Privacy Preserving Location Data Publishing: A Machine Learning Approach

Sina Shaham, Ming Ding, Bo Liu, Shuping Dang, Zihuai Lin, and Jun Li

Abstract—Publishing datasets plays an essential role in open data research and promoting transparency of government agencies. However, such data publication might reveal users' private information. One of the most sensitive sources of data is spatiotemporal trajectory datasets. Unfortunately, merely removing unique identifiers cannot preserve the privacy of users. Adversaries may know parts of the trajectories or be able to link the published dataset to other sources for the purpose of user identification. Therefore, it is crucial to apply privacy preserving techniques before the publication of spatiotemporal trajectory datasets. In this paper, we propose a robust framework for the anonymization of spatiotemporal trajectory datasets termed as machine learning based anonymization (MLA). By introducing a new formulation of the problem, we are able to apply machine learning algorithms for clustering the trajectories and propose to use k-means algorithm for this purpose. A variation of k-means algorithm is also proposed to preserve the privacy in overly sensitive datasets. Moreover, we improve the alignment process by considering multiple sequence alignment as part of the MLA. The framework and all the proposed algorithms are applied to T-Drive, Geolife, and Gowalla location datasets. The experimental results indicate a significantly higher utility of datasets by anonymization based on MLA framework.

Index Terms—k-anonymity, spatiotemporal trajectories, longitudinal dataset, machine learning, privacy preservation.

1 Introduction

Publication of data by different organizations and institutes is crucial for open research and transparency of government agencies. Just in Australia, since 2013, over 7000 additional datasets have been published on 'data.gov.au,' a dedicated website for the publication of datasets by the Australian government. Moreover, the new Australian government data sharing legislation encourage government agencies to publish their data, and as early as 2019, many of them will have to do so [2]. Unfortunately, the process of data publication can be highly risky as it may disclose individuals' sensitive information. Hence, an essential step before publishing datasets is to remove any uniquely identifiable information from them. However, such an operation is not sufficient for preserving the privacy of users. Adversaries can re-identify individuals in datasets based on common attributes called quasi-identifiers or may have prior knowledge about the trajectories traveled by the users. Such side information enables them to reveal sensitive information that can cause physical, financial, and reputational harms to people.

One of the most sensitive sources of data is location trajectories or spatiotemporal trajectories. Despite numerous use cases that the publication of spatiotemporal data can provide to users and researchers, it poses a significant threat to users' privacy. As an example, consider a person who has been using GPS navigation to travel from home to work every morning of weekdays. If an adversary has some prior knowledge about a user, such as the home address, it is possible to identify the user. Such an inference attack can compromise user privacy, such as revealing the user's health condition and how often the user visits his/her medical specialist. Therefore, it is crucial to anonymize spatiotemporal datasets before publishing them to the public. The privacy issue gets even more severe if the adversary links identified users to other databases, such as the database of medical records. That is the very reason why nowadays most companies are reluctant to publish any spatiotemporal trajectory datasets without applying an effective privacy preserving technique.

A widely accepted privacy metric for the publication of spatiotemporal datasets is k-anonymity. This metric can be summarized as ensuring that every trajectory in the published dataset is indistinguishable from at least $k-1$ other trajectories. The authors in [3], adopted the notion of k-anonymity for spatiotemporal datasets and proposed an anonymization algorithm based on generalization. Xu et al. [4] investigated the effects of factors such as spatiotemporal resolution and the number of users released on the anonymization process. Dong et al. [5] focused on improving the existing clustering approaches. They proposed an anonymization scheme based on achieving k-anonymity by grouping similar trajectories and removing the highly dissimilar ones. More recently, the authors in [6] developed an algorithm called k-merge to anonymize the trajectory datasets while preserving the privacy of users from probabilistic attacks. Local suppression and splitting

This work was submitted in part and accepted to appear in the proceedings of INFOCOM WORKSHOPS, 2019 [1].

S. Shaham and Z. Lin are with the Department of Engineering, The University of Sydney, Sydney, NSW, 2006 Australia (e-mail: sina.shaham, zihuai.lin}@sydney.edu.au).

