

k -scale: k -Anonymizing Millions of Trajectories

Abhishek Kumar Mishra

Inria, Lyon, France

Email: abhishek.mishra@inria.fr

Marco Fiore

IMDEA Networks Institute, Spain

Email: marco.fiore@networks.imdea.org

Abstract—Trajectory datasets collected by network operators and service providers offer detailed information about individual mobility and have wide application in business and research. However, managing such data raises privacy risks, as the unique movement patterns of individuals pose significant re-identification risks and make common countermeasures like pseudonymization ineffective. The privacy-preserving data publishing (PPDP) of trajectory datasets that maintains post-anonymization accuracy and truthfulness is an open problem –especially for large datasets with millions of records like those gathered by major actors in the telco ecosystem. We close this gap with **k -scale**, a framework that implements k -anonymity in massive mobile user trajectory datasets, removing uniqueness while safeguarding accuracy at the record level. Not only **k -scale** is the first model capable of scaling k -anonymization to a dataset of one million trajectories, but it does so while also outperforming state-of-the-art methods for trajectory data publishing in terms of preserved data quality, which we prove in real-world massive datasets and applications.

I. INTRODUCTION

Trajectory datasets collected by Mobile Network Operators (MNOs) or service providers offer detailed insights into the movements and activities of large user populations. Such data enable unprecedented investigations in demography, sociology, epidemiology, or networking [1]–[3] and can unlock new sources of revenues for MNOs by supporting original value-added services [4]. However, the mobility of individuals easily reveals private information like residence and workplace locations, commuting habits, daily schedules, religious beliefs, or visits to sensitive places such as clinics or nightclubs. Consequently, the collection, processing, and sharing of trajectory data are heavily regulated; for instance, the European Union (EU) General Data Protection Regulation (GDPR) directly prohibits companies –even those based outside Europe– from unwarrantedly storing mobility data of EU citizens [5].

Pseudonymization is insufficient. Legal frameworks and privacy concerns are barriers that limit the accessibility and circulation of trajectory data, curbing scientific investigation and industrial innovation. Data owners commonly alleviate the problem via *pseudonymization*, which replaces personal identifiers (names, phone numbers, or device codes like IMSI) with random or hashed values –hence keeping an association of all samples to the user without revealing her identity.

Unfortunately, as for other personal data types like medical and web databases [6], [7], pseudonymization is insufficient to prevent re-identification in trajectory data. Individuals exhibit movement patterns that are highly distinguishable even within large pseudonymized populations. Seminal experiments have indicated that 50% of the mobile users can be tracked down

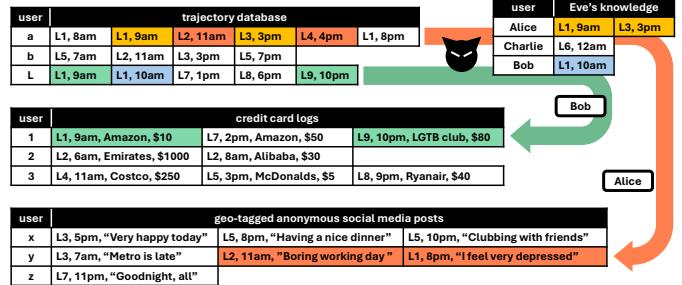


Fig. 1: Toy example illustrating record linkage attacks on trajectory data. Cell colors denote the spatiotemporal samples used to link unique trajectories with (i) Eve’s knowledge (in yellow for Alice and blue for Bob) and (ii) records in credit card or social media logs (orange for Alice and green for Bob)

in a database with 25 millions of trajectories just by knowing their three most frequently visited locations [8]. Or, pinpointing with near certainty the trajectory of any mobile subscriber among those of 1.5 million other customers only requires knowledge of five random observations of her movements [9].

Trajectories are prone to record linkage. The *uniqueness* of individual mobility patterns is a vulnerability that can be exploited for *record linkage* attacks, which aim at associating one specific trajectory to some external data. Fig. 1 exemplifies the operation of record linkage in a toy scenario where trajectories’ spatiotemporal samples consists of a numbered location (L) and an hour. The attacker, Eve, first re-identifies Alice and Bob in the target trajectory database by linking her adversarial knowledge (e.g., acquired by meeting or stalking the victims) to the unique records that contain the spatiotemporal samples in her possession; this is a risk per-se as Eve gains access to the full mobility data of Alice and Bob. Even worse, Eve can then run additional record linkage attacks to match Alice’s and Bob’s entries in the trajectory database to unique records in external databases that contain sensitive personal information like Alice’s mental condition or Bob’s sexual orientation. Notably, the fact that all databases are pseudonymized does not offer any protection against the attacks.

While this is an oversimplified example, record linkage attacks against real-world trajectory datasets have been demonstrated using external data with independent location and temporal resolutions [10] and have successfully linked pseudonymized mobile phone trajectories to personal information in publicly available social network metadata, credit card logs, and public transit databases [11]–[14].

Anonymization is a requisite. The privacy-preserving circulation of trajectories must hinge on more robust methods

than pseudonymization. Indeed, privacy regulations like the EU GDPR call for proper *anonymization* of trajectory data as the sole way to authorize a more open utilization and possibly public disclosure of trajectory databases [15].

As later reviewed in §V, several privacy criteria with diverse guarantees can be used to anonymize trajectory data, such as k-anonymity, l-diversity, t-closeness, or differential privacy. Yet, today *there is no practical solution to anonymize –according to any privacy criterion– production-scale trajectory databases that include the information of millions of individuals*, such as those available to major telco actors. All methods proposed to date are too expensive to be run on datasets beyond tens of thousands of trajectories or dramatically dilute data utility for datasets beyond such size [16].

Our contribution. We aim at removing uniqueness in trajectory databases to protect them from record linkage attacks like that illustrated in Fig. 1. Achieving this while maintaining the accuracy and truthfulness of individual records necessary for downstream analyses is a challenging open problem [17].

We close the gap above with **k-scale**, a novel approach that completely eliminates uniqueness in large-scale trajectory datasets via computationally efficient k-anonymity at the record level. Our approach relies on algorithmic innovations ensuring that the k-anonymity privacy principle is satisfied while the (usually high) complexity of fundamental operations such as trajectory grouping and merging is drastically reduced. We demonstrate with multiple real-world trajectory datasets that **k-scale** preserves higher data accuracy compared to previous proposals based on k-anonymization or differential privacy such as W4M [18], GLOVE [19], or SafePath [20]. Importantly, our solution scales to datasets with records in the order of millions, which is unprecedented in the literature.

We clarify that **k-scale** is no *silver bullet* to the problem of preserving privacy in trajectory data. The k-anonymity principle **k-scale** hinges upon has significant known limitations in presence of complex attacks beyond record linkage. However, our model renders k-anonymization a viable – and significantly safer – alternative to the pseudonymization approaches that currently dominate industry practice for protecting massive trajectory datasets, marking a key milestone toward a long-term solution to this challenge.

II. PROBLEM DEFINITION

Our goal is the Privacy-Preserving Data Publishing (PPDP) of trajectory databases, i.e., producing trajectory data that safeguard the privacy of data subjects and retain the usefulness of the data for any subsequent analysis –thus abiding by the well established fundamental principles of PPDP [17].

A. Attacker model

Attacks on published individual data belong to four classes: *record linkage*, *attribute linkage*, *table linkage*, and *probabilistic attacks* [17]. We aim at preventing record linkage attacks that match side information with a single record in the target database as in Fig. 1. The other types of attacks imply more powerful adversaries and are beyond the scope of this work.

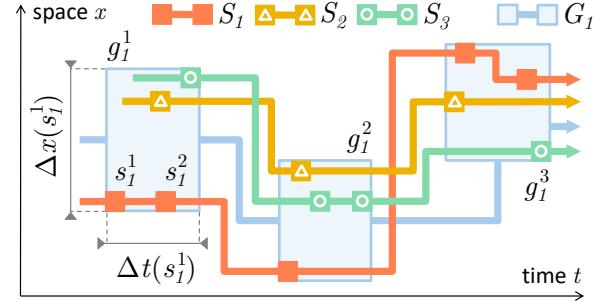


Fig. 2: Example of spatiotemporal generalization of trajectories S_1 , S_2 and S_3 into a generalized trajectory $G = \{g_1^1, g_1^2, g_1^3\}$. Space is unidimensional

The chances of success of a record linkage attack depend on the adversary’s knowledge: in the toy example in Fig. 1, Eve would have not been successful in re-identifying Alice in the trajectory database if she had possessed only one of the two samples associated to Alice –as either is found in two different trajectory records. Since no reliable model of the attacker’s knowledge exists in the context of spatiotemporal data [21], we assume that any subset of a user’s trajectory can potentially be part of the side information [22]. This is a worst-case scenario: our attacker model is more general and harder to counter than models limiting the knowledge of the adversary to, e.g., the most visited locations by the target individual [8] or a fixed subset of her trajectory samples [9].

