new functionalities specifically designed to evaluate privacy risks associated with data containing multiple-aspects. Within this extended framework, we present three key contributions: a comprehensive data model for handling multiple-aspects and attacks, a privacy risk assessment that considers the contribution of these aspects, and an evaluation of the privacy risk within

[3] https://github.com/oliveiragomesphd/TrajectGuard/

TABLE 2. Comparison between the related work.

Related Work	Quantitative	Raw Trajectory	Semantic Trajectory	Multiple-Aspects Trajectory	Individual Risk	Aspect Risk	Computation Improvements	Re-evaluation	Machine Learning	Quality
Song *et al.* (2014)	✓	✓			✓					
Achara *et al.* (2015)	✓				✓					
Basu *et al.* (2014)	✓				✓					
Pratesi *et al.* (2018)	✓	✓			✓			✓		✓
Pellungrini *et al.* (2017)	✓	✓			✓		✓		✓	✓
Naretto *et al.* (2020)	✓	✓			✓		✓		✓	
Mariani *et al.* (2021)	✓				✓					
Pellungrini *et al.* (2021)	✓				✓					
Khalfoun *et al.* (2021)	✓	✓			✓				✓	✓
Silva *et al.* (2022)	✓				✓				✓	
Naretto *et al.* (2023)	✓	✓			✓		✓		✓	
Guo *et al.* (2024)	✓	✓			✓				✓	
TrajectGuard	✓		✓	✓	✓	✓		✓		

anonymous datasets. Figure 1 depicts the similarities between the components within the blue rectangle and the newly introduced modules, contrasting them with the distinctive contributions of **TrajectGuard** highlighted within the red rectangle.

TrajectGuard, which is an extension of the PRUDEnce framework, comprises three main modules: the Multiple-Aspects Trajectories data model and attacks, *AspectGuard*, and *AnonimoGuard*.

1) *Multiple-Aspects Trajectories data model and attacks*: We define the multiple-aspects data model and outline the mathematical formulation for a range of privacy attacks on multiple-aspects trajectories by defining and analyzing potential threats that could compromise individuals' privacy.
2) *AspectGuard*: This specifically focuses on evaluating the impact of single aspects or their combinations on privacy risk. Unlike the current process, in which we assess the risk related to values, we address the risk to the aspect itself or a combination of aspects. This assessment leverages the minimal sample uniqueness of background knowledge combinations. Our analysis extends beyond single-aspect contributions to investigate the collective risk emerging from the interaction of multiple-aspects. Moreover, uniqueness emerged as the primary factor influencing re-identification risk. To offer an alternative perspective on evaluating re-identification risk, we introduce a novel evaluation metric based on the concept of *Minimal Sample Unique*, which compares the risk of each user's multiple-aspects trajectory data with the others.
3) *AnonimoGuard*: A privacy risk assessment for anonymized datasets. We retained the initial background knowledge and compared it with the anonymized version by applying it to both the trajectory data and the permanent aspect set.

A. DATA MODELING AND PRIVACY RISK

Our framework was designed to work the unique characteristics of multiple-aspect trajectories. According to [13], a background knowledge category represents an adversary's specific information regarding the particular dimensions of an individual's mobility data. The typical dimensions of mobility data are space and time. However, in the context of multiple-aspects trajectories, we have proposed one additional dimension: a moving object. These dimensions can be classified into three primary categories: spatial, temporal, and personal. Furthermore, the dimensions are classified according to their variability: volatile, long-term, and permanent. An instance of background knowledge is specific information possessed by an adversary. To provide an example of instances within background knowledge categories: location semantic (spatial volatile), year (temporal long-term), and city of birth (personal permanent).

Definition 15 (Probability of Re-Identification in TrajectGuard): Given an attack, function *matching*(D, b) indicating

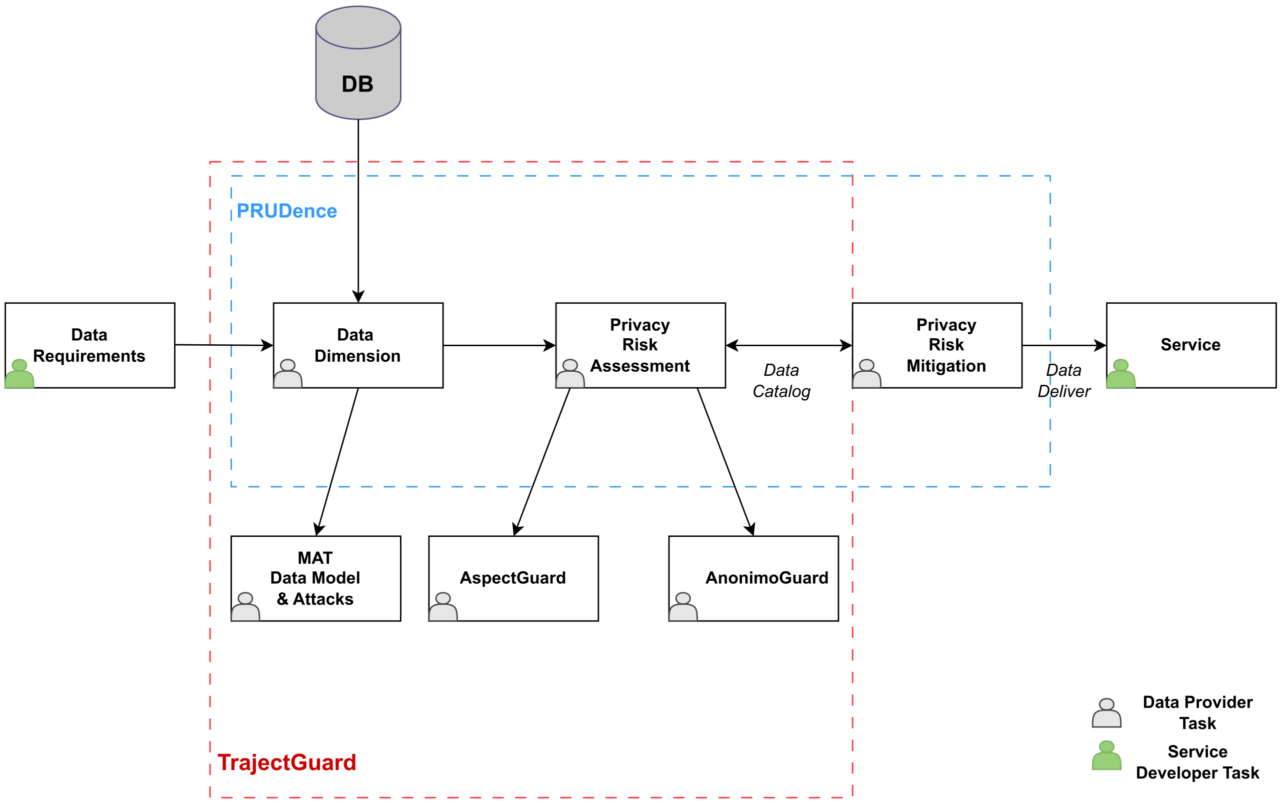


FIGURE 1. TrajectGuard framework.

whether a record $d \in D$ matches the instance of background knowledge configuration $b \in B_k$, and function $M(D, b) = \sum_{i=1}^{n} matching(D_i, b)$, we define the probability of re-identification of an individual u in dataset D as $PR_D(d = u|b) = \frac{M(D_u,b)}{M(D,b)}$.

Different from PRUDEnce, in Definition 13, the probability of re-identification associated with this knowledge is determined by the number of individual trajectories containing this specific instance divided by the total number of trajectories including the same instance. In datasets containing multiple aspect trajectories, each individual may have one or more trajectories containing the same background knowledge instance, and it is important to consider this in the evaluation process. For example, if a background knowledge instance is present in two trajectories, using PRUDEnce, the probability of re-identification would be 50%. However, in **TrajectGuard**, both trajectories could be from the same individual, and the probability of re-identification would be 100%.

The same situation exists in risk evaluation. It can be evaluated in two ways: the risk per individual's trajectory or the maximum risk value among the individual's trajectories. The choice between these approaches depends on the specific goals of the analysis.

Let U represent the set of individuals in the dataset, and D denote a dataset containing multiple-aspects trajectories. Each individual u can have at least one trajectory.

1) Risk per Individual's Trajectory: The risk per individual's trajectory is computed individually for each trajectory belonging to individual u.
2) Maximum Risk Among Individual Trajectories: The maximum risk value among individual trajectories is determined by selecting the highest risk value among all trajectories associated with individual u. This means that each trajectory of the user is evaluated for its privacy risk, and the highest risk value among all trajectories is taken as the final privacy risk.

The next subsection introduces the proposed privacy attacks on multiple-aspects trajectory datasets. It describes, evaluates, and defines them in detail.

B. PRIVACY ATTACKS

This subsection introduces privacy risk attacks on multiple aspect trajectories using previously presented definitions. We cover aspects attacks, location attacks, and visit attacks as well as multiple attacks.

We consider that for each individual u, we have a set $MAT_u = \langle u, PA_u, MAT_{LA_u} \rangle$ containing its permanent aspects, such as birthday and birthplace, and its trajectories with their long-term aspects, such as occupation and age of the individual and state or city of the trajectory as it does not change during it. In each trajectory related to each point, we also have volatile aspects such as location semantics and temperature.

1) PERMANENT ASPECT ATTACK

In this attack, the adversary knows one or more permanent aspects of their victim. Let h be the number of permanent aspects, denoted as $PA = \{pa_1, pa_2, \dots, pa_h\}$, known for individual u. The Permanent Aspect Attack uses background knowledge of k permanent aspects, where $k \leq h$. The set of possible configurations of background knowledge of length k is defined as $B_k = P(PA_u)^k$.

Each instance $b \in B_k$ is a subset of permanent aspects $b \subseteq P(PA_u)^k$ of length k, where $P(PA_u)^k$ represents all possible k-combinations of permanent aspects for individual u. Given $D_u = \{u \in D \mid MAT_u\}$ and $PA_u \subseteq MAT_u$, we define the evaluate function as:

$$\text{matching}(D, b) = \begin{cases} 1, & \text{if } b \subseteq P(PA_u)^k \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

For instance, if an attacker knows that the victim was born in France in 1950, they extract all trajectories of people with these aspects.

2) LONG-TERM ASPECT ATTACK

In this attack, the adversary knows one or more long-term aspects of an individual or its location from one or more trajectories. Let j be the number of trajectories and long-term aspects $MAT_{LA_u} = \langle LAT_{1_u}, \dots, LAT_{j_u} \rangle$ of an individual u known by the attacker, where the sequence of long-term trajectories is *temporal*, meaning the trajectories are ordered in time such that $t_1 < t_2 < \cdots < t_j$, where each t_i represents the timestamp, with t_1 being the time information related to the first point and t_j related to the last point. Each LAT_{i_u} is represented as (T_{i_u}, LA_{i_u}), where $LA_{i_u} = \{la^i_1, la^i_2, \dots, la^i_r\}$ is the i-th set of long-term aspects related to the i-th trajectory.

The *Long-Term Aspect Attack* is performed using background knowledge $B_k = \{(LA_{i_u})^k \mid LAT_{i_u} \in MAT_{LA_u}\}$, where $LT(LA_{i_u})^k$ represents all possible k-combinations of long-term aspects. Since each instance $b \in B_k$ is a subset of long-term aspects $b \subseteq LT(LA_{i_u})^k$, given $MAT_{LA_u} \in MAT_u \in D$ of individual u, $\forall LAT_{i_u} \in MAT_{LA_u}, LA_{i_u} \in LAT_{i_u}$, we define the matching function as:

$$\text{matching}(D, b, u) = \sum_{i=1}^{j} \text{evaluate}(LA_{i_u}, b) \tag{2}$$

$$\text{evaluate}(LA_{i_u}, b) = \begin{cases} 1, & \text{if } b \subseteq LT(LA_{i_u})^k \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

For instance, if an attacker knows that the victim is a female nurse, they extract all trajectories of people with these aspects.

3) VOLATILE ASPECT ATTACK

The adversary knows a set of volatile aspects related to an individual or location for a specific timestamp. Let j be the number of trajectories and long-term aspects $MAT_{LA_u} = \langle LAT_{1_u}, \dots, LAT_{j_u} \rangle$ of an individual u known by the attacker. Each $LAT_{i_u} = (T_{i_u}, LA_{i_u})$ represents the i-th trajectory, where each $T_{i_u} = (p_1, \dots, p_n)$ is a sequence of points. A point in the z-th position $p_z = \langle l_z, t_z, VA_z \rangle$ is composed of location l, timestamp t, and VA_z, which is a non-empty set of volatile aspects $VA_z = \{va_1, va_2, \dots, va_o\}$ of size o.