M. Ding is with Data61, Sydney, NSW, 1435 Australia (email: ming.ding@data61.csiro.au)

B. Liu is with University of Technology Sydney, NSW 2007, Australia (email: bo.liu@uts.edu.au)

S. Dang was with the R&D Center, Guangxi Huanan Communication Co., Ltd., Nanning 530007, China when completing the major work of this paper, and is now with Computer, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia (e-mail: shuping.dang@kaust.edu.sa).

J. Li is with NJUST, Nanjing, China (email: jleesr80@gmail.com)

techniques were also considered to protect privacy in [7].

However, there are three major problems with the aforementioned approaches.

- Lack of a well-defined method to cluster trajectories as there is not an easy way to measure the cost of clustering when considering the distances among trajectories rather than simply the locations.
- The existing literature focuses on pairwise sequence alignment, which results in a high amount of information loss [3], [6], [8]–[10].
- There is no unified metric to evaluate and compare the existing anonymization methods.

In this paper, we address the mentioned problems by proposing an enhanced anonymization framework termed machine learning based anonymization (MLA) to preserve the privacy of users in the publication of spatiotemporal trajectory datasets. MLA consists of two interworking algorithms: clustering and alignment. We have summarized our main contributions in the following bullet points.

- By formulating the anonymization process as an optimization problem and finding an alternative representation of the system, we are able to apply machine clustering algorithms for clustering trajectories. We propose to use k'-means [1] algorithm for this purpose, as part of the MLA framework.
- We propose a variation of k'-means algorithm to preserve the privacy of users in the publication of overly sensitive spatiotemporal trajectory datasets.
- We enhance the performance of sequence alignment in clusters by considering multiple sequence alignment instead of pairwise sequence alignment.
- We propose a utility metric to evaluate and compare the anonymization frameworks.

MLA and all algorithms associated with it are applied on two real-life GPS datasets following different distributions in time and spatial domains. The experimental results indicate a significantly higher utility levels while maintaining k-anonymity of trajectories.

The rest of this paper is organized as follows. First, a comprehensive review of the currently existing literature is presented in Section 2, followed by the system model used in Section 3. Next, the proposed framework is explained and analyzed in Sections 4 and 5, respectively. Several real-world applications of the framework are elaborated in Section 6, and finally, the paper is concluded in Section 7.

2 Related Work

Unfortunately, merely removing unique identifiers of users cannot protect their privacy, as databases can be linked to each other based on their quasi-identifiers. Doing so, adversaries can reveal sensitive information about the users and compromise their privacy. In this section, we review the existing approaches for the anonymization of spatiotemporal datasets.

1. The prime notation on the top of variable "k" is to distinguish between the variable k in the clustering algorithm and the variable k used in the definition of k-anonymity.

2.1 Generalization Technique

Generalization is currently one of the mainstream approaches for the anonymization of spatiotemporal trajectory datasets. The generalization technique is predicated on two interrelated mechanisms: clustering and alignment. Clustering aims at finding the best grouping of trajectories that minimizes a predefined cost function, and the alignment process aligns trajectories in each group.

The notion of k-anonymity was adopted in [8] for anonymization of spatiotemporal datasets . The authors proved that the anonymization process is NP-hard and followed a heuristic approach to cluster the trajectories. The use of 'edit distance' metric for anonymization of spatiotemporal datasets was proposed in [9]. In this work, the authors target grouping the trajectories based on their similarity and choose a cluster head for each cluster to represent the cluster. Also, dummy trajectories were added to anonymize the datasets further. Yarovoy et al. [10] proposed to use Hilbert indexing for clustering trajectories. The authors in [5], [11] chose to avoid alignment by selecting trajectories with the highest similarity as representatives of clusters. Poulis et al. [12] investigated applying restriction on the amount of generalization that can be applied by proposing a user-defined utility metric. Takahashi et al. [13] proposed an approach termed as CMAO to anonymize the real-time publication of spatiotemporal trajectories. The proposed idea is based on generalizing each queried location point with $k-1$ other queried location by other users, and hence, achieving k-anonmity.

The current state-of-art technique for applying gerelization to spatiotemporal datasets is based on generalization hierarchy (DGH) trees. In essence, DGH can be seen as a coding scheme to anonymize trajectories. We have categorized types of DGHs in the literature as:

- Full-domain generalization: This technique emphasizes on the level that each value of an attribute is located in the generalization tree. If a value of an attribute is generalized to its parent node, all values of that attribute in the dataset must be generalized to the same level [14]–[16].
- Subtree generalization: In this method, if a value of an attribute is generalized to its parent node, all other child nodes of that parent node need to be replaced with the parent node as well [17], [18].
- Cell generalization: This generalization technique considers each cell in the table separately. One cell can be generalized to its parent node while other values of that attribute remain unchanged [19]–[21].