B. Privacy model

By design, k-anonymity eliminates uniqueness from a database hence protects it against record linkage attacks [6]. The criterion commands that each entry in a k-anonymized dataset be *indistinguishable* from at least $k-1$ other records. The level of protection is proportional to k : a record linkage attack on a k-anonymized dataset returns at least k entries, leaving the adversary to guess the right target with a probability of success $1/k$. In the toy example of Fig. 1, k-anonymity would ensure that Eve finds k trajectories matching her knowledge on both Alice and Bob –thus rendering the attack ineffective.

We consider k-anonymity for full-length user trajectories, which formally is a quasi-identifier-blind anonymity of trajectory data [22]. This is the only way to ensure indistinguishability against a worst-case adversary who possesses an arbitrary subset of the target user’s trajectory as assumed in §II-A. In the toy example in Fig. 1, it guarantees that the attack fails under any Eve’s knowledge of Alice’s or Bob’s trajectory samples.

Finally, in order to implement k-anonymity we transform the original database via *spatiotemporal generalization*, which reduces data precision in space and time to make samples from different trajectories identical. We remark that generalization-based k-anonymity fully meet the PPDP principles [17].

C. Anonymization problem formulation

We now provide a formal definition of the anonymization problem that implements the privacy model in §II-B.

Let us consider a dataset Σ of U records S_1, \dots, S_U . The i -th record represents the spatiotemporal trajectory of network user i , and is a time-ordered sequence of spatiotemporal

samples (i.e., points in space and time), each mapping to a timestamped, geo-referenced event recorded for user i ; formally, $S_i = (s_i^1, \dots, s_i^{N_i})$, where every s_i^j is a triplet in the space $(\mathcal{X}, \mathcal{Y}, \mathcal{T})$ that represents the two spatial coordinates and the timestamp of the sample, respectively. Fig. 2 illustrates a toy example for three trajectories S_1, S_2 , and S_3 , considering a unidimensional space for simplicity of representation.

A spatiotemporal generalization Σ of Σ is a dataset of U records (G_1, \dots, G_U) , where the i -th record G_i represents the generalized version of the original spatiotemporal trajectory $S_i = (s_i^1, \dots, s_i^{N_i})$ of user i , and is a sequence of generalized spatiotemporal samples $G_i = (g_i^1, \dots, g_i^{M_i})$. Each generalized sample can be understood as a region in space and time: for instance, in Fig. 2, the sample $g_1^1 \in G_1$ generalizes $s_1^1 \in S_1$ by expanding it to a wider spatiotemporal region. One or more samples of S_i can be merged in a single generalized sample of G_i , such as s_1^1 and s_1^2 in Fig. 2, hence $M_i \leq N_i, \forall i$.

Let us now define a `gen` operation that returns a generalized trajectory rendering a set of input trajectories of cardinality L indistinguishable from each other¹. Formally,

$$G = \text{gen}(S_1, \dots, S_L), \quad (1)$$

ensures that, for each original trajectory S_1, \dots, S_L , each and every of its samples $s_i^1, \dots, s_i^{N_i}, \forall i = 1, \dots, L$, can be found in one of the generalized samples of G . For instance, in Fig. 2, G_1 is a valid `gen` output for the input (S_1, S_2, S_3) , since every g_1^k contains samples $s_i^j, \forall i = \{1, 2, 3\}$.

The generalization Σ of an original dataset Σ is said to be k -anonymous if every record it contains is indistinguishable from at least $k-1$ other records in Σ [6]. This condition is satisfied by definition if each trajectory $S_i \in \Sigma$ is generalized by $G_i \in \Sigma$ with at least $k-1$ other trajectories in Σ , or, formally

$$G_i = \text{gen}(S_i, S_{\pi_1(i)}, \dots, S_{\pi_{L_i}(i)}), \quad L_i \geq k-1, \quad \forall i. \quad (2)$$

Here, $\pi_1(i), \dots, \pi_{L_i}(i)$ are $L_i \geq k-1$ permutations of integers $1, \dots, U$ such that $\pi_l(i) \neq i$ and $\pi_l(i) \neq \pi_m(i), \forall i, l, m, l \neq m$. The constraints on the permutations prevent one index from appearing more than once in the set of sequences. In Fig. 2, a dataset of three generalized trajectories $\{G_1, G_2, G_3\}$ such that $G_1 = G_2 = G_3$ is 3-anonymous, as (2) is met for every i . Indeed, an adversary possessing an arbitrary subset $S_i^- \subseteq S_i$ of the samples of S_1, S_2 , or S_3 (as per our attacker model in §II-B), always finds S_i^- to match all 3 generalized records.

Many different choices of sets $\mathcal{S}(i) = \{S_{\pi_1(i)}, \dots, S_{\pi_{L_i}(i)}\}$ used to generalize each $S_i, \forall i$ fulfill (2). However, these choices do not yield a same *generalization cost*. Let $\mathcal{C}(G_i)$ denote the cost² of generalizing S_i into G_i . The optimal generalization problem of identifying the dataset $\tilde{\Sigma}^*$ that meets (2) and minimizes the overall generalization cost is

$$\arg \min_{\mathcal{S}(i)} \sum_{i=1}^U \mathcal{C}(\text{gen}(S_i, \mathcal{S}(i))), \quad \text{s.t. } |\mathcal{S}(i)| \geq k-1, \quad \forall i. \quad (3)$$

¹Multiple definitions of the `gen` operator are possible, and our problem formulation and its solution are agnostic of the specific implementation. We elaborate in §III-A about our choice of operator.

²As for the `gen` operator, different expressions for the cost $\mathcal{C}(\cdot)$ can be used. Details on the cost we adopt are in §III-A.

Solving (3) requires exploring all possible permutations $\pi_1(i), \dots, \pi_{L_i}(i)$ with $L_i \geq k-1$ for all original trajectories, while considering that the choice of \mathcal{S}_i for an input trajectory S_i affects that of \mathcal{S}_j for all other trajectories S_j . In fact, the optimal generalization problem can be mapped to a multidimensional assignment problem (MAP) with a cost hypermatrix $C = \{\mathcal{C}(G_i)\}$, which is known to be NP-hard [23].

III. K-SCALE DESIGN AND OPERATION

We propose a novel solution to the problem in (3), named `k-scale`. The full operation of our approach is illustrated in Fig. 3. Overall, `k-scale` receives as input the original dataset Σ of individual spatiotemporal trajectories, and produces a k -anonymized version $\tilde{\Sigma}$ of the same. To this end, it processes the dataset in two main phases, which are outlined next and whose internal steps are then detailed in §III-A–III-E.

Efficient cost matrix calculation. Any solution to (3) builds on the hypermatrix $C = \{\mathcal{C}(G_i)\}$ that contains the costs of generalizing each input trajectory S_i with any possible combination of $k-1$ other trajectories. Yet, computing C is unfeasible for massive datasets of millions of records –even in the simplest case where $k = 2$. Instead, `k-scale` hinges on the observation that only a tiny fraction of the full C is actually relevant to the solution of the problem in (3). An accurate approximation of the optimal solution can in fact be obtained from a very sparse matrix \tilde{C} that only contains those generalization costs that are *small*, i.e., are issued by combinations of similar trajectories that have an actual chance of being part of the optimal solution of (3).

To determine the sparse matrix of useful costs, `k-scale` first defines an optimal `gen` operator (§III-A) as well a computationally lightweight distance metric \mathcal{D} that approximates it (§III-B). We use \mathcal{D} to calculate a K -nearest neighbors (KNN) matrix than only contains the K most similar trajectories³ for each trajectory in Σ . Finally, a cost matrix \tilde{C} limited to pairwise costs and with dimensions of $U \times K$ (with $K \ll U$) is derived by running `gen` for all KNN elements (Sec III-C).