The Volatile Aspect Attack is performed using background knowledge $B_k = \{V(T_{i_u})^k \mid LAT_{i_u} \in MAT_{LA_u}, p_z \in T_{i_u}\}$, where $V(T_{i_u})^k$ represents all possible k-combinations of volatile aspects sets from trajectory i. Each instance $b \in B_k$ is a subset of volatile aspects $b \subseteq V(T_{i_u})^k$. Given $MAT_{LA_u} \in MAT_u \in D$ of individual u, $\forall LAT_{i_u} \in MAT_{LA_u}, T_{i_u} \in LAT_{i_u}$, and $D_u = \{u \in D \mid MAT_u\}$, we define the matching function as:

$$\text{matching}(D, b) = \sum_{i=1}^{j} \text{evaluate}(T_{i_u}, b) \tag{4}$$

$$\text{evaluate}(T, b) = \begin{cases} 1, & \text{if } b \subseteq V(T_{i_u})^k \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

For instance, if the attacker knows that the victim went to the supermarket and church at a temperature of 25°C, they extract all trajectories of people that contain these aspects.

For all attacks in sequence, the matching function is calculated as follows:

$$\text{matching}(D, b) = \sum_{i=1}^{j} \text{evaluate}(LA_{i_u}, b)$$

The evaluation will change for each attack, as seen in the following subsections.

4) LOCATION SEQUENCE ATTACK

In this attack, introduced in [13], [38], and [39], the adversary knows a subset of the locations visited by the individual and the temporal order of the visits.

Let j be the number of trajectories and long-term aspects $MAT_{LA_u} = \langle LAT_{1_u}, \dots, LAT_{j_u} \rangle$ of an individual u known by the attacker, where the sequence of long-term trajectories is temporal. Each $LAT_{i_u} = (T_{i_u}, LA_{i_u})$ represents the i-th trajectory, where $LA_{i_u} = \{la^i_1, la^i_2, \dots, la^i_r\}$ is the i-th set of long-term aspects related to the i-th trajectory with length r. Each T_{i_u} is a temporal sequence of points $p_1, \dots, p_n$, such that $t_1 < t_2 < t_n$ and $p_z = \langle l_z, t_z, VA_z \rangle$, where z represents the point position in the trajectory and l_z is the location.

The Location Sequence Attack background knowledge is a set of configurations based on k locations. We define the set of possible configurations of background knowledge as $B_k = \{LS(T_{i_u})^k \mid T_{i_u}\}$, where $LS(T_{i_u})^k$ denotes the set of all possible k-subsequences of the locations in set T_{i_u}. We indicate with $a \preccurlyeq_{ls} b$ that a is a subsequence of b. The symbol $\preccurlyeq_{ls}$ represents a subsequence of locations. Each instance $b \in B_k$ is a subsequence of location $l_u \preccurlyeq_{ls} T_{i_u}$, where $l_u = \langle (l_1) \subset p_1, \dots, (l_k) \subset p_k \rangle$. Given $MAT_{LA_u} \in MAT_u \in D$ of individual u, $\forall LAT_{i_u} \in MAT_{LA_u}, T_{i_u} \in LAT_{i_u}$, and $D_u = \{u \in D \mid MAT_u\}$, we define the evaluation function as:

$$\text{evaluate}(T, b) = \begin{cases} 1, & b \preccurlyeq_{ls} LS(T)^k \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

In this attack, the attacker knows that a person went first to a supermarket and then to work but did not know the sequence of places.

5) VISIT ATTACK

In this attack, introduced in [6], [13], [40], [41], and [42], an adversary knows a subset of the locations visited by the individual and the time the individual visits these locations.

Let j be the number of trajectories and long-term aspects $MAT_{LA_u} = \langle LAT_{1_u}, \ldots, LAT_{j_u} \rangle$ of an individual u known by the attacker, where each $LAT_{i_u} = (T_{i_u}, LA_{i_u})$ represents the i-th trajectory. Each T_{i_u} is a temporal sequence of points $(p_1, \ldots, p_n)$, where n represents the numeric position of the last point in the trajectory, $(t_1 < t_2 < t_n)$, and $p_n = \langle l_n, t_n, VA_n \rangle$, where l_n is the location and t_n is the temporal information.

A Visit Attack is performed using background knowledge $B_k = \{LV(T_{i_u})^k \mid T_{i_u}\}$, where $LV(T_{i_u})^k$ denotes the set of all possible k-subsequences of the spatio-temporal points in set T_{i_u}. We indicate with $a \preccurlyeq_{lv} b$ that a is a subsequence of b. The symbol $\preccurlyeq_{lv}$ is used to represent a subsequence of spatio-temporal points (or visits). Each instance $b \in B_k$ is a subsequence of spatio-temporal points $b \preccurlyeq_{lv} T_{i_u}$, where $b = \langle (l_1, t_1) \subset p_1^1, \ldots, (l_k, t_k) \subset p_k \rangle$. Sub-trajectory b positively matches a specific trajectory i if the latter supports b in the spatial and temporal dimensions. Given $MAT_{LA_u} \in MAT_u \in D$ of individual u, $\forall LAT_{i_u} \in MAT_{LA_u}, T_{i_u} \in LAT_{i_u}$, and $D_u = \{u \in D \mid MAT_u\}$, we define the evaluation function as:

$$\text{evaluate}(T, b) = \begin{cases} 1, & b \preccurlyeq_{lv} LV(T)^k \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

In this attack, the attacker knows that, for instance, a person first went to a supermarket at 1 pm and then to work at 2 pm. They know when a person visited and the place.

We also propose an evaluation of combined attacks, that is, attacks where background knowledge combines two previously defined attacks. Table 4 presents a complete list of the attacks and their matching criteria. Table 3 presents a comprehensive list of mathematical symbols and their meanings.

In Figure 2, we illustrate the risk assessment process in **TrajectGuard**. The input is a dataset containing multiple aspects of trajectories. In this example, the attacker attempts a location sequence attack with permanent aspects, which in this case is the user's birthplace. In the first step, all background knowledge configurations are generated with a knowledge size of two, meaning combinations are created with $k = 2$. The second table shows the resulting background knowledge instances. Next, each configuration's probability of re-identification is calculated, as shown in the third table, where each instance has its associated risk evaluated. In the final step, the maximum risk value for each user is identified by getting the maximum trajectory risk value. However, it's also possible to determine the highest risk per user's individual trajectory.

TABLE 3. Multiple attacks symbols.

Symbol	Background Knowledge
	Instance
b_l	Subsequence of locations
b_{lv}	Subsequence of spatiotemporal points
b_p	Subset of permanent aspects
b_{lt}	Subset of long-term aspects
b_v	Subset of volatile aspects
	Configurations
$LS(T)^{k_l}$	k-subsequences of locations from trajectory T
$LV(T)^{k_{lv}}$	k-subsequences of spatiotemporal points from trajectory T
$P(PA_u)^{k_p}$	k-combinations of permanent aspects from individual u
$LT(T)^{k_{lt}}$	k-combinations of long-term aspects from trajectory T
$V(T)^{k_v}$	k-combinations of volatile aspects from trajectory T

C. AspectGuard

In this subsection, we introduce *AspectsGuard*, an enhancement of the PRUDEnce framework, and the contribution of this study, a module specifically focused on evaluating the impact of single aspects or their combinations on privacy risk. Unlike the current process based on [12], in which we assess the risk related to the background knowledge configuration values, we address the risk to the aspect itself or a combination of aspects. This assessment leverages minimal sample uniqueness in background knowledge combinations. Our analysis extends beyond single-aspect contributions to scrutinize the collective risk emerging from the interaction of multiple-aspects.

Moreover, through the comprehensive privacy risk assessment conducted in Section V, we discerned that uniqueness emerged as the primary factor influencing an individual's re-identification risk.

The *AspectGuard Risk Assessment* assesses re-identification risk by considering the aspects of risk contribution at and dataset levels. To enhance this evaluation, we present a Special Uniques Detection Algorithm [43] that measures the contribution of each aspect to re-identification. This algorithm relies on the concept of special uniqueness within records, where a set is considered special unique if it is uniquely representative in the sample across all aspects and concurrently contains at least one *Minimal Sample Unique* (MSU), which is a unique attribute set without any unique subsets [23].

1) CONTRIBUTION CALCULATION

To comprehensively assess the re-identification risk, *AspectGuard* employs a contribution calculation. This calculation quantifies how each aspect contributes to re-identification by counting the number of MSUs involved in each aspect. This approach provides valuable insights into the aspects significantly contributing to re-identification risk.

This definition describes a dataset in the context of multiple aspects of trajectories. It is crucial to emphasize that this

TABLE 4. Multiple attacks summary.

Attack	Matching
Location Sequence Attack with Permanent Aspects	$b_l \preccurlyeq_l LS(T)^{k_l} \wedge b_p \subseteq P(PA_u)^{k_p}$
Location Sequence Attack with Long-Term Aspects	$b_l \preccurlyeq_l LS(T)^{k_l} \wedge b_{lt} \subseteq LT(T)^{k_{lt}}$
Location Sequence Attack with Volatile Aspects	$b_l \preccurlyeq_l LS(T)^{k_l} \wedge b_v \subseteq V(T)^{k_v}$
Location Sequence Attack with Permanent and Long-Term Aspects	$(b_l \preccurlyeq_l LS(T)^{k_l}) \wedge (b_p \subseteq P(PA_u)^{k_p}) \wedge (b_{lt} \subseteq LT(T)^{k_{lt}})$
Location Sequence Attack with Permanent and Volatile Aspects	$(b_l \preccurlyeq_l LS(T)^{k_l}) \wedge (b_p \subseteq P(PA_u)^{k_p}) \wedge (b_v \subseteq V(T)^{k_v})$
Location Sequence Attack with Long-Term and Volatile Aspects	$(b_l \preccurlyeq_l LS(T)^{k_l}) \wedge (b_{lt} \subseteq LT(T)^{k_{lt}}) \wedge (b_v \subseteq V(T)^{k_v})$
Location Sequence Attack with Permanent, Long-Term and Volatile Aspects	$(b_l \preccurlyeq_l LS(T)^{k_l}) \wedge (b_p \subseteq P(PA_u)^{k_p}) \wedge (b_{lt} \subseteq LT(T)^{k_{lt}}) \wedge (b_v \subseteq V(T)^{k_v})$
Visit Attack with Permanent Aspects	$(b_{lv} \preccurlyeq_{lv} LV(T)^{k_{lv}}) \wedge (b_p \subseteq P(PA_u)^{k_p})$
Visit Attack with Long-Term Aspects	$b_{lv} \preccurlyeq_{lv} LV(T)^{k_{lv}} \wedge b_{lt} \subseteq LT(T)^{k_{lt}}$
Visit Attack with Volatile Aspects	$b_{lv} \preccurlyeq_{lv} LV(T)^{k_{lv}} \wedge b_v \subseteq V(T)^{k_v}$
Visit Attack with Permanent and Long-Term Aspects	$(b_{lv} \preccurlyeq_{lv} LV(T)^{k_{lv}}) \wedge (b_p \subseteq P(PA_u)^{k_p}) \wedge (b_{lt} \subseteq LT(T)^{k_{lt}})$
Visit Attack with Permanent and Volatile Aspects	$(b_{lv} \preccurlyeq_{lv} LV(T)^{k_{lv}}) \wedge (b_p \subseteq P(PA_u)^{k_p}) \wedge (b_v \subseteq V(T)^{k_v})$
Visit Attack with Long-Term and Volatile Aspects	$(b_{lv} \preccurlyeq_{lv} LV(T)^{k_{lv}}) \wedge (b_{lt} \subseteq LT(T)^{k_{lt}}) \wedge (b_v \subseteq V(T)^{k_v})$
Visit Attack with Permanent, Long-Term and Volatile Aspects	$(b_{lv} \preccurlyeq_{lv} LV(T)^{k_{lv}}) \wedge (b_p \subseteq P(PA_u)^{k_p}) \wedge (b_{lt} \subseteq LT(T)^{k_{lt}}) \wedge (b_v \subseteq V(T)^{k_v})$

method is not limited to trajectory data and can be applied to any aspect.

Definition 16 represents a data structure denoted as $DS_{aspects}$ containing n records $(A_1, A_2, \ldots, A_n)$. A set of aspects represented each record A_i in the dataset. For example, if a dataset follows a multiple-aspects trajectory data model, A_i contains the smallest granularity, the coordinates and timestamp, and other related aspects. In the given example, these aspects are (x, y, t, PA, LA, VA). It is crucial to emphasize that this method is not limited to trajectory data; it can be applied to any set of aspects.

In simpler terms, it outlines the structure of the dataset. It consists of individual records, each characterized by a set of aspects. The specific aspects mentioned in the example are related to a multiple-aspects trajectory data model and include information such as coordinates (x, y), time (t), and attributes such as permanent aspects (PA), long-term aspect (LA), and volatile aspects (VA).

Definition 16 (Data Structure): Consider a dataset $DS_{aspects}$ comprising n records, denoted as $DS_{aspects} = \{A_1, A_2, \ldots, A_n\}$. A set of aspects characterizes each record A_i in this dataset. For example, in the context of a multiple-aspects trajectory data model, a record A_i includes the minimum granularity aspects such as coordinates (x, y), timestamp t, and additional attributes like permanent aspects (PA), long-term aspects (LA), and volatile aspects (VA).

Definition 17 (Aspect Subset): An aspect subset sb_i is associated with a specific record A_i, where $sb_i \subseteq A_i$.