2.2 Other Anonymization Techniques

Aside from the generalization technique, we have categorized the existing methods for the anonymization spatiotemporal datasets into three major groups:

- **Perturbation** anonymizes location datasets by addition of noise to data;
- **ID swapping** swaps user IDs in road junctions to anonymize location datasets;
- **Splitting** divides trajectories into shorter lengths to anonymize location datasets.

The authors in [22] proposed an algorithm that swaps the IDs of users in trajectories once they reach an intersection. Doing so, the algorithm prevents adversaries from identifying a particular user. Cicek et al. [23] made a distinction between sensitive and insensitive location nodes of trajectories. Their proposed algorithm only groups the paths around the sensitive nodes and exploits generalization to create supernodes.

Moreover, Cristina et al. [24] shifted the burden of privacy preservation in data publishing to the user side. The authors attempted to anonymize the data on the mobile phones before storage on the database as they would have more control over their privacy. Instead of clustering trajectories for anonymization, Cicek et al. in [23] focused on the obfuscation of underlying map for sensitive locations. Brito et al. [25] minimized the information loss during the data anonymization by suppressing key locations. The Local suppression and splitting techniques were considered for trajectory anonymization in [7]. Although the proposed approach is useful for a predefined number of locations, it cannot be generalized to system models in which the users can make queries from an arbitrary location on the map. Naghizadeh et al. [26] focused on the stop points along trajectories. A sensitivity measure is introduced in this work, which relies on the amount of time users spend in different locations. Sensitive locations are replaced or displaced with a less sensitive location to preserve the privacy of users. Jiang et al. [27] considered the perturbation of locations by adding noise to preserve the privacy of users. Adding noise can generate fake trajectories that do not correspond to realistic scenarios.

3 System Model

We assume that a map has been discretized into an $\epsilon \times \epsilon$ grid and the time is discretized into bins with length ϵ_t. Therefore, each point in the dataset represents a snapshot of a real-world location query including x-coordinate, y-coordinate, and time. The datasets with continuous time or space data can fit into our model using interpolation. The level of spatial-temporal granularity in discretization does not affect the effectiveness of the proposed model. In our model, we consider a spatiotemporal trajectory datasets denoted by T. The dataset consists of trajectories $tr_1, ..., tr_n$ where n represents the number of trajectories in the dataset ($T = \{tr_1, ..., tr_n\}$, $|T| = n$). The i-th trajectory tr_i is an ordered set of l_i spatiotemporal 3D points (i.e., $tr_i = \{p_1, ..., p_{l_i}\}$, $|tr_i| = l_i$). Each point p_j is defined by a triplet $< x_j, y_j, t_j >$, where x_j, y_j, t_j indicate the x-coordinate, y-coordinate, and the time of query, respectively.

3.1 Generalization Model

Our proposed framework is based on the generalization technique to anonymize the spatiotemporal datasets. To apply this technique, we use the domain generalization hierarchy (DGH) trees and quantify the information loss accordingly.

3.1.1 Domain Generalization Hierarchies

DGH tree is defined formally in Definition 1. To clarify the construction of DGHs, an example of such a tree for spatiotemporal datasets is provided in Example 1. In our model, we utilize three dimensions: x-coordinate, y-coordinate, and the time of queries in hours.

Definition 1. A DGH tree for an attribute $\mathcal{A}$, denoted as $H_{\mathcal{A}}$, is a partially ordered tree structure, which maps specific and generalized values of the attribute $\mathcal{A}$. The root of the tree is the most generalized value and is returned by the function RT.

Example 1. Consider an 4×8 map shown in Example 1. As can be seen in the figure, the generalization technique is applied by three DGH trees, each of them corresponding to one of the attributes. For instance, the x-coordinate attribute can have 8 possible values $(0, 1, ..., 7)$. At the lowest level of the tree, each coordinate needs three bits of information to be shown that indicates the maximum information bits. As we go higher up the DGH tree, more information loss incurs, and less number of bits are used to represent the coordinates.