Selection of the k -anonymization sets. The second phase of the `k-scale` framework implements the k -anonymization process starting from \tilde{C} . Our proposal hinges upon an original *anonymity constraint*, i.e., the condition that bounds the choice of the trajectories $\mathcal{S}(i)$ that shall be used to generalize each trajectory S_i in Σ . Specifically, our anonymity constraint implements the k -anonymity principle in an innovative way, which dramatically reduces the complexity of deriving the sets $\mathcal{S}(i)$ for all $i = \{1, \dots, U\}$ while also granting high flexibility in the trajectory selection –which ultimately leads to lower-cost generalized samples (§III-D).

Applying the `gen` function on the selected sets $\mathcal{S}(i)$ for all input trajectories produces the generalized trajectories G_i that constitute the final k -anonymized dataset $\tilde{\Sigma}$ (§III-E).

³The K parameter of the KNN matrix is semantically different from and shall not be confounded with the k parameter of k -anonymization.

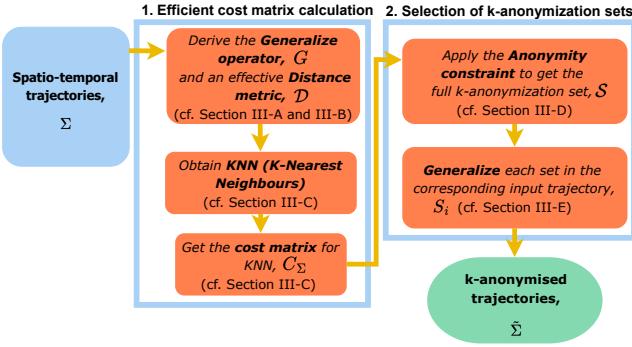


Fig. 3: Overview of k-scale operation

A. Generalization operator

As a main objective of k-scale is extreme scalability, our solution leverages expressions of the generalization operator gen and of its cost $\mathcal{C}(\cdot)$ that are simple yet retain interesting properties in terms of data quality.

Let $\Delta t(\cdot)$, $\Delta x(\cdot)$, and $\Delta y(\cdot)$ be the span in time and in the two space dimensions of the generalized sample to which the sample in the argument is mapped. For instance, in Fig. 2 the first spatiotemporal sample of S_1 , s_1^1 , is generalized into a sample g_1^1 that spans $\Delta t(s_1^1)$ in time and $\Delta x(s_1^1) + \Delta y(s_1^1)$ in space as illustrated. We define the cost of generalizing an input trajectory S_i into a k-anonymous trajectory G_i as

$$\mathcal{C}(G_i) = \frac{1}{N_i} \sum_{n=1}^{N_i} \Delta t(s_i^n) \cdot (\Delta x(s_i^n) + \Delta y(s_i^n)), \quad (4)$$

where we recall that $N_i = |S_i|$ is the number of spatiotemporal samples in the original trajectory. We note that more than one sample of the original trajectory may be mapped to the same generalized sample, in which case that generalized sample is counted more than once in the cost computation.

The expression in (4) implies that the cost of generalizing a single sample s_i^n is given by the product of the temporal expansion $\Delta t(s_i^n)$ by the total spatial expansion $\Delta x(s_i^n) + \Delta y(s_i^n)$. The total cost $\mathcal{C}(G_i)$ is then the average cost of all input samples. For instance, the cost of G_1 in Fig. 2 is the mean area of the light blue rectangles of g_1^1 , g_1^2 , and g_1^3 .

We remark that the cost definition in (4) enhances similar previous proposals of cost that also leverage the product of the temporal and spatial expansions of samples [24]. However, priori formulations simply sum up costs over all s_i^n , which does not reflect the loss of accuracy incurred by having several samples together instead of having them separate⁴.

Having defined $\mathcal{C}(\cdot)$, an optimal gen operator shall match the samples of any number L of input trajectories S_1, \dots, S_L

⁴An example that illustrates the problem is the following: let G_1 be a generalized trajectory with a single sample of size $\Delta x, \Delta y$ comprising a long period Δt ; let us also consider another generalized trajectory G_2 with a large number n of samples within the exact same space $\Delta x, \Delta y$, each covering a small time interval of size δt , such that $n \cdot \delta t = \Delta t$ (i.e., all the smaller intervals are adjacent and together cover the same period as the first trajectory). Previous expressions yield the same cost for G_1 and G_2 , yet the first is clearly less accurate [24]. Our formulation (4) solves the issue.

so as to minimize the costs $\mathcal{C}(G_1), \dots, \mathcal{C}(G_L)$, formally

$$G = \arg \min_{\text{gen}(S_1, \dots, S_L)} \mathcal{C}(\text{gen}(S_1, \dots, S_L)). \quad (5)$$

To solve this problem, we rely on an optimal and computationally efficient algorithm, k-merge [24], which can identify G in (5) from any number L of input trajectories when the generalization cost is expressed as a product of time and space expansions as in (4). The k-merge algorithm takes advantage of time coherence (i.e., the fact that generalized samples cannot overlap to avoid spatiotemporal ambiguity) to efficiently navigate the space of solutions. In practice, this results in a complexity $\mathcal{O}(\sum_{i=1}^L N_i)$, i.e., a linear cost in the total number of samples in all the sequences to be generalized.

B. Proxy distance metric

Computing the approximate cost matrix \tilde{C} requires running the gen operator billions of times over the high-dimensional data represented by trajectories spanning weeks and featuring thousands of samples each. While optimal and efficient, k-merge is too expensive to derive \tilde{C} for the production-scale trajectory databases we target. We introduce a proxy metric that gives a cheap approximation of the cost $\mathcal{C}(\text{gen}(S_i, S_j))$ of generalizing two trajectories with k-merge.

Let CM_i be the spatial center of mass (CM) of the spatiotemporal samples of a trajectory S_i , and RG_i their radius of gyration (RG)⁵. The geographical square centered at CM_i and with side $2 * \text{RG}_i$ is denoted by B_i . Our distance metric, \mathcal{D} , is then defined for a given pair of trajectories S_i and S_j as

$$\mathcal{D}(S_i, S_j) = \begin{cases} |\text{CM}_i - \text{CM}_j|_M & \text{if } m_{ij} = 1 \\ m_{ij}, & \text{otherwise,} \end{cases} \quad (6)$$

where

$$m_{ij} = \frac{1 - c_{ij}}{1 + c_{ij}} \quad \text{and} \quad c_{ij} = 2 \frac{B_i \cap B_j}{B_i \cup B_j}. \quad (7)$$

The term m_{ij} lies in the range $[0, 1]$, where 0 denotes perfectly matching CM and RG for i and j , and 1 indicates completely disjoint squares. When $m_{ij} < 1$ there is an overlap between squares B_i and B_j , which is used as the value for \mathcal{D} . When $m_{ij} = 1$, \mathcal{D} is the Manhattan distance $|\cdot|_M$ of the two CM.

C. Approximate cost matrix calculation

Using the proxy metric above, we propose and parametrize a dedicated KNN construction methodology. Our KNN computation leverages a BallTree [25] method that uses \mathcal{D} in (6) as the distance measure. We find this methodology to be especially efficient in our experiments and to outperform other approaches like KDTree [26] in high dimensions.

An important design choice concerns the number of nearest neighbours K in the KNN computation. The value controls a key trade-off: a lower K implies faster calculations of \tilde{C}

⁵The center of mass of a trajectory S_i is defined as $\text{CM}_i = 1/N_i \cdot \sum_{n=1}^{N_i} r_n$ and the radius of gyration as $\text{RG}_i = \sqrt{1/T_i \cdot \sum_{n=1}^{N_i} t_n(r_n - \text{CM}_i)^2}$. In both definitions, r_n represents spatial positions (x, y) of the samples s_i , while t_n is the time spent by the user i at location r_n out of total time T_i .

but limits the variety of candidate trajectories for downstream k-anonymization. Empirical evaluations omitted here due to space limitations show that a value of K that is two orders of magnitude larger than that the k used by k-anonymization is sufficiently large to encompass most of the closest trajectories required to generalize at low cost every input trajectory.

After having computed the KNN structure, we use the `gen` operator to compute the actual cost of generalizing every input trajectory with its set of K nearest neighbours identified by the KNN. The results is the sparse, approximate, and pairwise cost matrix of generalization costs C .