In Definitions 16 and 17, A_i represents the set of aspects. We define $\mathcal{P}(A_i)$ as the power set of A_i, which is the set of all possible subsets of A_i. Each subset $(sb_1, sb_2, \ldots, sb_m) \in \mathcal{P}(A_i)$ represents an aspect set. Thus, sb_x is the x-th possible subset generated from A_i and is considered as a candidate MSU.

Minimal Sample Unique (MSU) refers to a set of attributes associated with a specific entity or record in a dataset that possesses two key properties:

1) **Uniqueness**: The subset sb_x must uniquely identify the entity among all the other entities in the dataset. For an attribute set A_i, a subset $sb_x \subseteq A_i$ satisfies:

$$\forall j \neq i,\ sb_x \nsubseteq A_j$$

2) **Subset Uniqueness**: The subset sb_x does not contain any proper subset that uniquely identifies the entity A_i. Formally, for all proper subsets $B \subset sb_x$, it holds that:

$$\exists j \neq i \text{ such that } B \subseteq A_j$$

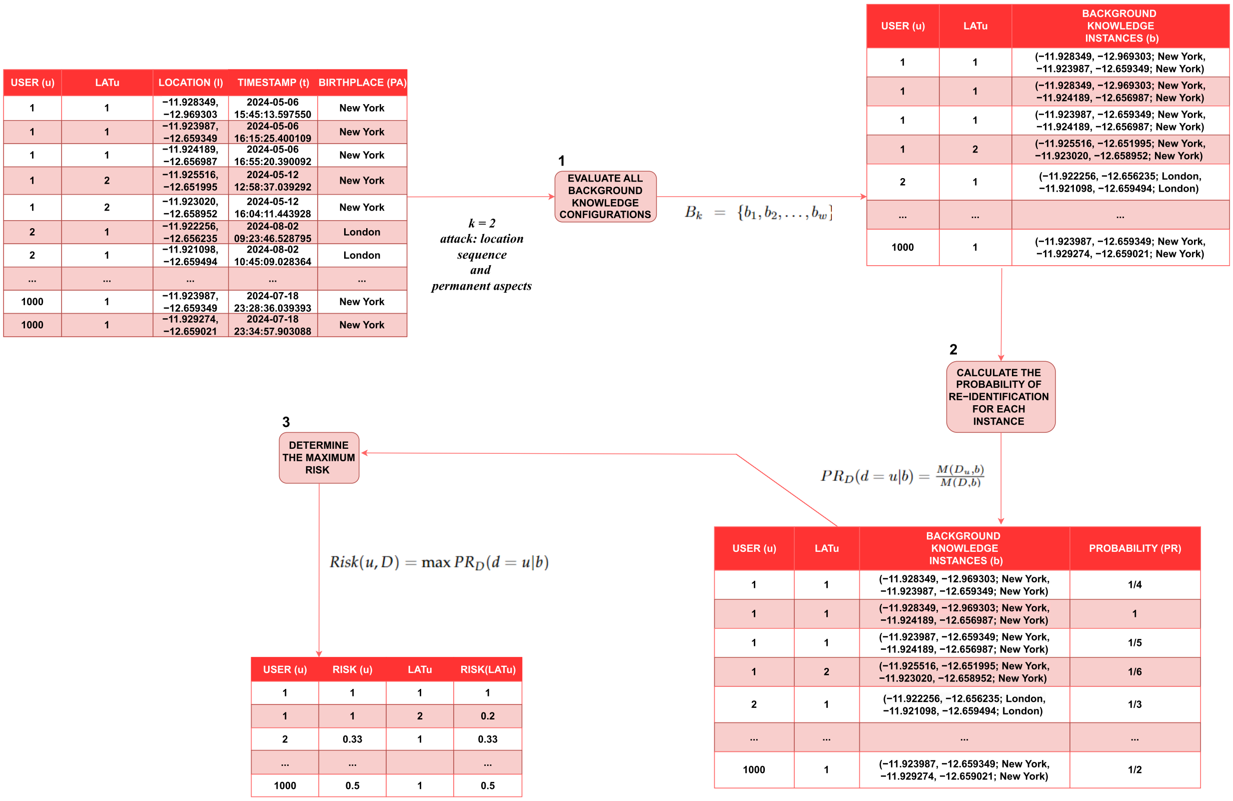
ensuring that every proper subset B of the subset sb_x is contained in at least one other attribute set A_j of a different entity R_j.

In summary, an MSU set represents the minimal combination of attributes required to distinguish a specific entity from all other entities in the dataset. It also has the additional property that no subset of this set can uniquely identify an entity.

Definition 18 (Contribution Calculation): Let $C(a)$ be the contribution of attribute a to re-identification. Calculate $C(a)$ as follows:

$$C(a) = \sum_{x=1}^{n} \begin{cases} 1 & \text{if } a \in sb_x \text{ and } sb_x \text{ is an MSU} \\ 0 & \text{otherwise} \end{cases}$$

1) **Identification of Aspect Sets:**
   - For each record R_i in the dataset D_{aspect}, identify all possible aspect subsets $\mathcal{P}(A_i)$.
2) **Evaluation of MSUs:**
   - For each aspect set sb_x, determine whether it is a Minimal Sample Unique (MSU). A set sb_x is an MSU if:


1
EVALUATE ALL BACKGROUND KNOWLEDGE CONFIGURATIONS
k = 2
attack: location sequence and permanent aspects
$B_k = \{b_1, b_2, \ldots, b_w\}$
2
CALCULATE THE PROBABILITY OF RE−IDENTIFICATION FOR EACH INSTANCE
$PR_D(d = u|b) = \frac{M(D_u,b)}{M(D,b)}$
3
DETERMINE THE MAXIMUM RISK
$Risk(u, D) = \max PR_D(d = u|b)$

USER (u)	LATu	LOCATION (l)	TIMESTAMP (t)	BIRTHPLACE (PA)
1	1	−11.928349, −12.969303	2024-05-06 15:45:13.597550	New York
1	1	−11.923987, −12.659349	2024-05-06 16:15:25.400109	New York
1	1	−11.924189, −12.656987	2024-05-06 16:55:20.390092	New York
1	2	−11.925516, −12.651995	2024-05-12 12:58:37.039292	New York
1	2	−11.923020, −12.658952	2024-05-12 16:04:11.443928	New York
2	1	−11.922256, −12.656235	2024-08-02 09:23:46.528795	London
2	1	−11.921098, −12.659494	2024-08-02 10:45:09.028364	London
...	...	...	...	...
1000	1	−11.923987, −12.659349	2024-07-18 23:28:36.039393	New York
1000	1	−11.929274, −12.659021	2024-07-18 23:34:57.903088	New York

USER (u)	LATu	BACKGROUND KNOWLEDGE INSTANCES (b)
1	1	(−11.928349, −12.969303; New York, −11.923987, −12.659349; New York)
1	1	(−11.928349, −12.969303; New York, −11.924189, −12.656987; New York)
1	1	(−11.923987, −12.659349; New York, −11.924189, −12.656987; New York)
1	2	(−11.925516, −12.651995; New York, −11.923020, −12.658952; New York)
2	1	(−11.922256, −12.656235; London, −11.921098, −12.659494; London)
...	...	...
1000	1	(−11.923987, −12.659349; New York, −11.929274, −12.659021; New York)

USER (u)	LATu	BACKGROUND KNOWLEDGE INSTANCES (b)	PROBABILITY (PR)
1	1	(−11.928349, −12.969303; New York, −11.923987, −12.659349; New York)	1/4
1	1	(−11.928349, −12.969303; New York, −11.924189, −12.656987; New York)	1
1	1	(−11.923987, −12.659349; New York, −11.924189, −12.656987; New York)	1/5
1	2	(−11.925516, −12.651995; New York, −11.923020, −12.658952; New York)	1/6
2	1	(−11.922256, −12.656235; London, −11.921098, −12.659494; London)	1/3
...	...	...	...
1000	1	(−11.923987, −12.659349; New York, −11.929274, −12.659021; New York)	1/2

USER (u)	RISK (u)	LATu	RISK(LATu)
1	1	1	1
1	1	2	0.2
2	0.33	1	0.33
...	...		...
1000	0.5	1	0.5

FIGURE 2. **TrajectGuard framework data-flow.**

- sb_x is unique in the sample.
- None of its subsets is unique in the sample.

3) **Contribution Calculation:**

- The contribution of each aspect in the dataset is calculated by counting the number of MSUs it is part of.

Drawing parallel with privacy risk assessment, identifying all possible aspect subsets is similar to calculating all background knowledge configurations. Therefore, MSUs are unique background knowledge configurations in which none of their aspects are unique.

2) AspectGuard RISK ASSESSMENT

In the *AspectGuard* Risk Assessment, we evaluate the re-identification risk by considering the risk contribution of aspects at the dataset level. We can understand the impact of each aspect and its relation to re-identification risk.

We explore the core contribution of *AspectGuard*, where we quantify the distinct impact of each aspect or combination of aspects on the risk of re-identification. The aspect risk is calculated by considering the number of Minimal Sample Uniques per aspect combination and dividing it by the total number of Minimal Sample Uniques. The formula for aspects risk is as follows:

$$\text{Risk}_{\text{aspect}}(a) = \frac{C(a)}{\text{Total MSUs}}, \tag{8}$$

where (a) represents the aspect(s) under consideration. This metric provides insights into the overall risks associated with the different aspects of the dataset.

These risk assessments form a crucial part of our approach, providing a comprehensive understanding of the re-identification risks related to the aspects risk contribution at both individual and dataset levels.

D. *AnonimoGuard*

In this subsection, we introduce the last contribution: *AnonimoGuard*, which measures the privacy risk for anonymized datasets. *AnonimoGuard Risk Assessment* acknowledges the particularities of evaluating privacy risk in anonymized mobility datasets. The main factor is a detailed comparison between the background knowledge configurations between the original and anonymized datasets. Here, we use the background knowledge instances from the original dataset as the adversary's knowledge, ensuring a realistic perspective of the re-identification risk.

In simpler terms, *AnonimoGuard* ensures fairness by considering the background knowledge instances from the original and anonymized datasets as the foundation for evaluating re-identification risk. For instance, AnonimoGuard could be used to assess the privacy risk of an anonymized trajectory dataset, where multiple aspects such as location, time, personal information, and frequency of visits are recorded. It assesses whether combining these aspects might allow an attacker to re-identify individuals based on their movement patterns. By performing this evaluation, AnonimoGuard ensures that the dataset is genuinely anonymized, confirming that re-identification is impossible even when considering multiple trajectory-related attributes.

1) PRIVACY RISK CALCULATION

In the *AnonimoGuard Risk Assessment* process, background knowledge configurations are identified in both the original dataset D and anonymized dataset D' using Definition 12. Find the intersection of background knowledge configurations between datasets, representing the shared knowledge available to the adversaries. The individual privacy risk is calculated based on this intersection set, considering the maximum ratios of matching records if the set is not empty, and the risk is set to zero if the set is empty. Next, we describe these steps and definitions.

1) **Adversary Background Knowledge**: Identify the background knowledge configurations in the original dataset D using Definition 12. To calculate the background knowledge configurations in the original dataset for each individual's trajectory data identified in the original dataset D, we identify the set of background knowledge configurations B_k^u.
2) **Background Knowledge Configurations After Anonimization**: Identify the background knowledge configurations in the anonymized dataset D'.
   Definition 19 (Background Knowledge Configurations in Anonymized Dataset D'): Similarly, for each individual's trajectory data identified in anonymized dataset D', we identify the set of background knowledge configurations $B_k^{u'}$.
3) **Background Knowledge Intersection**: Find the intersection set between the background knowledge configurations in the original and anonymized datasets ($B_k^u \cap B_k^{u'}$). This set represents the common knowledge available to the adversary in both datasets, as defined in Definition 20.
   Definition 20 (Intersection of Background Knowledge Configurations): For each u, find the intersection set $B_k^u \cap B_k^{u'}$ between the background knowledge configurations in the original and anonymized datasets.
4) **Anonymity Risk Assessment** The privacy risk for each trajectory is calculated based on the intersection set. If the intersection set is not empty, the risk is calculated as the maximum ratio of matching records. If the intersection set is empty, the risk is zero, indicating no common background knowledge and no re-identification risk.
   a) Calculate Background Knowledge Configurations in Original Dataset D:
      For each u in the original dataset D, identify the background knowledge configurations B_k^u using Definition 12.
   b) Privacy Risk Calculation:
      The anonymity risk of re-identification (or anonymity privacy risk), as described in Definition 21, for an individual, denoted as u, is determined by considering the intersection set of background knowledge configurations in both the original and anonymized datasets, represented as $B_k^u \cap B_k^{u'}$. If this intersection set is not empty, the risk is calculated as the maximum probability of re-identification for each background knowledge configuration b' within the intersection set, i.e., $\max\left(PR_{D'}(d = u \mid b')\right)$.
      However, the risk is set to zero if the intersection set is empty, indicating no common background knowledge between the original and anonymized datasets.
      Definition 21 (Risk of Re-Identification or Privacy Risk in AnonimoGuard): The risk of re-identification (or privacy risk) of an individual u given an intersection set of background knowledge configurations $B_k^u \cap B_k^{u'}$ is defined as the maximum probability of re-identification:

$$Risk(u, D') = \begin{cases} \max\left(PR_{D'}(d = u \mid b')\right), & \forall b' \in (B_k^u \cap B_k^{u'}) \\ 0, & B_k^u \cap B_k^{u'} = \emptyset \end{cases}$$

V. EXPERIMENTS

This section presents the datasets and experiments conducted using the **TrajectGuard** framework.[4] The individual will be named as a user due to the dataset's characteristics. We evaluated the privacy risk, aspect risk (*AspectGuard*), and risk after applying data protection techniques (*AnonimoGuard*). Finally, each subsection concludes with a discussion of the results.

A. MULTIPLE-ASPECTS TRAJECTORY DATASETS

This subsection introduces the datasets used in our study: Foursquare and the U.S. Census Bureau, Breadcrumb, and WiFi UFSC. These datasets were categorized as public, semi-private, and private, respectively. All datasets were preprocessed using the scikit-mobility Python library [44].

1) FOURSQUARE AND THE U.S. CENSUS BUREAU

Our Foursquare data set comprises check-ins in NYC collected from April 12, 2012, to February 16, 2013, for almost ten months. It contains 227,428 check-ins. Each check-in is associated with a user's ID, timestamp in minutes, GPS coordinates (latitude and longitude), and semantic meaning. Venue categories from Foursquare characterize it. This data set was authored by [45]. The data were compressed to a radius of 100 m. To create a multiple-aspects trajectory data set, we included information on users using NYC U.S. Census Bureau information. The NYC U.S. Census provides

[4] https://github.com/oliveiragomesphd/TrajectGuard/

population information, such as gender and race, represented by percentage values per census tract. The census tracts were territorial areas established by the Bureau of the Census for population analysis. We use the latitude and longitude from the foursquare data set to obtain census tracts and census tracts to obtain the Bureau of Census information.

Data Information: The aspects related to the location and moving object, other than spatiotemporal coordinates and time, are:

- Volatile Aspects: Semantic location (venue category) and temperature.
- Long-term aspect: Trajectory size, weekday.
- Permanent aspect: Gender, employed, citizen, race.

2) BREADCRUMBS

The Breadcrumbs dataset [46] was created using the data obtained during a campaign conducted in Lausanne in the spring of 2018. Eighty participants were recruited through a specialized Labex unit at the University of Lausanne. Access to the dataset was granted through a secure data-sharing license. For our analysis, we specifically utilized the GPS data.

Data Information: The aspects related to the location and moving object, other than spatiotemporal coordinates and time, are:

- Volatile Aspects: Location semantic (street) and temperature
- Long-term aspect: Weekday and trajectory size.
- Permanent aspect: Gender, age, job, family status, nationality, and exercise.

3) WIFI

The Wi-Fi UFSC dataset is sourced from user device associations with wireless access points within the university's wireless network. The devices to be connected must be within a specified radio range to receive signals from these access points. Each access point is associated with its geographic coordinates, indicating its installation location. The dataset used in the experiment captured a single three-day log of 14,360 undergraduate students.

Data Information: The aspects related to the location and moving object, other than spatiotemporal coordinates and time, are:

- Volatile Aspects: Semantic location and temperature
- Long-term aspect: Weekday and trajectory size.
- Permanent aspect: Age, course, and gender.

When we refer to *permanent aspects*, in all datasets, we highlight the static nature of these characteristics within the user trajectories. Although attributes such as gender, age, job, family status, nationality, and exercise habits may not be permanent aspects of an individual's life, they exhibit permanence within the trajectory records. These features remained constant throughout the recorded trajectories, representing stable and unchanged attributes during the analysis.

B. PRIVACY RISK ASSESSMENT

This subsection assesses the privacy risk of our previously defined attacks on real-life mobility datasets with multiple-aspects, permanent, long-term, and volatile. We applied the attacks to the three datasets introduced earlier and quantified the privacy risk using different background knowledge. The main goal of these experiments was to understand the impact of the multiple-aspects of privacy risk in other datasets to guide the data provider to interpret the risk values when using the *TrajectGuard framework*.

The experiments were conducted on a machine with 16 vCPUs and 128 GB of RAM. We conducted tests on location and time information using different formats: raw or generalized values, with or without geographical coordinate knowledge. The legends in the graph are abbreviated for simplicity and are listed in Table 5. The chart legends contain abbreviations of the aspects used in the attack; legend can be checked in Table 6.

TABLE 5. Graphs - aspects abbreviation.

Aspect	Abbreviation
latitude	lat
longitude	long
location semantic	s
time	t
temperature	tmp
permanent aspects	PA
long-term aspects	LA
volatile aspects	VA

To comprehensively assess associated privacy risks, we performed thorough testing by employing diverse attacks on three distinct datasets. As shown in Table 7, for permanent and long-term aspect attacks, the x-axis contains the risk percentage, and the y-axis represents the number of users. We evaluated the risk of volatile, location sequence, visit aspects attacks, and their multiple aspect attack variations by assessing the user's data risk for each user's trajectory. The x-axis indicates the risk percentage, and the y-axis represents the percentage of users.

1) WI-FI

The graph in Figure 3 presents the risk evaluation with the *Permanent Aspects Attack*: *Age*, *Course*, and *Gender*. Parameter k represents the number of permanent aspects the attacker knows regarding the user's data. The figure illustrates a direct relationship wherein an increase in knowledge size corresponds to an increase in risk values. When an attacker has more knowledge of aspects related to the victim, it isn't easy to find many users sharing the same characteristics. However, this behavior is expected. What was possible to note as a characteristic of this dataset is that most users were evaluated as having less than 50% risk. This implies that most data are not unique, indicating that users share the same permanent aspect values. Of the approximately 14,360 users, only approximately 600 have a maximum risk value

TABLE 6. Graphs - aspects abbreviation.

Graph Legend	Attack
lat, long	Location Sequence
lat, long, t	Visit
lat, long, PA	Location Sequence and Permanent Aspects
lat, long, LA	Location Sequence and Long-Term Aspects
lat, long, VA	Location Sequence and Volatile Aspects
lat, long, PA, LA	Location Sequence, Permanent and Long-Term Aspects
lat, long, LA, VA	Location Sequence, Long-Term and Volatile Aspects
lat, long, PA, VA	Location Sequence, Permanent and Volatile Aspects
lat, long, PA, LA, VA	Location Sequence, Permanent, Long-Term and Volatile Aspects
lat, long, t, PA	Visit and Permanent Aspects
lat, long, t, LA	Visit and Long-Term Aspects
lat, long, t, VA	Visit and Volatile Aspects
lat, long, t, PA, LA	Visit, Permanent and Long-Term Aspects
lat, long, t, LA, VA	Visit, Long-Term and Volatile Aspects
lat, long, t, PA, VA	Visit, Permanent and Volatile Aspects
lat, long, t, PA, LA, VA	Visit, Permanent, Long-Term and Volatile Aspects

TABLE 7. Charts axis.

Attack	X-axis	Y-axis
Permanent and Long-Term Attacks	Risk (%)	Quantity Users
Other Attacks	Risk (%)	Users's Trajectories

with a knowledge size of two and more than a thousand with a knowledge size of three. This is approximately 4% and 7%, respectively, which are small values. The maximum risk value implies the user has at least one unique background knowledge instance. The absence of this unique characteristic in the Wi-Fi dataset suggests a certain level of data protection for most users when only permanent aspects are analyzed.

Figure 4 presents the risk assessment for *Long-Term Aspects Attack*. The long-term aspects are: *Weekday*, *Trajectory Size* and *Trajectory Size and Weekday*. With a knowledge size equal to one, the risk values were observed to be closer to zero across all aspects. Notably, the *Weekday* aspect exhibited lower risk values because of the dataset's limited number of days (only three). Consequently, many trajectories share the same weekday, reducing the risk of re-identification. Conversely, the *Trajectory Size* and *Trajectory Size and Weekday* aspects tend to have risk values closer to zero, which shows that many trajectories share the same values, reducing the risk. In the Wi-Fi dataset, the risk values were low when analyzing only the long-term aspects. This means that it is difficult to re-identify users using them.

Figure 5, representing the *Long-Term Aspects Attack* with knowledge size 2, illustrates a significant increase in risk for *Trajectory Size* and *Trajectory Size and Weekday* aspects. At the same time *Weekday* consistently retained low-risk values. It indicates that the increase in knowledge size does not bring high risk if the aspect is "Weekday." However, it adds risk when the attacker knows aspects include *Trajectory Size* or *Trajectory Size and Weekday*. This indicates that "Size" has more unique values, and its combination with "Weekday" makes it even more unique.

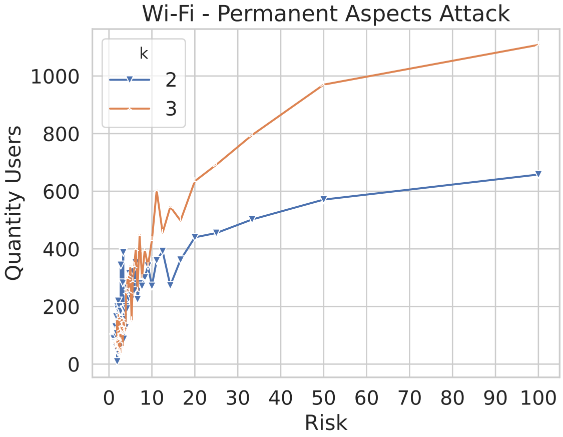


FIGURE 3. Wi-Fi - permanent aspects attack.

Approximately 15% of the users had a unique "Trajectory Size and Weekday" combination.

Regarding the *Volatile Aspects Attack*, in the Wi-Fi dataset, we consider *Location Semantic* and *Temperature* as volatile aspects. We tested the attack using distinct possible knowledge: *Location Semantic* and *Temperature*, *Location Semantic* only, and *Temperature* only. Figure 7 shows results for a knowledge size of one. Notably, the majority of the values were below the 50% risk threshold. This suggests a lower contribution to re-identification from the volatile aspects in this specific dataset when the knowledge size is small. We consider a small area covered by the dataset: a small campus. Individuals are more likely to connect to the same location in a confined space. Additionally, the *Temperature* was expected to remain constant across all locations, varying only with time.

In Figure 8, where size knowledge equals 2, the behavior shifts, with the distribution of values for both *Location Semantic* and only *Temperature* aspects trending towards higher risk values. Additionally, trajectories involving both

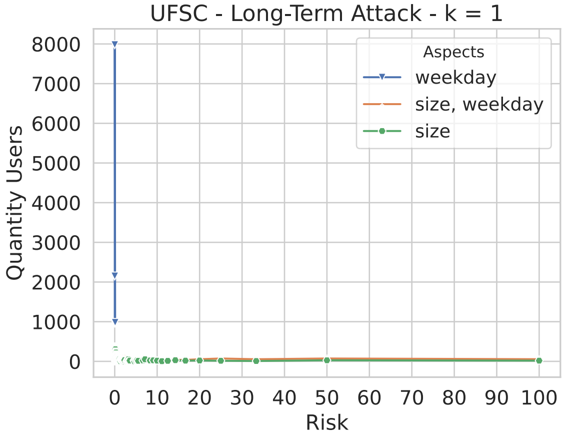


FIGURE 4. Wi-Fi - long-term aspects attack - $k = 1$.

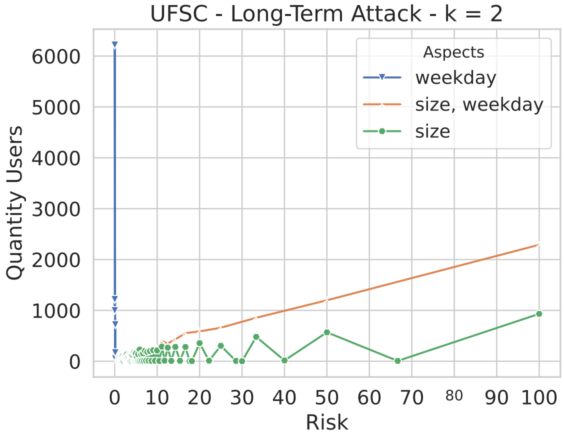


FIGURE 5. Wi-Fi - long-term aspects attack - $k = 2$.

Location Semantic and *Temperature* aspects consistently moved towards a risk value close to 100% for most instances. As knowledge size increased, there was a tendency for risk values to approach maximum risk. However, it is important to note that in this case, the *Temperature* is more correlated with time than space, as it encompasses a small area. Although we did not observe much risk when the knowledge size was one, when we increased the risk size, it became more difficult to have the same knowledge of temperature values across several trajectories, as it depended on them being present at the same time or at a time with the same *Temperature*.