D. Anonymity constraint

As discussed in §II-C, the optimal k-anonymization of all trajectories in a dataset requires solving an NP-hard multidimensional assignment problem to select the best combination of sets $\mathcal{S}(i)$, i.e., the ensemble of $k-1$ trajectories that shall be generalized with every trajectory S_i to ensure its k-anonymity. The anonymity constraint defines the exact conditions that the combinations of $\mathcal{S}(i)$ must meet to solve the problem. Next, we present two previous constraints and our original proposal.

Full consistency. The classic approach to coordinate the choice of $\mathcal{S}(i)$ is enforcing a *full consistency*, where including a trajectory S_j in the anonymization set $\mathcal{S}(i)$ of trajectory S_i automatically lets S_i be part of the anonymization set of S_j also. Formally, *full consistency* means that

$$S_i \in \mathcal{S}(j) \iff S_j \in \mathcal{S}(i) \quad \forall S_i, S_j \in \Sigma, i \neq j, \quad (8)$$

which de facto forces the construction of *disjoint clusters*, each containing at least k trajectories that are then generalized into each other. As a visual example, for the four raw trajectories in the bottom plot of Fig. 4a, *full consistency* achieves 3-anonymity by creating a single cluster where all users are anonymized with each other. In a graph representation where each directed edge implies that the out-node trajectory is generalized with the in-node one, this constraint creates fully connected sub-graphs of k or more trajectories, as in Fig. 4a.

The k-pick constraint. An alternative, less restrictive condition is the *k-pick* constraint proposed by [24]. Here, each trajectory S_i picks exactly $k-1$ other trajectories in the dataset to form its anonymization set $\mathcal{S}(i)$, under the constraint that every trajectory in Σ must be picked by at least $k-1$ different trajectories to achieve k-anonymity. In practice, in a graph representation, implementing *k-pick* requires forming *closed cycles* of k or more trajectories picking one another [24], which grants higher flexibility in choosing $\mathcal{S}(i)$ compared to the clustering of *full consistency*. The top-right plot in Fig. 4a illustrates the graph of closed cycles returned by this strategy in the 3-anonymity toy example.

We argue however that *k-pick* is still unnecessarily removing degrees of freedom in the anonymization set selection process. Our claim is that k-anonymity is achieved by solely having each trajectory $S_i \in \Sigma$ be picked by $k-1$ other trajectories S_j and included in their anonymization set $\mathcal{S}(j)$; instead, forcing trajectories to pick $k-1$ others as done in *k-pick* is gratuitous. In other words, there is no need

that the generalization always occurs among $k-1$ trajectories to guarantee k-anonymity, rather it is sufficient that each trajectory appears in $k-1$ generalizations.

The nu-k-pick constraint. The last observation above lets us formulate a new anonymity constraint, which we name *nu-k-pick* (from *non-uniform*, see next), which commands precisely (and only) that each trajectory S_i be picked by at least $k-1$ other trajectories. This results in graphs of selected trajectories that are not constrained to closed cycles, as in the middle plot of Fig. 4a. We now formally validate that *nu-k-pick* solves the problem of k-anonymity in §II-C.

Proposition: Each sample of trajectory S_i is guaranteed to be generalized with samples of trajectories $S_{j_1}, S_{j_2}, \dots, S_{j_{k-1}}$.

Claim: An attacker possessing any subset of points of trajectory S_i will find those points in the generalized trajectories $G_i, G_{j_1}, G_{j_2}, \dots, G_{j_{k-1}}$, thereby satisfying k -anonymity.

Proof of Claim: By the definition of k -anonymity, any subset of points from trajectory S_i must be indistinguishable from points in at least k trajectories. Given the $k-1$ trajectories $S_{j_1}, S_{j_2}, \dots, S_{j_{k-1}}$ selected to generalize with trajectory S_i using *nu-k-pick*, each sample of trajectory S_i is generalized with samples from these $k-1$ trajectories utilizing the generalization operator G . Therefore, any subset of points $P_i \subseteq S_i$ from trajectory S_i will be included in the generalized trajectories $G_i, G_{j_1}, G_{j_2}, \dots, G_{j_{k-1}}$, ensuring that an attacker possessing such a subset will find those points distributed among at least k trajectories. Formally,

$$\forall P_i \subseteq S_i, \exists G_i, G_{j_1}, \dots, G_{j_{k-1}} \text{ s.t. } P_i \subseteq \bigcap_{m=0}^{k-1} G_{j_m}. \quad (9)$$

The points in P_i are therefore guaranteed to be indistinguishable from those in the generalized trajectories $G_i, G_{j_1}, G_{j_2}, \dots, G_{j_{k-1}}$, which satisfies k -anonymity.

Toy example. To illustrate the rationale behind the new constraint, let us consider the toy example depicted in Fig. 4, which shows spatiotemporal trajectories of users S_1, S_2, S_3 , and S_4 . For simplicity, we consider that each trajectory is composed by two samples, yet the discussion generalizes to trajectories of any cardinality. Also, let the k-anonymity parameter be $k = 3$. The three constraints operate as follows.

- **full consistency.** A cluster with at least 3 trajectories must be formed to meet the constraint, hence the only option is for each trajectory to include all others in its anonymization set. All trajectories have then identical (large and high-cost) generalized trajectories $\{G_1, G_2, G_3, G_4\}$ that contain samples of all raw trajectories as per Fig. 4b.
- **k-pick.** An optimal solution is one where S_1 picks $\{S_2, S_4\}$, S_2 picks $\{S_1, S_3\}$, S_3 picks $\{S_2, S_4\}$, and S_4 picks $\{S_1, S_3\}$. This results in S_1 generalizing with $\{S_2, S_4\}$, S_2 with $\{S_1, S_3\}$, S_3 with $\{S_2, S_4\}$, and S_4 with $\{S_1, S_3\}$ –see the graph in Fig. 4a. The generalized trajectories G_2, G_3, G_4 include a lower number of samples each, hence have reduced total cost with respect to the *full consistency* case as shown in Fig. 4c.

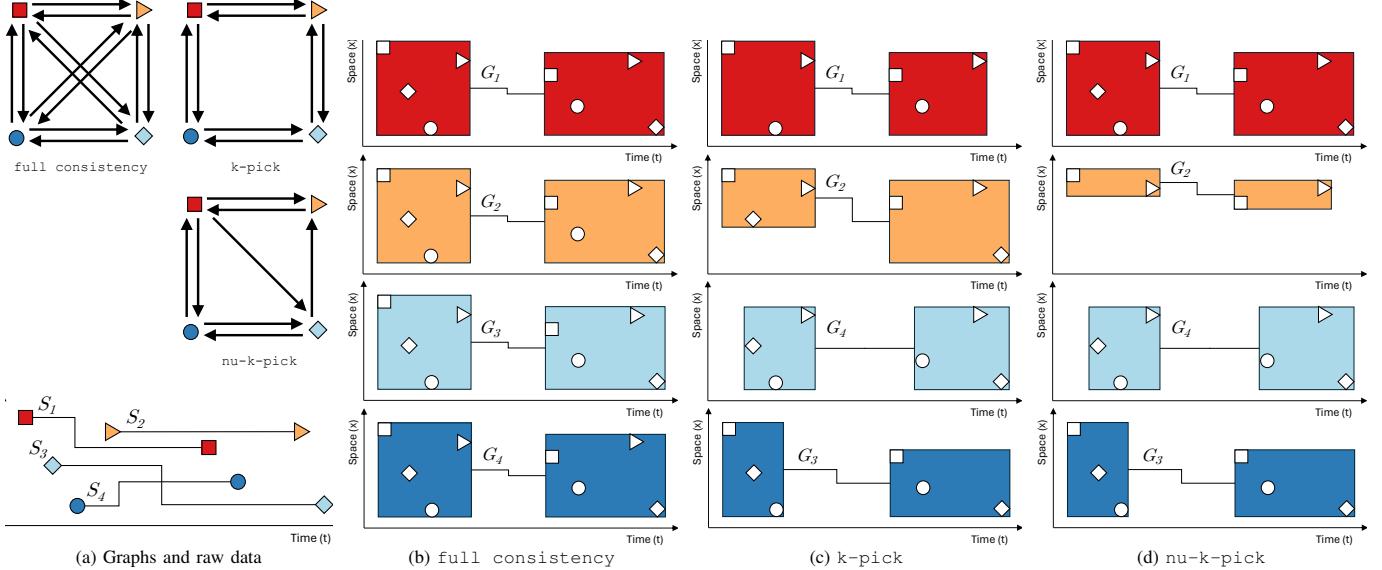


Fig. 4: Visual example of anonymity constraints. Space is unidimensional to ease the representation. (a) Top: graphs illustrating selections of the anonymization sets abiding by each constraint, where a directed edge implies that the out-node trajectory is generalized with the in-node one. Bottom: toy example of four two-sample trajectories to be 3-anonymized. (b), (c), (d) Resulting generalized trajectories from G_1 (top) to G_4 (bottom) under each anonymity constraint

- nu-k-pick. The optimal solution lets S_1 be picked by $\{S_2, S_4\}$, S_2 by $\{S_1, S_3\}$, S_3 by $\{S_1, S_4\}$, and S_4 by $\{S_1, S_3\}$. This results in generalization of S_1 with $\{S_2, S_3, S_4\}$, of S_2 with $\{S_1\}$, of S_3 with $\{S_2, S_4\}$, and of S_4 with $\{S_1, S_3\}$. Removing uniformity in the number of trajectories involved in each generalization allows reducing substantially the cost of $\{G_2\}$ without penalizing the other generalized trajectories as in Fig. 4d.