Concerning the *Location Sequence Attacks* with all multiple-aspects variations, the observed behavior remains consistent in Figures 8 and 9, with risk values trending towards maximum risk as the k value increases. Additionally, it is notable that including more aspects of the attack leads to higher risk values. Notably, the location information has very small maximum risk values when the knowledge size is one. However, when more aspects are added, the risk values increase. By comparing the types of aspects, the combination of location and permanent aspects poses the highest risk,

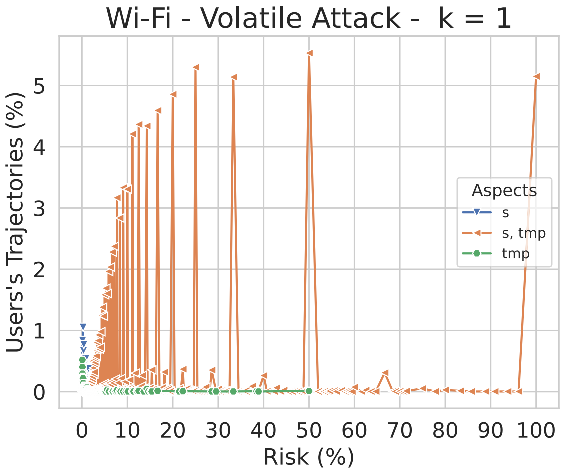


FIGURE 6. Wi-Fi - volatile attack - k = 1, legend in Table 5.

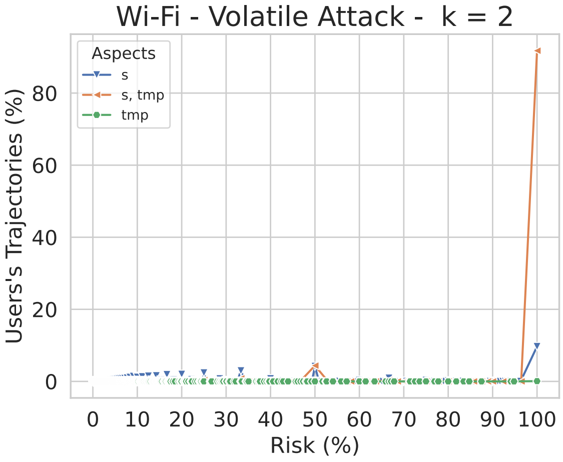


FIGURE 7. Wi-Fi - volatile attack - k = 2. Legend in Table 5.

followed by location and long-term aspects, and finally, location with volatile aspects. The same pattern was observed for a knowledge size of two. This type of evaluation helps the data provider to understand the dimensions that have the greatest impact on re-identification risk, highlighting areas requiring more attention regarding data protection.

Regarding *Visit attacks*, consistent results are observed in Figures 10 and 11. All attacks exhibited very high high-risk values, indicating that the location and time information remained highly unique even with the introduction of multiple-aspects. Based on the *Location Sequence Attacks* charts in Figures 8 and 9, when considering only the location, some lower risk values are still present. However, this risk is considerably elevated when combined with time. The uniqueness of this dataset arises from the locations representing access points where multiple individuals can connect simultaneously. However, these behaviors are distinguished by connections to the same access point at the same time intervals, resulting in fewer concurrent connections.

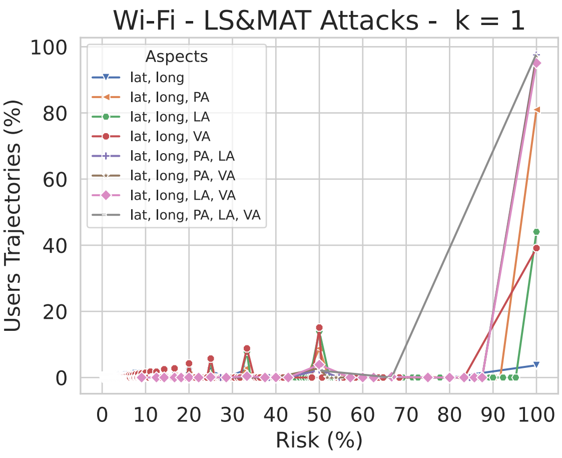


FIGURE 8. Wi-Fi - LS&MAT attacks - $k = 1$. Legend in Table 6.

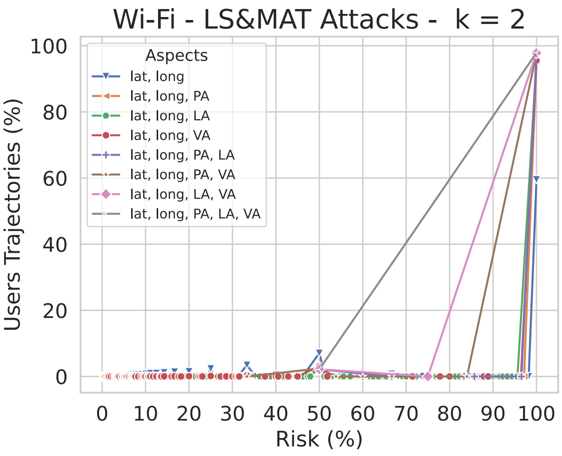


FIGURE 9. Wi-Fi - LS&MAT attacks - $k = 2$. Legend in Table 6.

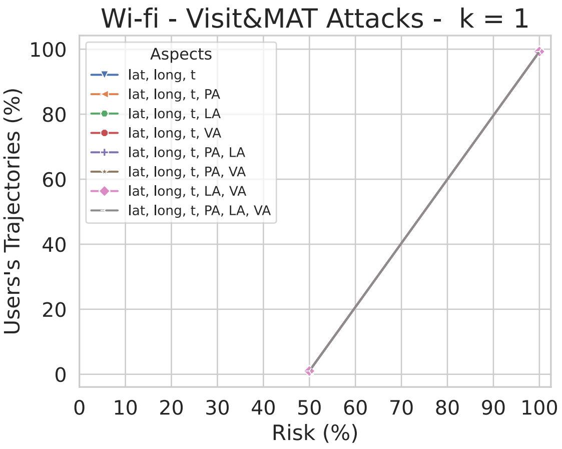


FIGURE 10. Wi-Fi - Visit&MAT attacks - $k = 1$. Legend in Table 6.

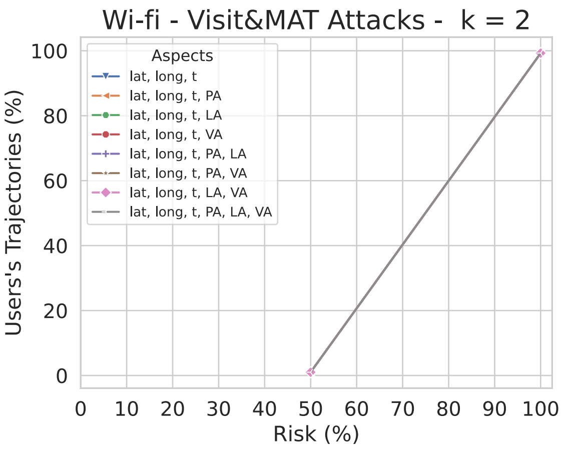


FIGURE 11. Wi-Fi - Visit&MAT attacks - $k = 2$. Legend in Table 6.

2) BREADCRUMBS

The graphs in Figure 12 present the risk evaluation with the *Permanent Aspects Attack* used are: *Gender*, *Age*, *Job*, *Family Status*, *Nationality*, and *Exercise*. The chart in Figure 12 illustrates a direct relationship wherein an increase in the knowledge size corresponds to a rise in the risk values. Like the Wi-Fi dataset, most of the risk values for the 78 users are not unique until the knowledge size equals 5. This implies that the attacker needs to have a lot of information about the user in order to have a chance of identifying them. By analyzing only the permanent aspects, most users have a certain level of data protection due to the lack of uniqueness in the initial knowledge sizes. However, as the knowledge size grows, more users are re-identified due to the increased uniqueness.

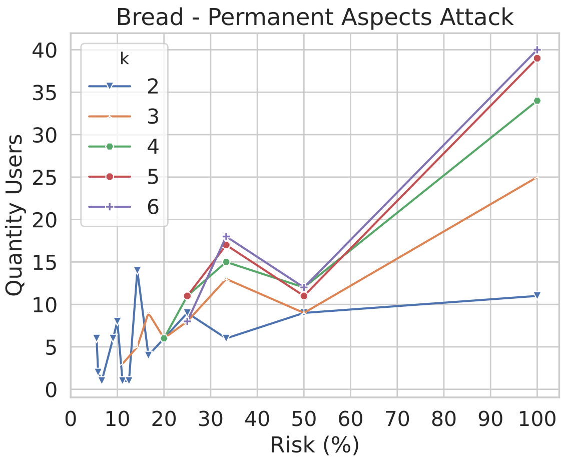


FIGURE 12. Breadcrumbs - permanent aspects attack.

Figure 13 presents the risk assessment for the *Long-Term Aspects Attack*. The long-term aspects are: *Weekday*, *Trajectory Size* and *Trajectory Size and Weekday*. The *Weekday* aspect is less risky than the *Trajectory Size* aspect. However, when both aspects were combined, a significant increase in the risk was observed. This highlights the impact of multiple aspect on risk values. As the value of k increases to two in Figure 14, we observe that the risk values approach 100%, for all aspect sets.

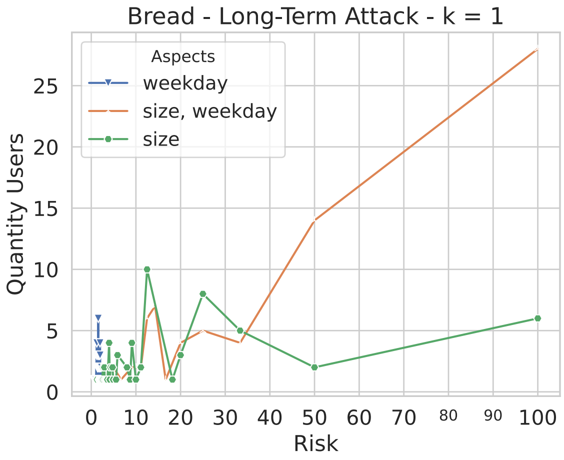


FIGURE 13. Breadcrumbs - long-term attack - $k = 1$.

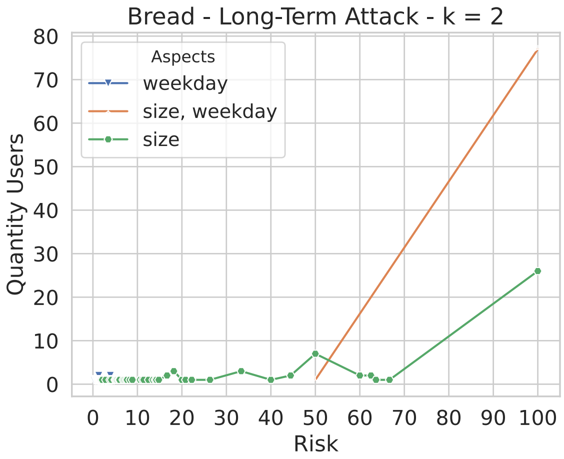


FIGURE 14. Breadcrumbs - long-term attack - $k = 2$.

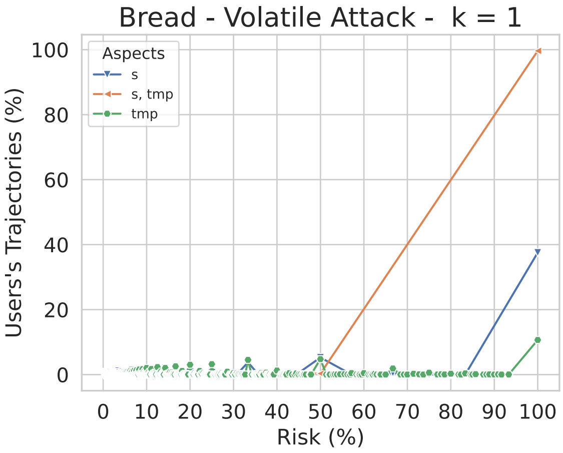


FIGURE 15. Breadcrumbs - volatile attack - $k = 1$, legend in Table 5.

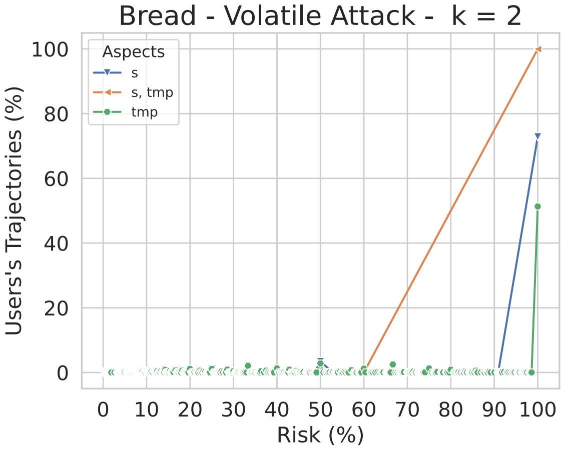


FIGURE 16. Breadcrumbs - volatile attack - $k = 2$. Legend in Table 5.

In the Breadcrumbs dataset, where *Location Semantic* and *Temperature* are considered volatile aspects, we conducted attacks using distinct knowledge combinations: *Location Semantic* and *Temperature*, only *Location Semantic*, and only *Temperature*. Figure 15 exhibits a distinct pattern compared to the results of the Wi-Fi *Volatile Aspects Attack*. Notably, there was a significant increase in high-risk values, with a concentration above the 50% risk threshold. This suggests a higher contribution to re-identification from the volatile aspects in this specific dataset. In Figure 16, where the knowledge size is equal to two, the behavior follows the same direction, with the distribution of values for both, only *Location Semantic* or *Temperature* aspects trending towards higher risk values. Additionally, trajectories involving both semantic and temperature aspects consistently moved towards a risk value close to the maximum value in most instances. This means that both aspects *Location Semantic* and *Temperature* have unique values that can contribute to the re-identification risk. We can also see that multiple-aspects contribute to increasing risk value.

Regarding the *Location Sequence Attacks* and *Visit Attacks*, with knowledge sizes one and two, all attacks exhibit a maximum risk value of 100%, indicating that even with the introduction of multiple-aspects, the location values were so unique that their addition did not affect the risk value.

3) FOURSQUARE

The graphs in Figure 17 present the risk evaluation with the *Permanent Aspects Attack*: *Gender*, *Employed*, *Citizen*, and *Race*. Although the graph shows a similar behavior of most values not being unique, unlike the other datasets, the unique values are minimal, representing no more than 2% with a knowledge size of 5. This can be explained by the fact that permanent aspects were generated from aggregated data.

Figures 18 and 19 show the risk assessment for the *Long-Term Aspects Attack*. The long-term aspects are: *Weekday*, *Trajectory Size* and *Trajectory Size and Weekday*. When analyzing the *Long-Term Aspects Attack*, compared to the other two datasets, Foursquare exhibits a distinct behavior. *Trajectory Size* and *Weekday* knowledge showed

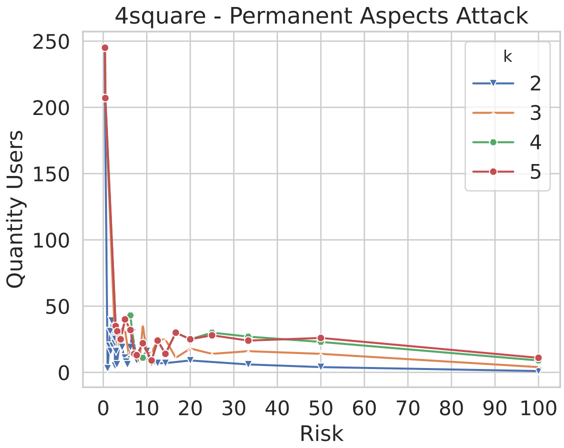


FIGURE 17. Foursquare - permanent aspects attack.

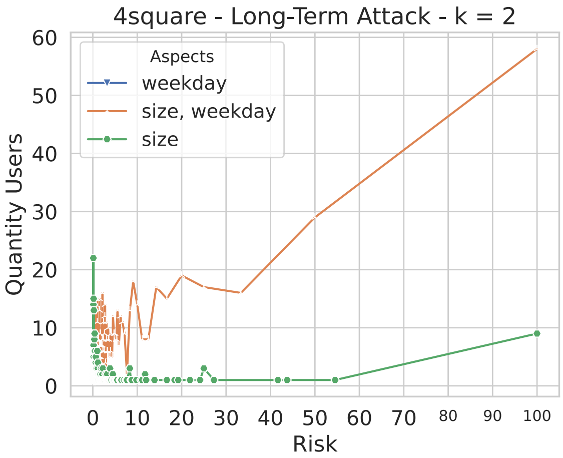


FIGURE 19. Foursquare - long-term attack - $k = 2$.

very low-risk values. This behavior was observed in the Foursquare dataset because of its low density, indicating a small number of daily check-ins and a large collection period (from April 12, 2012, to February 16, 2013). The *Trajectory Size and Weekday* combination followed a similar pattern. As knowledge size increased, there was a trend towards higher risk values approaching 100%. The *Trajectory Size and Weekday* knowledge combination significantly increased risk.

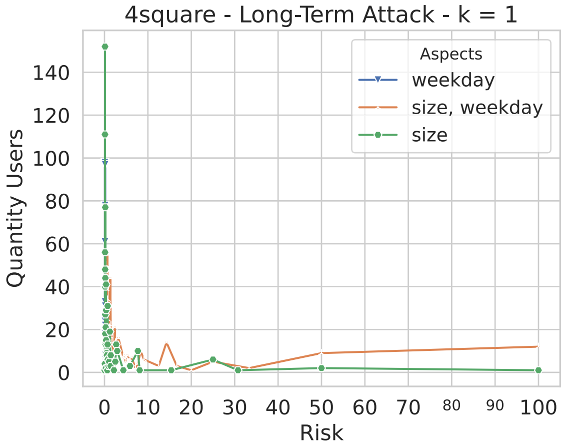


FIGURE 18. Foursquare - long-term attack - $k = 1$.

In the Foursquare dataset, where we examine *Location Semantic* and *Temperature* as volatile aspects, our *Volatile attack* evaluation involves diverse knowledge configurations: *Location Semantic* and *Temperature*, only *Location Semantic*, and only *Temperature*. Figure 20 reveals that the majority of values are concentrated below the 50% risk threshold for only the *Location Semantic* or *Temperature* knowledge. This implies a small contribution to this dataset's re-identification of volatile aspects. However, when both knowledge types were combined, we observed maximum risk for all users.

In Figure 21, when the knowledge size is equal to two, the behavior shifts for *Location Semantic* and *Temperature* knowledge when isolated. The distribution of values for only *Location Semantic* trends towards higher risk values. For *Temperature*, all the user data had a risk value of 100%.

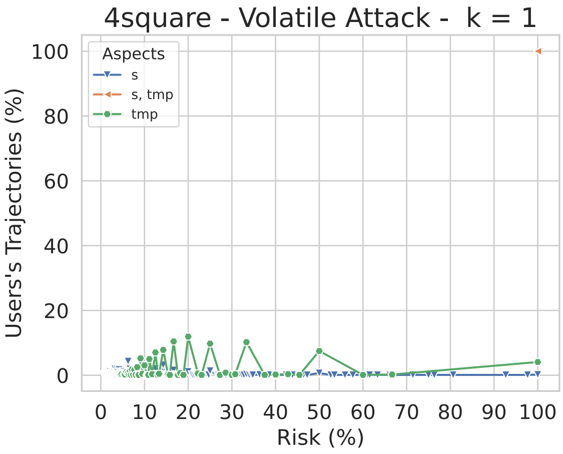


FIGURE 20. Foursquare - volatile attack - $k = 1$. Legend in Table 5.

Concerning the *Location Sequence Attacks* and knowledge size of one, the observed trend of the risk values approaching the maximum value is shown in Figure 22. Notably, all the attacks presented very high-risk values. Only the coordinates were sufficient to yield high-risk values. Very few instances exist in which the risk is not one in the Location Sequence Attack and the Location Sequence Attack with Long-Term Aspects. In Location Sequence Attacks with knowledge size two, all attacks yielded the maximum risk for all user trajectories.

Regarding Visit attacks, all attacks exhibited risk values equal to 100%. This indicates that because the location and time information are sufficiently unique, multiple-aspects do not contribute significantly to the increase in risk values.

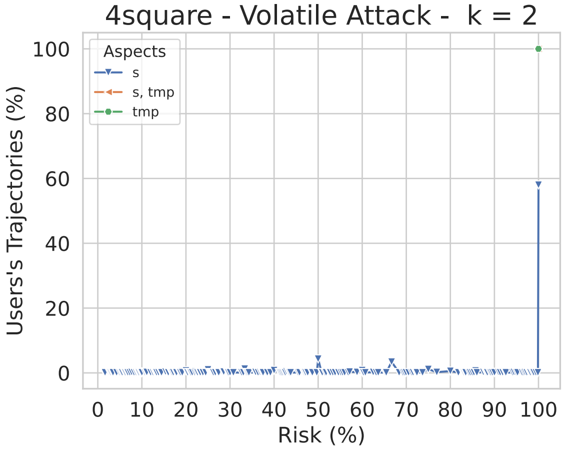


FIGURE 21. Foursquare - volatile attack - $k = 2$. Legend in Table 5.

Next, we summarize and discuss the most important findings.

4) RESULTS HIGHLIGHTS AND DISCUSSION

The experiments were conducted on three distinct datasets using the data model and attacks delineated earlier in this work, including Wi-Fi, Breadcrumbs, and Foursquare, each characterized by distinct aspects such as location semantics, trajectory size, weekdays, temperature, and moving object aspects.

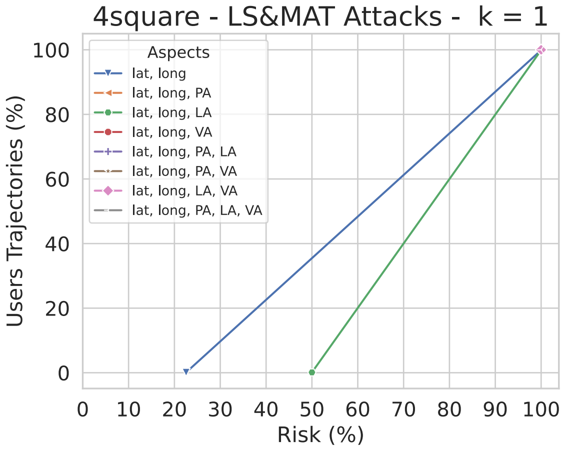


FIGURE 22. Foursquare - LS&MAT attacks - $k = 1$. Legend in Table 6.

The findings indicate that the coverage area, temporal distribution, and aspect values significantly influence the risk of re-identification. While permanent, long-term, and volatile aspects alone may not pose a significant risk to users in some datasets, exploring various attack scenarios involving multiple-aspects reveals that including additional aspects refines data filtering, increasing its uniqueness and facilitating re-identification. However, uniqueness is the primary factor affecting risk values. It can arise from location data alone, a combination of location and time information, or from the inclusion of multiple-aspects. Therefore, the interpretation of the results closely aligns with the dataset characteristics.

The significant impact of time and coordinate values on risk is notable, primarily because of their inherent uniqueness. Despite the distinctive nature of trajectory data, it is evident that the addition of more aspects facilitates re-identification. In particular, aspect knowledge is necessary to enable a successful attack if an attacker lacks a unique location or location-time knowledge source.

The study's findings offer valuable insights for data providers to protect trajectory data from multiple-aspects. By uncovering how permanent, long-term, and volatile aspects influence re-identification risks across different datasets, this research provides practical guidance for enhancing data protection strategies in trajectory data management and analysis.

The next subsection will present the experiments with *AspectGuard*.

C. *AspectGuard*

In this subsection, we present the results of implementing *AspectGuard* on the three distinct datasets. Our evaluation encompasses both user and aspect risks and comprehensively assesses the framework's effectiveness.

1) RESULTS

We examined how different aspects affect the risk of re-identification and exploration of the nuanced dynamics in each dataset. Our analysis offers valuable insights into how specific attributes impact privacy risk and how combining multiple-aspects shapes the overall risk of re-identification. A thorough exploration of *AspectGuard*'s findings provides a nuanced understanding of privacy risk in multiple-aspects trajectory data for data providers.

2) ASPECTS RISK

In the aspect risk evaluation, we thoroughly explored aspects-based risks within three distinct datasets: Breadcrumbs, Wi-Fi, and Foursquare. Each dataset presents unique characteristics and poses different challenges regarding re-identification risk assessment. The percentage values in Tables 9, 8, and 10 represent the percentage of minimal sample unique sets containing that aspect or a combination of aspects. The frequency of occurrence of an aspect or a combination of aspects reveals the uniqueness they bring to the data, directly influencing the risk of re-identification. It is important to highlight that it is not just about how unique this attribute is but how this aspect impacts the overall uniqueness of the background knowledge configurations.

a: Wi-Fi DATASET

Regarding the Wi-Fi dataset delineated in Table 8, a nuanced perspective on re-identification risk has emerged, showcasing the multifaceted interplay of various aspects in shaping privacy vulnerabilities. While temporal factors, as represented by *time*, indeed wield significant influence, accounting for a dominant 54.56% of the overall risk, it is imperative to recognize the substantial contributions of other aspects beyond mere spatial and temporal dimensions.

Indeed, the Wi-Fi dataset reveals that diverse attributes, beyond traditional demographic factors, significantly impact re-identification risk. Notably, the role of location semantic attributes, denoted as *location semantic* made a substantial contribution of 33.36%.

Moreover, spatial attributes, beyond mere location granularity, also exhibit a considerable influence. For instance, the combination of *location* and *size* contributes approximately 40.95%, showing how the spatial scope or scale of activity areas can amplify the re-identification risk. Similarly, environmental factors, such as *temperature*, played a non-trivial role, contributing 29.20% to the overall risk. This highlights the significance of considering environmental context, when assessing privacy vulnerabilities in multiple-aspects of trajectory data. However, with the privacy risk assessment, we should note that *temperature* did not bring a lot of uniqueness by itself. We can see that *temperature* is present in many unique sets.

In Table 8, we observe that demographic attributes such as *age* and *gender* make considerable contributions to re-identification risk, with *age* contributing 27.89% and *gender* contributing 5.24% to the overall risk. These findings underscore the significance of demographic information in shaping privacy risks in trajectory data scenarios. This highlights the importance of considering the contributions of other contextual aspects beyond *location* and *time*.

These findings collectively emphasize the need for a comprehensive privacy protection approach encompassing diverse contextual dimensions, including temporal, spatial, semantic, and environmental factors. Effective privacy preservation strategies can be devised to mitigate the heightened risk posed by multifaceted re-identification attacks in multiple-aspects of trajectory data scenarios.