The plots in Fig. 4b–4d also intuitively illustrate how all anonymity constraints effectively realize 3-anonymity. In all cases, an attacker with partial or full knowledge of the raw samples of any of the input trajectories would find such samples in 3 different generalized trajectories. This makes it impossible to distinguish an original trajectory from $k-1$ others [6] and fulfills the k -anonymity definition in §II-C.

E. Trajectory database k -anonymization

Algorithm 1 Algorithm for k -anonymization

```

1: procedure K-SCALE( $\Sigma, \tilde{C}_\Sigma, \tilde{C}, k$ )
2:    $\tilde{\Sigma} = []$ 
3:   A = [False] $_{|\Sigma| \times |\Sigma|}$ 
4:    $\tilde{C}_\Sigma = GetFullCostMatrix(\tilde{C})$ 
5:   for col  $\leftarrow \tilde{C}_\Sigma$  do            $\triangleright$  Pick lowest costs from full cost matrix
6:      $V_{col} = PickLowestCosts(col, k-1)$ 
7:      $MarkTrue(A, V_{col})$ 
8:   end for
9:   for index, row  $\leftarrow \tilde{C}_\Sigma$  do            $\triangleright$  Trajectory generalization
10:    S = row[SelectTrueVals(A, index)]
11:    if then $|S| > 0$             $\triangleright$  Mitigate granularity-related attacks
12:       $\tilde{\Sigma}.add(gen(S, C, k))$ 
13:    else
14:       $\tilde{\Sigma}.add(gen(\{index, MinNb(C_\Sigma, index)\}, C, k))$ 
15:    end if
16:   end for
17: return  $\tilde{\Sigma}$ 
end procedure

```

A major advantage of nu-k-pick is that it is remarkably simpler to implement than other anonymity constraints. It removes the need for sets of at least k trajectories of the other two approaches and allows each trajectory to have an anonymization set of arbitrary size. This represents a huge computational advantage, since it makes the choice of each $S(i)$ fully *independent* of that of the other $S(j), \forall j$.

Building on the property above, we devise an efficient algorithm to achieve k -anonymization of a target trajectory database as per Alg. 1. We first obtain a full $|\Sigma| \times |\Sigma|$ cost matrix \tilde{C}_Σ by expanding the cost matrix \tilde{C} of size $|\Sigma| \times K$: all the non-neighbour trajectories whose generalization cost is not present in C are marked with an ∞ cost. We also introduce a logical $|\Sigma| \times |\Sigma|$ matrix A initiated with false values and storing the anonymization sets. We then parse each column of \tilde{C}_Σ , marking in A the $k-1$ entries with minimum gen cost. This effectively lets each trajectory (column) be picked by exactly $k-1$ other trajectories (rows), which meets our new anonymity constraint. As anticipated, nu-k-pick can be fulfilled with high computational efficiency, by parsing \tilde{C}_Σ once with cost $\mathcal{O}(|\Sigma|^2)$ or, equivalently, $\mathcal{O}(U^2)$. In other words, our nu-k-pick constraint allows dramatically reducing the complexity of the problem problem in (3), from NP-hard under a full consistency to quadratic.

We then generalize the original trajectories in Σ based on A. For each trajectory S_i we find in A the set of other trajectories which S_i shall be merged with. If the set is not empty, we apply the gen operator in §III-A to S_i and all the trajectories in the set. If instead the anonymization set of S_i is empty, no generalization of its samples is necessary to achieve k -anonymity. Still, an adversary could infer that S_i is a raw trajectory from its (unchanged) spatiotemporal granularity, especially in datasets where all raw samples have the same resolution. To mitigate this risk, we also always generalize S_i

Id	Samples	Users	Resolution	User density	Duration
sen	144M	1M	117 km ² /BS	5.1 user/km ²	2 weeks
civ	4.6M	49.3k	263 km ² /BS	0.15 user/km ²	2 weeks
mob	21.4M	100k	0.25 km ² grids	10 user/km ²	75 days

TABLE I: Datasets considered for k-anonymization

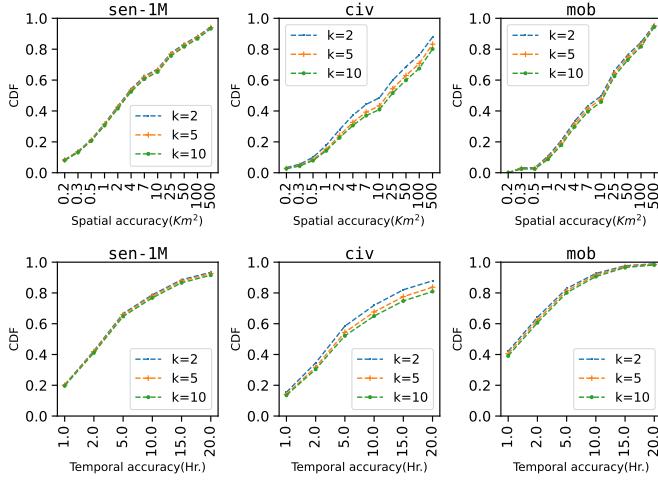


Fig. 5: Accuracy of datasets k-anonymized with $k\text{-scale}$

with its minimum cost neighbor in \tilde{C}_Σ .

The output of $k\text{-scale}$ is then a k-anonymized dataset $\tilde{\Sigma}$ where every subset of the samples of an original trajectory S_i of Σ are found in k generalized trajectories G_{j_1}, \dots, G_{j_k} .

IV. PERFORMANCE EVALUATION

We validate the performance and scalability of $k\text{-scale}$ with three distinct large trajectory databases. Two are Call Detail Record (CDR) datasets collected in Senegal (`sen`) and Ivory Coast (`civ`) by Orange as part of the D4D Challenge [27]. The third dataset (`mob`) consists of metropolitan-scale GPS trajectories in Japan [28]. The properties of the selected databases, in Tab. I, show their heterogeneity in terms of population density, user movements, and spatiotemporal granularity. For `sen` dataset, we randomly sample subsets of various sizes from 1,000 to 1 million (`sen-1k`, ..., `sen-1M`).

We run all experiments in a Centos server with 32 AMD EPYC 7262 8-Core Processors having 3.2 GHz CPU and 512 KB cache. Most $k\text{-scale}$ operations are parallelizable hence our implementation uses all available CPU cores.

Accuracy upon k-anonymization. We define the *anonymization accuracy* of a generalized dataset as the loss of resolution incurred by the data. The *spatial accuracy* is the geographical span of a generalized sample (in Km²), whereas the *temporal accuracy* is the duration of a generalized sample (min or Hr.).