TABLE 8. Wi-Fi - aspect contribution to the re-identification risk.

Combination	Contribution (%)
age	27.8908
course	34.08765
gender	5.24493
location semantic	33.35922
location	35.1258
time	54.55987
temperature	29.19769
weekday	3.38977
size	40.95193
age, location semantic	9.61978
age, location	10.07527
age, time	8.22016
age, temperature	11.56758
course, location semantic	11.35475
course, temperature	12.07462
¨temperature, size	16.81284

b: BREADCRUMBS DATASET

We present aspect contributions to overall re-identification risk in the Breadcrumbs dataset in Table 9. Notably, temporal information emerged as a dominant factor, with the *time* aspect accounting for a substantial portion of the risk at approximately 54.56%. This underscores the importance of the temporal context in trajectory data, where the timing of location visits significantly affects the risk of re-identification. Moreover, spatial attributes such as *location* make significant contributions, indicating the role of geographic context in privacy risk assessment.

Interestingly, including fine-grained location details, as represented by the *location semantic* aspect, also contributed to the risk. This suggests that even variations in location granularity can affect re-identification vulnerability. The substantial contribution of the *location semantic, temperature* combination underscores the importance of considering multiple-aspects in privacy risk analysis. By recognizing the impact of spatial and environmental attributes, privacy protection strategies can be refined to mitigate the heightened risks posed by such combinations. In the Breadcrumbs dataset, demographic information was provided with a certain degree of generalization, which inherently reduces its impact on re-identification risk compared to attributes such as *time* and *location*.

TABLE 9. Bread - aspect contribution to the re-identification risk.

Combination	Contribution (%)
time	30.12103
location	30.41431
location semantic	1.90685
temperature	0.22153
time, location semantic	0.16654
time, gender	0.14874
time, age	0.16078
time, job	0.14036
location semantic, gender	0.10579
location semantic, temperature	14.36135

c: FOURSQUARE DATASET

We found a nuanced interplay of aspect-based risks in the Foursquare dataset, as shown in Table 10. Here, along with temporal and spatial attributes, we observed significant contributions from factors such as location semantic and environmental conditions (represented by *temperature*). Including location-related information introduces a new dimension to privacy risk assessment, highlighting the relevance of contextual factors beyond traditional temporal and spatial dimensions. Moreover, the influence of environmental conditions emphasizes the importance of contextualizing re-identification risk within broader environmental contexts, where weather conditions can impact individuals' behavioral patterns and, subsequently, their risk of re-identification. In the context of demographic aspects, their limited influence on privacy risk can be attributed to their nature as derived from statistical aggregates. Instead, they tend to represent more general population trends rather than providing distinct and discriminative information at the individual level. As a result, their contribution to re-identification risk is generally less important compared to more granular and context-specific attributes like *location* and *time*. This observation highlights the importance of considering data attributes' contextual relevance and uniqueness when assessing privacy risks in trajectory datasets.

TABLE 10. **4square - aspect contribution to the re-identification risk.**

Combination	Contribution (%)
location	4.26908
time	23.41448
temperature	0.0081
location semantic, location	0.15603
location semantic, time	0.27135
location semantic, temperature	10.99296
location, temperature	18.93811
location, weekday	5.69651
location, size	4.01507
gender, temperature	0.06494%
age, temperature	0.08432%
job, temperature	0.07227%

Through these experiments, we demonstrated that aspects beyond time information and location significantly contribute to the re-identification risk. This finding underscores the nuanced interplay of various contextual factors in shaping privacy vulnerabilities within the trajectory datasets. By acknowledging the considerable impact of attributes such as age, gender, and others, we expand our understanding of the multifaceted nature of privacy risk. This insight is crucial for data providers to execute comprehensive risk assessments on multiple-aspects datasets and to implement effective mitigation strategies to safeguard individual privacy in data-driven environments.

In the next subsection, we will present the experiments with *AnonimoGuard*.

D. *AnonimoGuard*

This subsection presents a simulation of *AnonimoGuard* behavior within the three datasets. The objective was to illustrate the differences between the current privacy risk anonymization analysis in PRUDEnce data flow and the use of *AnonimoGuard*. We compared these in terms of the risk evaluation results.

Since we do not have any anonymization techniques specifically designed or proven to work with multiple-aspects of trajectories, we applied some generalizations to certain trajectory aspects. The purpose was to achieve a certain level of data protection to evaluate the behavior of *AnonimoGuard* with data after the anonymization process, which often involves generalization [6], [47].

As shown in Figure 23, we applied some generalizations when the location data had few visits. The blue line represents the results obtained using the *AnonimoGuard* method, with background knowledge configurations from the original dataset. The orange line represents PRUDEnce, which generates all possible combinations from the anonymized dataset and evaluates the risk.

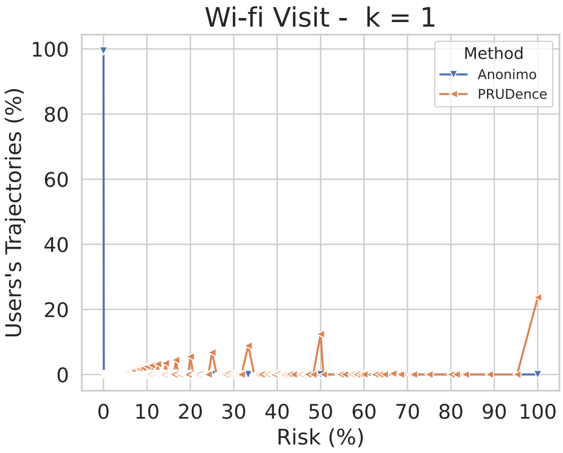


FIGURE 23. **Wi-fi AnonimoGuard x PRUDEnce k = 1.**

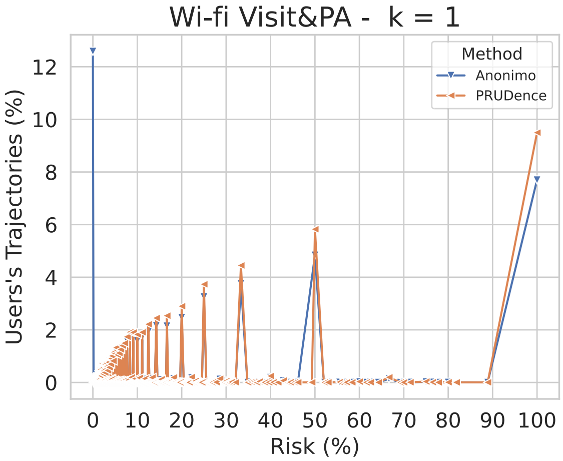


FIGURE 24. **Wi-fi AnonimoGuard x PRUDEnce k = 1.**

We observed a much lower risk with *AnonimoGuard* than with PRUDEnce. This is because the altered data from the original dataset no longer matched. When using PRUDEnce, we assume that the attacker's knowledge is equivalent to anonymized data, which is inaccurate. Thus, the only real risk exists when the anonymization method does not change the original data.

When different attacks and aspects are used, the same behavior can be observed across datasets. We have included some examples, as depicted in Figure 24 with *Visit and Permanent Aspects Attack* in the Wi-Fi dataset, and in Figure 25 with *Visit Attack* and Breadcrumbs dataset, as well as *Visit and Volatile Attack* with Foursquare 26.

As can be seen in the charts, *AnonimoGuard* provides a more realistic overview of privacy risk in datasets with applied data protection, showing a real decrease in risk and an evaluation with real background knowledge data.

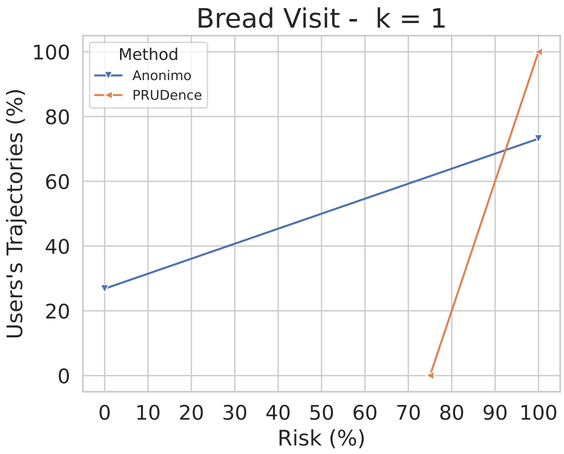


FIGURE 25. Breadcrumbs AnonimoGuard x PRUDEnce k = 1.

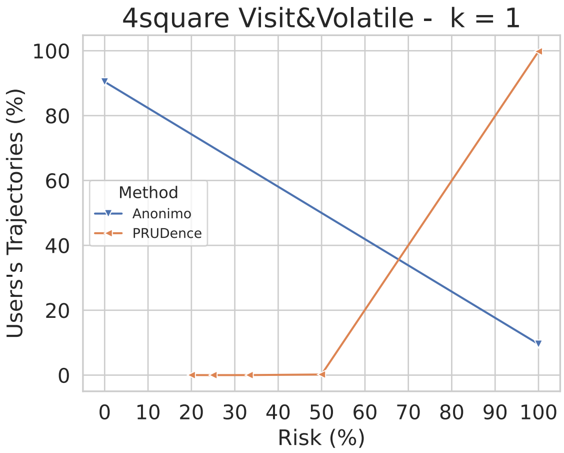


FIGURE 26. Foursquare AnonimoGuard x PRUDEnce k = 1.

VI. CONCLUSION AND FUTURE WORKS

We introduced the **TrajectGuard** privacy risk assessment framework as an extension of the state-of-the-art PRUDEnce for semantically enriched trajectories, the so-called multiple-aspects trajectories.

TrajectGuard is a valuable tool for data providers to safeguard privacy in mobility datasets characterized by multiple semantic aspects. It not only formulates and assesses privacy risk for these trajectories but also introduces attacks and nuanced risk evaluation through *AspectGuard*. Equitable privacy assessments were conducted on the anonymized datasets using *AnonimoGuard*.

To understand the impact of these aspects on re-identification, we conducted several experiments involving different and complementary semantically enriched trajectory datasets: Wi-Fi, Breadcrumbs, and Foursquare.

The results show that certain aspects, such as permanent, long-term, or volatile, might not pose a significant risk in some datasets. When we consider various attack scenarios involving multiple-aspects, we find that including more aspects refines how data are refined, making it more unique and easier to re-identify. However, the key factor influencing the risk is uniqueness, which can arise from location data alone, a combination of location and time details, or the inclusion of multiple-aspects. Therefore, understanding the results depends on the specific characteristics of the dataset. Factors such as the size of the coverage area, how data are spread over time, and the distribution of aspects values significantly affect the risk of re-identifying individuals. These insights from privacy risk analysis can be useful for data providers to check for increasing privacy risks as they enrich data with more semantic information.

AspectGuard's risk evaluation model focuses on protecting sensitive aspects and making informed data-sharing decisions aligned with data minimization principles. This approach helps data providers conduct comprehensive risk assessments and implement effective mitigation strategies to safeguard individual privacy in data-driven environments.

AnonimoGuard is a novel approach that ensures fairness in privacy risk evaluation by considering background knowledge from original and anonymized datasets. Due to the lack of established anonymization methods for multi-aspect trajectories, we simulated *AnonimoGuard* using simple suppression and generalization techniques. The results show a decrease in risk compared to using new data unfamiliar to the attacker, demonstrating *AnonimoGuard*'s effectiveness in assessing re-identification risk.

While the framework offers strong privacy evaluations, we recognize the computational challenges posed by using combinations to assess privacy risks across multiple dimensions. This process, particularly with large datasets, can be resource-intensive and time-consuming. To mitigate these issues, we propose optimizations such as more efficient algorithms, improved data structures, and strategies for scaling in environments like cloud and edge computing. Parallelization and load distribution can enhance TrajectGuard's efficiency in handling large datasets. In future work, we will explore these optimizations in greater detail, focusing on enhancing TrajectGuard's computational efficiency and scalability to make it more suitable for large-scale, real-world applications.

Looking ahead, there are many open questions about data privacy for trajectories with multiple aspects that require exploration. Specifically, there is a notable gap in research focused on anonymizing these multi-aspect trajectories. Developing an algorithm for this purpose would enable us to effectively test *AnonimoGuard*. Methods to assess the quality of anonymized multiple-aspects trajectories are currently lacking and need development. Test the framework with different, bigger, and more complex multiple-aspects trajectories datasets. Include additional attacks, such as those based on frequency and probability, in the framework. Additionally, we will explore TrajectGuard's compliance with privacy regulations beyond GDPR and LGPD, ensuring the framework aligns with broader global privacy standards.

REFERENCES

[1] S. Spaccapietra, C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto, and C. Vangenot, ''A conceptual view on trajectories,'' *Data Knowl. Eng.*, vol. 65, no. 1, pp. 126–146, Apr. 2008.

[2] R. D. S. Mello, V. Bogorny, L. O. Alvares, L. H. Z. Santana, C. A. Ferrero, A. A. Frozza, G. A. Schreiner, and C. Renso, ''MASTER: A multiple aspect view on trajectories,'' *Trans. GIS*, vol. 23, no. 4, pp. 805–822, Aug. 2019.