Fig. 5 provides an overview of spatiotemporal accuracy for the three reference datasets, i.e., `sen-1M`, `civ` and `mob`, upon k-anonymization with $k\text{-scale}$ with $k = \{2, 5, 10\}$. Let us focus first on the `sen-1M` dataset, highlighting that *this is the first time that k-anonymization is achieved for a 1M trajectory dataset of real-world weeks-long user movements*. In this case, ~30 percent of the generalized samples have spatial accuracy

Metric	Dataset	W4M	Glove	k-scale
Mean spatial error [km]	sen-1k	30.1	19.9	5.8
	sen-2k	26.2	14.9	4
	sen-5k	17.4	-	4
	sen-1M	11.8	-	3.1
Mean time error [min]	sen-1k	2654	392	237
	sen-2k	2664	221	221
	sen-5k	2800	-	209
	sen-1M	2619	-	183
Created samples [%]	sen-1k	19.7	0	0
	sen-2k	21.2	0	0
	sen-5k	21.7	-	0
	sen-1M	19.3	-	0
Deleted samples [%]	sen-1k	23	1.2	0
	sen-2k	23.3	1.826	0
	sen-5k	24.4	-	0
	sen-1M	19.1	-	0

TABLE II: Comparative analysis of k-anonymization for $k = 2$

≤ 1 km² and $\sim 70\%$ have an area ≤ 10 km². Such accuracy is aligned with the low spatial resolution of the original dataset in Tab. I and retains sufficient utility for downstream analyses of the individual trajectories at city and country scales, as we will demonstrate later in this section. Similar considerations hold for the spatial accuracy when looking at the `civ` and `mob` datasets. The differences are mainly in the temporal dimension, with the k-anonymized `mob` dataset retaining higher resolution in time: the uniform GPS sampling of the original data helps finding temporally close samples for generalization.

A striking property of $k\text{-scale}$ is its robustness to k . Even $k = 10$ does not cause significant drops in spatiotemporal accuracy, proving that $k\text{-scale}$ allows protecting large-scale trajectory datasets with practical values of k .

Complexity analysis. The $k\text{-scale}$ framework has three components: (i) the construction of the KNN expedited by our proxy metric has a cost $\mathcal{O}(N_i \cdot U \log U)$; (ii) the computation of the cost matrix is of order $\mathcal{O}(U \cdot K \cdot N_i)$; and (iii) the processes of anonymization and generalization have a complexity of $\mathcal{O}(U \cdot N_i)$. In practical terms, the overall complexity of $k\text{-scale}$ stays less than quadratic in the number of trajectories of the input dataset, making the solution extremely scalable. As a result, k-anonymizing the `sen-1M` dataset took in around 10 hours, while datasets of up to 200,000 records required less than two hours –in the low-end server mentioned before. Considering the one-time nature of the anonymization operation, $k\text{-scale}$ is a computationally viable solution for very large telco players to protect their trajectory data subjects.

Comparison against k-anonymity frameworks. We compare $k\text{-scale}$ against the two best solutions for the k-anonymization of trajectory databases available to date [16]: `W4M` in a variant with linear spatiotemporal distance and chunking (LC) specifically designed for large datasets [18] and `Glove` [19]. We apply `W4M` with $\delta = 2$ km and 10% trashing, based on extensive parametrization experiments. We also optimize `Glove` by replacing its original trajectory generalization method with the more advanced `k-merge`.

Tab. II compares all methods across `sen` datasets of different sizes, with $k = 2$. Note that we are bounded to $k = 2$ by

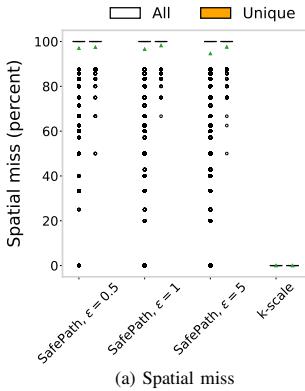


Fig. 6: Comparative analysis of k -scale ($k = 30$) and SafePath

the benchmarks, which experience a dramatic drop of accuracy under higher k ; and, even in that case *Glove* cannot scale beyond a relatively small dataset of 2,000 trajectories.

The results make the superior performance of k -scale apparent in terms of accuracy. The spatial resolution retained by our method is 71–78% more accurate than that of the best competitor for each dataset size. The temporal accuracy of k -scale is around ten times better than that of *W4M* and comparable with that of *Glove* for the small datasets that the benchmark can anonymize. Interestingly, these numerical results let us appreciate how larger datasets yield increasingly more accurate generalized trajectories, implying that the utility of the data may be even higher in production-scale databases of, e.g., tens of millions of trajectories.

In addition, we remark that *W4M* achieves its accuracy by allowing both the generation of new synthetic samples and the removal of hard-to-anonymize ones, whereas *Glove* only permits the latter. In Tab. II, *W4M* generates a substantial number of synthetic samples that amount to around 20% of the original data and removes an even higher fraction of the input dataset; *Glove* suppresses around 2% of the original trajectories. New fictitious samples not corresponding to actual user movements violate the PPDP principles and can create movement patterns that do not exist in the real world, hence biasing subsequent analyses [17]; removal of data implies a loss of original mobility information. We highlight that k -scale achieves better accuracy than the benchmark without adding or suppressing any sample.

Comparison against differential privacy. SafePath [20] is the state-of-the-art tool for the generation of differentially private (DP) versions of trajectory databases (see §V). Unfortunately, we were not able to apply SafePath to our reference datasets, as it does not scale to country-wide samples due to its high time complexity. We opted instead for comparing SafePath and k -scale on the data used for the evaluation of the DP tool in its original paper [20]: the dataset consists of 100,000 metro commuters trajectories along 68 possible locations (i.e., metro stations), with a time granularity of one hour. The reason why SafePath can operate on such data is that the limited cardinality of the spatial dimension lets many users exhibit the exact same trajectory: in fact, there are only 1,811 unique spatiotemporal trajectories

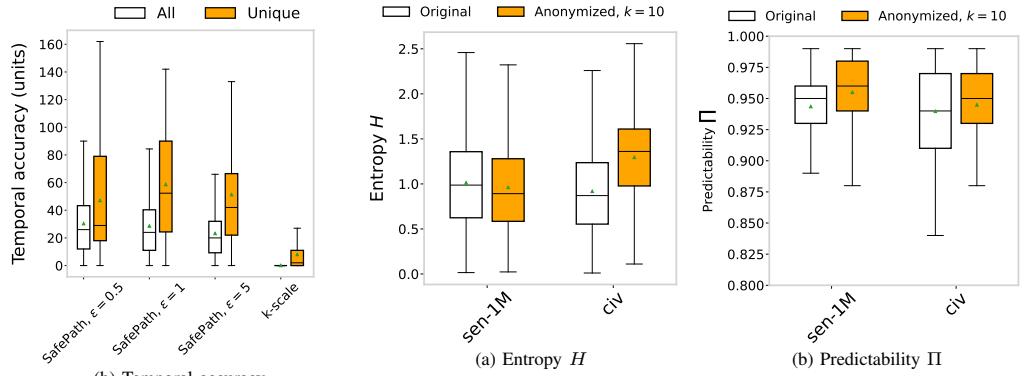


Fig. 7: Entropy and predictability of original and k -anonymized ($k = 10$) data

in the raw data, each repeated for many users. We retain the parametrization of SafePath used in the original paper and vary the privacy budget $\epsilon \in \{0.5, 1, 5\}$: lower ϵ results in more noise being introduced to the actual trajectory, hence yields stronger privacy guarantees but reduced utility of the sanitized data. For k -scale, the redundancy of the raw data lets us set $k = 30$.

Fig. 6 summarizes the results, for the two cases where the input dataset contains *all* 100,000 trajectories or just the 1,811 *unique* ones. Given that train stations are associated to identifiers with no geographical dimension, we define the spatial accuracy as the percentage of locations in the original trajectory that are absent in the output DP or k -anonymized trajectory. Fig. 6a illustrates how SafePath misses on average more than 90 percent of the original spatial information in each DP trajectory it generates. On the contrary, the k -anonymized dataset produced by k -scale does not miss –by design– any original sample; also, all generalized samples contain just 1 location (hence fully retain the original data resolution) in the dataset with all trajectories and less than 4 locations for 60% of the samples in the dataset of unique raw trajectories.

Fig. 6b shows the temporal accuracy, defined as in previous experiments as the duration of the generalized samples. The accuracy attained by k -scale is between two and three orders of magnitude better than that of SafePath, independently of the privacy budget of the latter.

Utility of the k -anonymized data. To prove that k -scale-anonymized data retains sufficient utility, we conduct classical data mining tasks for the original *sen-1M* dataset and its k -anonymized versions, for $k = 10$. In tasks requiring single-location and precise-time representation, we approximate the geographical location of a generalized sample as the center of the area it spans and its occurrence time to the middle of the generalized interval. We focus our utility analysis on two seminal results derived from large-scale trajectory data about (i) the high predictability [29] and (ii) the mobility laws [30] that govern in human movement.

Fig. 7a shows that the entropy rate (H) of the probability of finding a particular time-ordered subsequence in the user trajectory (Π) is similar in the original and anonymized *sen-1M* and *civ* datasets. In Fig. 7b, we observe that the high theoretical predictability averaging around 95% found

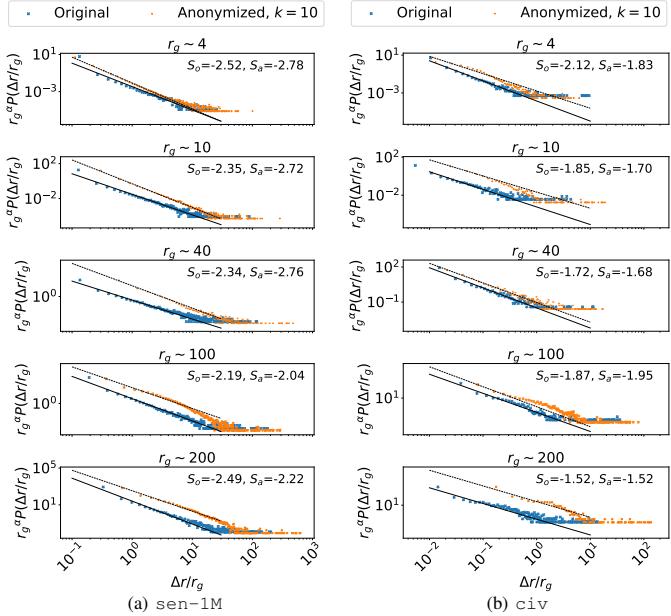


Fig. 8: Human mobility laws in original and k -anonymized ($k = 10$) data

in [29] is preserved by a k -anonymization with $k = 10$.

Fig. 8 illustrates the probability density function of individual travel distances $P(\Delta r / r_g)$ for users with a radius of gyration $r_g \sim 4, 10, 40, 100$, and 200 km with the parameter α set to 1.3 . We note that the normalized power-law distributions of displacements exhibit linear behavior, consistent with the findings of [30], with slopes showing a similar range of values round 2 across users with varying r_g as well as between the original (S_o) and anonymized (S_a) datasets. Specifically, the variance of S_o and S_a for each dataset remains relatively small across different values of r_g . Additionally, the difference between the slopes ($|S_o - S_a|$) is negligible.

We also assess the reliability of k -scale-anonymized data for calculations of the center of mass, radius of gyration, convex hull, home and work location of individual trajectories. Results are omitted due to space limitations but show again consistency between the original and k -anonymized data.

V. RELATED WORK

Our work addresses the privacy-preserving data publishing (PPDP) of spatiotemporal trajectory micro-data. This is not to be confused with the problem of ensuring privacy of positioning data in Location-Based Services (LBS) [31]–[41], where the aim is safeguarding queries to mobility databases and not the complete spatiotemporal trajectories that we target.

Approaches based on k -anonymity. Various techniques have been proposed for k -anonymizing spatiotemporal trajectories. Some solutions do not fulfill PPDP principles by creating fictitious mobility [42] or permuting samples across trajectories [43]. Others exclusively rely on limited approaches based exclusively on sample [44] or make strong assumptions on uniform and regular sampling of user mobility over time [45] that do not hold in practical databases. State-of-the-art-methods that can operate on practical trajectory data include TGA [46], W4M [18], and GLOVE [19]. As TGA is extremely expensive

form a computational viewpoint, we consider the last two models as a baseline for performance evaluation in §IV.

Approaches based on l -diversity or t -closeness. Advanced privacy principles build on k -anonymity to extend its robustness to attacks more complex than record linkage. The principle of l -diversity [47] assumes that a set of sensitive attributes is associated to each trajectory, and forces any trajectory to be indistinguishable from l others whose sensitive attributes are also different from those of the original trajectory. The t -closeness principle [48] extends the notion above, and ensures that there is no substantial statistical difference between the attribute values in every set of indistinguishable users and those in the whole user population. While these privacy criteria have been investigated for relational databases, only one study addressed them in the context of trajectory data [49]. The proposed solution uses GLOVE at its core. While it is beyond the scope of our paper to seek solutions for l -diversity or t -closeness, we argue that k -scale could, e.g., replace GLOVE as the core component of method above to enhance its accuracy and scalability. A similar consideration holds for a prior model targeting $k^{\tau, \epsilon}$ -anonymity of individual portions of a trajectory [24], which was developed on top of GLOVE and could be re-designed with k -scale at its center.

Approaches based on differential privacy. The vast majority of PPDP research based on the differential privacy principle for spatiotemporal trajectory data has focused on publishing aggregate statistics about the mobility patterns of the whole user population in a dataset, rather than the trajectories themselves. Solutions like DP quadtrees, spatiotemporal density, transit graphs, or histograms do not produce spatiotemporal trajectories [50]–[53]. Even in these settings, achieving user-level differential privacy is difficult: the privacy loss grows linearly with the dataset duration, which entails exponentially weaker guarantees and has raised significant questions on the practical viability of DP solutions in the literature [54].

Generating DP trajectory data is even more challenging. While well-understood mechanisms exist for differential privacy of counting data, datasets that are high-dimensional like spatiotemporal trajectories present unique problems. Recent studies highlight how DP struggles with correlations between spatiotemporal points, which weakens its guarantees and allows inferring sensitive locations [55]. Some existing approaches relax the requirements of actual differential privacy, e.g., allowing some spatiotemporal points to be disclosed as in (ϵ, δ) -differential privacy [56], which clearly weakens the privacy guarantees. Other solutions generate synthetic trajectories based on DP aggregate statistics computed from the original data, using techniques like prefix-trees [57], n-grams [57], spatial distributions [58], and transition probabilities [59]. While it preserves global properties of the dataset, the approach does not yield individual trajectories that represent real individual mobility [19]. State-of-the-art approaches model trajectories as a noisy prefix tree and publish ϵ -differentially-private trajectories [20], [60]. Among these, we consider SafePath [20] as a state-of-the-art benchmark in our performance evaluation.

VI. CONCLUSIONS

Our proposed solution, *k-scale*, advances the state of the art in the anonymization of trajectories, breaking through barriers such as scalability to millions of trajectories and preservation of the data spatiotemporal quality. Namely, *k-scale* stands as the first work not only to *k*-anonymize a million trajectories but to do so with a high value of $k = 10$ that is of practical use. This is a leap forward with respect to previous solutions that *k*-anonymize datasets of tens of thousands trajectories at most and with $k < 5$. We do not claim that *k-scale* is the ultimate solution to the problem of PPDP of trajectory data, as many open problems remain to be solved. Instead, *k-scale* is a cornerstone in the path toward such a definitive solution.