[3] F. Lettich, C. Pugliese, C. Renso, and F. Pinelli, ''Semantic enrichment of mobility data: A comprehensive methodology and the MAT-BUILDER system,'' *IEEE Access*, vol. 11, pp. 90857–90875, 2023, doi: 10.1109/ACCESS.2023.3307824.

[4] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, ''L-diversity: Privacy beyond k-anonymity,'' in *Proc. 22nd Int. Conf. Data Eng. (ICDE)*, Mar. 2006, p. 24.

[5] H. Zang and J. Bolot, ''Anonymization of location data does not work: A large-scale measurement study,'' in *Proc. 17th Annu. Int. Conf. Mobile Comput. Netw.*, Sep. 2011, pp. 145–156.

[6] O. Abul, F. Bonchi, and M. Nanni, ''Never walk alone: Uncertainty for anonymity in moving objects databases,'' in *Proc. IEEE 24th Int. Conf. Data Eng.*, Apr. 2008, pp. 376–385.

[7] M. Deng, K. Wuyts, R. Scandariato, B. Preneel, and W. Joosen, ''A privacy threat analysis framework: Supporting the elicitation and fulfillment of privacy requirements,'' *Requirements Eng.*, vol. 16, no. 1, pp. 3–32, Mar. 2011.

[8] OWASP. (2023). *Owasp Risk Rating Methodology*. [Online]. Available: https://owasp.org/www-community/OWASP_Risk_Rating_Methodology

[9] A. Goguen and A. Fringa, ''Risk management guide for information technology systems,'' Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep. QC 100 u57 800-30 2002 c.2, 2002, doi: 10.6028/NIST.SP.800-30.

[10] C. J. Alberts, S. G. Behrens, R. D. Pethia, and W. R. Wilson, ''Operationally critical threat, asset, and vulnerability evaluation (OCTAVE) framework, version 1.0,'' Carnegie-Mellon Univ. Pittsburgh PA Softw. Eng. Inst., Pittsburgh, PA, USA, Tech. Rep. CMU/SEI-99-TR-017 ESC-TR-99-017, 1999.

[11] J. Meier, *Improving Web Application Security: Threats and Countermeasures*. Redmond, WA, USA: Microsoft Press, 2003.

[12] F. Pratesi, A. Monreale, R. Trasarti, F. Giannotti, D. Pedreschi, and T. Yanagihara, ''Prudence: A system for assessing privacy risk vs utility in data sharing ecosystems,'' *Trans. Data Privacy*, vol. 11, pp. 139–167, Aug. 2018.

[13] R. Pellungrini, L. Pappalardo, F. Pratesi, and A. Monreale, ''A data mining approach to assess privacy risk in human mobility data,'' *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 3, pp. 1–27, May 2018.

[14] F. Naretto, R. Pellungrini, F. M. Nardini, and F. Giannotti, ''Prediction and explanation of privacy risk on mobility data with neural networks,'' in *ECML PKDD 2020 Workshops: Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020, Ghent, Belgium, September 14–18, 2020, Proceedings*. Springer, 2020, pp. 501–516.

[15] E. Ghasemi Komishani, M. Abadi, and F. Deldar, ''PPTD: Preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression,'' *Knowl.-Based Syst.*, vol. 94, pp. 43–59, Feb. 2016.

[16] G. Qiu, D. Guo, Y. Shen, G. Tang, and S. Chen, ''Mobile semantic-aware trajectory for personalized location privacy preservation,'' *IEEE Internet Things J.*, vol. 8, no. 21, pp. 16165–16180, Nov. 2021.

[17] L. Yao, Z. Chen, H. Hu, G. Wu, and B. Wu, ''Sensitive attribute privacy preservation of trajectory data publishing based on l-diversity,'' *Distrib. Parallel Databases*, vol. 39, no. 3, pp. 785–811, Sep. 2021.

[18] V. Bogorny, C. Renso, A. R. de Aquino, F. de Lucca Siqueira, and L. O. Alvares, ''CONSTAnT—A conceptual data model for semantic trajectories of moving objects,'' *Trans. GIS*, vol. 18, no. 1, pp. 66–88, Feb. 2014.

[19] V. Bogorny and M. Wachowicz, ''A framework for context-aware trajectory,'' in *Data Mining For Business Applications*. Boston, MA, USA: Springer, 2009, pp. 225–239.

[20] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, Y. Theodoridis, and Z. Yan, ''Semantic trajectories modeling and analysis,'' *ACM Comput. Surv.*, vol. 45, no. 4, pp. 1–32, Aug. 2013.

[21] L. O. Alvares, V. Bogorny, B. Kuijpers, J. A. F. de Macedo, B. Moelans, and A. Vaisman, ''A model for enriching trajectories with semantic geographical information,'' in *Proc. 15th Annu. ACM Int. Symp. Adv. Geographic Inf. Syst.*, Nov. 2007, p. 22.

[22] A. Monreale, R. Trasarti, D. Pedreschi, C. Renso, and V. Bogorny, ''C-safety: A framework for the anonymization of semantic trajectories,'' *Trans. Data Privacy*, vol. 4, no. 2, pp. 73–101, 2011.

[23] M. Elliot, C. J. Skinner, and A. Dale, ''Special uniques, random uniques and sticky populations: Some counterintuitive effects of geographical detail on disclosure risk,'' *Res. Off. Statist.*, vol. 1, pp. 53–67, 1998.

[24] M. Templ, B. Meindl, A. Kowarik, and S. Chen, ''Introduction to statistical disclosure control (SDC),'' IHSN, Work. Paper 007, 2014.

[25] Y. Song, D. Dahlmeier, and S. Bressan, ''Not so unique in the crowd: A simple and effective algorithm for anonymizing location data,'' in *Proc. PIR@ SIGIR*, 2014, pp. 19–24.

[26] J. P. Achara, G. Acs, and C. Castelluccia, ''On the unicity of smartphone applications,'' in *Proc. 14th ACM Workshop Privacy Electron. Soc.*, vol. 69, Oct. 2015, pp. 27–36.

[27] A. Basu, A. Monreale, J. C. Corena, F. Giannotti, D. Pedreschi, S. Kiyomoto, Y. Miyake, T. Yanagihara, and R. Trasarti, ''A privacy risk model for trajectory data,'' in *Proc. IFIP Int. Conf. Trust Manag.*, Singapore. Cham, Switzerland: Springer, Jul. 2014, pp. 125–140.

[28] A. Armando, M. Bezzi, N. Metoui, and A. Sabetta, ''Risk-based privacy-aware information disclosure,'' *Int. J. Secure Softw. Eng.*, vol. 6, no. 2, pp. 70–89, Apr. 2015.

[29] A. Narayanan and V. Shmatikov, ''De-anonymizing social networks,'' in *Proc. 30th IEEE Symp. Secur. Privacy*, May 2009, pp. 173–187.

[30] A. Ramachandran, Y. Kim, and A. Chaintreau, '''I knew they clicked when I saw them with their friends': Identifying your silent web visitors on social media,'' in *Proc. 2nd ACM Conf. Online Social Netw. (COSN)*, Dublin, Ireland, 2014, pp. 239–246.

[31] A. Cecaj, M. Mamei, and F. Zambonelli, ''Re-identification and information fusion between anonymized CDR and social network data,'' *J. Ambient Intell. Humanized Comput.*, vol. 7, no. 1, pp. 83–96, Feb. 2016.

[32] B. Khalfoun, S. Ben Mokhtar, S. Bouchenak, and V. Nitu, ''EDEN: Enforcing location privacy through re-identification risk assessment: A federated learning approach,'' *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 2, pp. 1–25, Jun. 2021.

[33] R. Pellungrini, ''Privacy risk and data utility assessment on network data,'' in *Proc. Int. Symp., From Data Models Back*. Cham, Switzerland: Springer, 2021, pp. 93–106.

[34] G. Mariani, A. Monreale, and F. Naretto, ''Privacy risk assessment of individual psychometric profiles,'' in *Proc. 24th Int. Conf. Discovery Sci.*, Halifax, NS, Canada. Cham, Switzerland: Springer, Oct. 2021, pp. 411–421.

[35] P. Silva, C. Gonçalves, N. Antunes, M. Curado, and B. Walek, ''Privacy risk assessment and privacy-preserving data monitoring,'' *Expert Syst. Appl.*, vol. 200, Aug. 2022, Art. no. 116867.

[36] H. Guo, X. Wu, J. Liu, B. Mao, and X. Chen, ''Adaptive and reliable location privacy risk sensing in Internet of Vehicles,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 9, pp. 12696–12708, Sep. 2024.

[37] F. Naretto, R. Pellungrini, S. Rinzivillo, and D. Fadda, ''Exphlot: Explainable privacy assessment for human location trajectories,'' in *Proc. Int. Conf. Discovery Sci.* Cham, Switzerland: Springer, 2023, pp. 325–340.

[38] N. Mohammed, B. C. M. Fung, and M. Debbabi, ''Walking in the crowd: Anonymizing trajectory data for pattern analysis,'' in *Proc. 18th ACM Conf. Inf. Knowl. Manage.*, Nov. 2009, pp. 1441–1444.

[39] A. Monreale, D. Pedreschi, R. G. Pensa, and F. Pinelli, ''Anonymity preserving sequential pattern mining,'' *Artif. Intell. Law*, vol. 22, no. 2, pp. 141–173, Jun. 2014.

[40] R. Yarovoy, F. Bonchi, L. V. S. Lakshmanan, and W. H. Wang, ''Anonymizing moving objects: How to hide a MOB in a crowd?'' in *Proc. 12th Int. Conf. Extending Database Technol., Adv. Database Technol.*, vol. 26, Mar. 2009, pp. 72–83.

[41] A. Monreale, G. L. Andrienko, N. V. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo, and S. Wrobel, ''Movement data anonymity through generalization,'' *Trans. Data Priv.*, vol. 3, no. 2, pp. 91–121, 2010.

[42] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, ''Unique in the crowd: The privacy bounds of human mobility,'' *Sci. Rep.*, vol. 3, no. 1, pp. 1–5, Mar. 2013.

[43] M. J. Elliot, A. M. Manning, and R. W. Ford, ''A computational algorithm for handling the special uniques problem,'' *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 493–509, Oct. 2002.

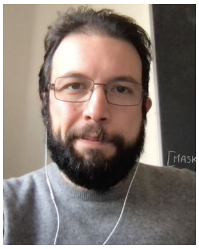
[44] L. Pappalardo, F. Simini, G. Barlacchi, and R. Pellungrini, "Scikit-mobility: A Python library for the analysis, generation and risk assessment of mobility data," 2019, *arXiv:1907.07062*.

[45] D. Yang, D. Zhang, V. W. Zheng, and Z. Yu, "Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 45, no. 1, pp. 129–142, Jan. 2015.

[46] A. Moro, V. Kulkarni, P.-A. Ghiringhelli, B. Chapuis, K. Huguenin, and B. Garbinato, "Breadcrumbs: A rich mobility dataset with point-of-interest annotations," in *Proc. 27th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, vol. 45, Nov. 2019, pp. 508–511.

[47] M. Gramaglia and M. Fiore, "Hiding mobile traffic fingerprints with GLOVE," in *Proc. 11th ACM Conf. Emerg. Netw. Exp. Technol.*, Dec. 2015, pp. 1–13.

FERNANDA OLIVEIRA GOMES is currently pursuing the joint Ph.D. degree with the University of Pisa and Santa Catarina's Computer Science Programs, Federal University. Her research interests include privacy, mobility data, semantic enrichment of trajectories, and blockchain. She has authored peer-reviewed publications in these fields.



ROBERTO PELLUNGRINI is currently a Research Fellow with Scuola Normale Superiore, Pisa, Italy. He is also a member of the Knowledge Discovery and Data Mining Laboratory (KDDLab), a joint research group with the Information Science and Technology Institute, National Research Council, Pisa. His research interests include mainly in data privacy, explainable AI, and human-AI collaboration.

ANNA MONREALE is currently an Associate Professor with the Computer Science Department, University of Pisa, and a member of the Knowledge Discovery and Data Mining Laboratory (KDD-Laboratory), a joint research group with the Information Science and Technology Institute, National Research Council, Pisa. Her research interests include big data analytics, social networks, and privacy issues raised in mining these kinds of social and human sensitive data. She has authored several peer-reviewed publications in these fields.

CHIARA RENSO received the Ph.D. degree in computer science. She is currently a Senior Researcher with the ISTI Institute of CNR, Italy. She has more than 100 peer-reviewed publications in the area of mobility analysis, machine learning and artificial intelligence methods for mobility data, analysis of geolocated social media, semantic enrichment of trajectories, and privacy.

JEAN EVERSON MARTINA is currently a Senior Lecturer in computer security with the Federal University of Santa Catarina. He has authored more than 80 peer-reviewed publications in the field of computer security. His research interests include cryptography, security protocols, applied formal methods, privacy, digital identity management, digital documents, and blockchain.

Open Access funding provided by 'Consiglio Nazionale delle Ricerche-CARI-CARE-ITALY' within the CRUI CARE Agreement