REFERENCES

- [1] V. D. Blondel *et al.*, “A survey of results on mobile phone datasets analysis,” *EPJ data science*, vol. 4, 2015.
- [2] D. Naboulsi *et al.*, “Large-scale mobile traffic analysis: a survey,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, 2015.
- [3] S. Wang *et al.*, “A survey on trajectory data management, analytics, and learning,” *ACM Comput. Surv.*, vol. 54, Mar. 2021.
- [4] S. Tong *et al.*, “Personalized mobile marketing strategies,” *J. of the Acad. Mark. Sci.*, vol. 48, 2020.
- [5] E. Parliament *et al.*, “Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec -general data protection regulation (gdpr).” <https://eur-lex.europa.eu/eli/reg/2016/679/oj>, 2016. Accessed: 2025-03-12.
- [6] L. Sweeney, “*k*-anonymity: A model for protecting privacy,” *International journal of uncertainty, fuzziness and knowledge-based systems*, vol. 10, no. 05, 2002.
- [7] A. Narayanan *et al.*, “Robust de-anonymization of large sparse datasets,” in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, IEEE, 2008.
- [8] H. Zang *et al.*, “Anonymization of location data does not work: A large-scale measurement study,” in *Proceedings of the 17th annual international conference on Mobile computing and networking*, 2011.
- [9] Y.-A. De Montjoye *et al.*, “Unique in the crowd: The privacy bounds of human mobility,” *Scientific reports*, vol. 3, no. 1, 2013.
- [10] L. Rossi *et al.*, “Spatio-temporal techniques for user identification by means of gps mobility data,” *EPJ Data Sci.*, vol. 4, no. 11, 2015.
- [11] A. Cecaj *et al.*, “Re-identification of anonymized cdr datasets using social network data,” in *2014 IEEE PerCom WORKSHOPS*, IEEE, 2014.
- [12] C. Riederer *et al.*, “Linking users across domains with location data: Theory and validation,” in *Proceedings of the 25th international conference on world wide web*, 2016.
- [13] D. Kondor *et al.*, “Towards matching user mobility traces in large-scale datasets,” *IEEE Transactions on Big Data*, vol. 6, no. 4, 2018.
- [14] H. Wang *et al.*, “De-anonymization of mobility trajectories: Dissecting the gaps between theory and practice,” in *NDSS’18*, 2018.
- [15] European Union, “General Data Protection Regulation, Recital 26 – Personal Data and Identifiability.” <https://eur-lex.europa.eu/eli/reg/2016/679/oj>, 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council.
- [16] M. Fiore *et al.*, “Privacy in trajectory micro-data publishing: a survey,” *Transactions on Data Privacy*, vol. 13, 2020.
- [17] B. C. Fung *et al.*, “Privacy-preserving data publishing: A survey of recent developments,” *ACM Computing Surveys (Csur)*, 2010.
- [18] O. Abul *et al.*, “Anonymization of moving objects databases by clustering and perturbation,” *Information systems*, vol. 35, no. 8, 2010.
- [19] M. Gramaglia *et al.*, “Glove: towards privacy-preserving publishing of record-level-truthful mobile phone trajectories,” *ACM/IMS Transactions on Data Science (TDS)*, vol. 2, no. 3, 2021.
- [20] K. Al-Hussaeni *et al.*, “Safepath: Differentially-private publishing of passenger trajectories in transportation systems,” *Computer Networks*, vol. 143, 2018.
- [21] R. Trujillo-Rasua *et al.*, “On the privacy offered by (k, δ) -anonymity,” *Information Systems*, vol. 38, no. 4, 2013.
- [22] F. Bonchi *et al.*, “Trajectory anonymity in publishing personal mobility data,” *ACM Sigkdd Explorations Newsletter*, vol. 13, no. 1, 2011.
- [23] D. Karapetyan *et al.*, “Local search heuristics for the multidimensional assignment problem,” *Journal of Heuristics*, vol. 17, 2011.
- [24] M. Gramaglia *et al.*, “Preserving mobile subscriber privacy in open datasets of spatiotemporal trajectories,” in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, IEEE, 2017.
- [25] S. M. Omohundro, *Five balltree construction algorithms*. International Computer Science Institute Berkeley, 1989.
- [26] J. L. Bentley, “Multidimensional binary search trees used for associative searching,” *Communications of the ACM*, vol. 18, no. 9, 1975.
- [27] V. D. Blondel *et al.*, “Data for development: the d4d challenge on mobile phone data,” *arXiv preprint arXiv:1210.0137*, 2012.
- [28] T. Yabe *et al.*, “Yjmob100k: City-scale and longitudinal dataset of anonymized human mobility trajectories,” *Scientific Data*, 2024.
- [29] C. Song *et al.*, “Limits of predictability in human mobility,” *Science*, vol. 327, no. 5968, 2010.
- [30] M. C. Gonzalez *et al.*, “Understanding individual human mobility patterns,” *nature*, vol. 453, no. 7196, 2008.
- [31] M. Gruteser *et al.*, “Anonymous usage of location-based services through spatial and temporal cloaking,” in *Proceedings of the 1st international conference on Mobile systems, applications and services*, 2003.
- [32] H. Kido *et al.*, “Protection of location privacy using dummies for location-based services,” in *ICDEW’05*, IEEE, 2005.
- [33] B. Gedik and L. Liu, “Protecting location privacy with personalized *k*-anonymity: Architecture and algorithms,” *IEEE Transactions on Mobile Computing*, vol. 7, no. 1, 2007.
- [34] M. Herrmann *et al.*, “Practical privacy-preserving location-sharing based services with aggregate statistics,” in *ACM WiSec*, 2014.
- [35] A. R. Beresford *et al.*, “Mix zones: User privacy in location-aware services,” in *IEEE PerCom workshops, 2004*, IEEE, 2004.
- [36] C. Kalaiarasy *et al.*, “Location privacy preservation in vanet using mix zones—a survey,” in *2019 ICCCI*, IEEE, 2019.
- [37] X. Ding *et al.*, “Privacy preserving similarity joins using mapreduce,” *Information Sciences*, vol. 493, 2019.
- [38] I. Memon *et al.*, “Dpmm: dynamic pseudonym-based multiple mix-zones generation for mobile traveler,” *Multimedia Tools and Applications*, vol. 76, 2017.
- [39] B. Niu *et al.*, “Achieving *k*-anonymity in privacy-aware location-based services,” in *IEEE INFOCOM 2014-IEEE conference on computer communications*, IEEE, 2014.
- [40] Z. Wu *et al.*, “Covering the sensitive subjects to protect personal privacy in personalized recommendation,” *IEEE Transactions on Services Computing*, vol. 11, no. 3, 2016.
- [41] Z. Wu *et al.*, “Constructing dummy query sequences to protect location privacy and query privacy in location-based services,” *World Wide Web*, vol. 24, 2021.
- [42] O. Abul *et al.*, “Never walk alone: Uncertainty for anonymity in moving objects databases,” in *2008 IEEE 24th international conference on data engineering*, Ieee, 2008.
- [43] J. Domingo-Ferrer *et al.*, “Microaggregation-and permutation-based anonymization of movement data,” *Information Sciences*, vol. 208, 2012.
- [44] B. C. Fung *et al.*, “Privacy protection for rfid data,” in *Proceedings of the 2009 ACM symposium on Applied Computing*, 2009.
- [45] R. Yarovoy *et al.*, “Anonymizing moving objects: How to hide a mob in a crowd?”, in *EDBT*, 2009.
- [46] M. E. Nergiz *et al.*, “Towards trajectory anonymization: a generalization-based approach,” in *ACM SIGSPATIAL*, 2008.
- [47] A. Machanavajjhala *et al.*, “*l*-diversity: Privacy beyond *k*-anonymity,” *ACM TKDD*, vol. 1, no. 1, 2007.
- [48] N. Li *et al.*, “*t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity,” in *IEEE ICDE*, IEEE, 2006.
- [49] Z. Tu *et al.*, “Protecting trajectory from semantic attack considering *k*-anonymity, *l*-diversity, and *t*-closeness,” *IEEE Transactions on Network and Service Management*, vol. 16, no. 1, 2018.
- [50] G. Cormode *et al.*, “Differentially private spatial decompositions,” in *2012 IEEE 28th International Conference on Data Engineering*, IEEE.
- [51] G. Acs *et al.*, “A case study: Privacy preserving release of spatio-temporal density in paris,” in *ACM SIGKDD*, 2014.
- [52] M. Alaggan *et al.*, “Sanitization of call detail records via differentially private bloom filters,” in *DBSec 2015*, Springer, 2015.

- [53] S. Brunet *et al.*, “Novel differentially private mechanisms for graphs,” *Cryptology ePrint Archive*, 2016.
- [54] F. Houssiau *et al.*, “On the difficulty of achieving differential privacy in practice: user-level guarantees in aggregate location data,” *Nature communications*, vol. 13, no. 1, 2022.
- [55] Miranda-Pascual *et al.*, “Sok: Differentially private publication of trajectory data,” *Proceedings on Privacy Enhancing Technologies*, 2023.
- [56] D. Shao *et al.*, “Publishing trajectory with differential privacy: A priori vs. a posteriori sampling mechanisms,” in *DEXA 2013*, Springer, 2013.
- [57] R. Chen *et al.*, “Differentially private transit data publication: a case study on the montreal transportation system,” in *ACM SIGKDD*, 2012.
- [58] D. J. Mir *et al.*, “Dp-where: Differentially private modeling of human mobility,” in *2013 IEEE international conference on big data*, IEEE.
- [59] M. E. Gursoy *et al.*, “Differentially private and utility preserving publication of trajectory data,” *IEEE Transactions on Mobile Computing*, vol. 18, no. 10, 2018.
- [60] C. Chen *et al.*, “Optimization of privacy budget allocation in differential privacy-based public transit trajectory data publishing for smart mobility applications,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 12, 2023.