Each node on a DGH tree can be generalized by moving up one or multiple levels of the DGH. The process of generalizing node_i to one of its parent nodes node_j is denoted using $\text{node}_i \rightarrow \text{node}_j$. A special case of generalization, in which the node is generalized to the root of the DGH, is referred to as suppression.

For generalizing two nodes, it is necessary to find the lowest common ancestor (LCA). The LCA is a critical point in the generalization process due to its corresponding subtree that entails both the nodes and achieves the lowest information loss for the generalization of two nodes. The definition of LCA is given in Definition 2.

Definition 2. The LCA of node_i and node_j in $H_{\mathcal{A}}$ is defined as the lowest common parent root of the two nodes. Function LCA returns the LCA.

For instance, in Example 1, if two leaf nodes '000' and '010' are to be generalized, their LCA corresponds to the parent node '0'. Hence, in the dataset, the x-coordinates '000' and '010' will be replaced by '0' to prevent adversaries from distinguishing between these two nodes.

3.1.2 Information Loss

The information loss incurred by generalizing node_i to node_j in DGH $H_{\mathcal{A}}$ is defined as

$$LS(\text{node}_i, \text{node}_j) = \log_2 LF(\text{node}_j) - \log_2 LF(\text{node}_i) \text{ bits}, \quad (1)$$

where $LF(.)$ function returns the number of leaves in the subtree generated by a node, and $LS(.)$ function returns the loss incurred by the generalization of nodes. The calculation of information loss is elaborated in Example 2.

Example 2. Consider the x-coordinate DGH tree given in Fig. 1, the information loss incurred by generalizing node '10' to '1' can be calculated as $\log_2 4 - \log_2 2 = 1$ bits.

Moreover, Lemma 1 can be used to derive the total loss incurred by the generalization of two nodes to their LCA.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964658, IEEE Transactions on Knowledge and Data Engineering

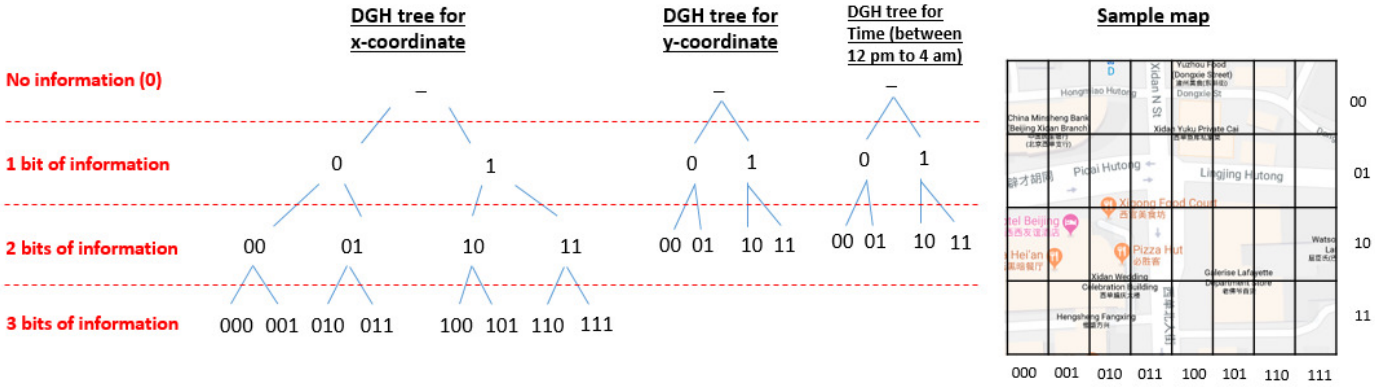


Figure 1: An example of DGHs for the attributes of spatiotemporal datasets.

Lemma 1. *The total loss incurred by generalizing $node_i$ and $node_j$ in $H_{\mathcal{A}}$ to their LCA, $node_p$, can be calculated as*

$$LS(node_i + node_j, node_p) = LS(node_i, node_p) + LS(node_j, node_p). \quad (2)$$

The total loss incurred during anonymization of a trajectory and a dataset are defined in Definitions 3 and 4, respectively.

Definition 3. The total loss rendered by the generalization of trajectory tr to achieve the anonymized trajectory $\overline{tr}$ with respect to attribute $\mathcal{A}$ can be calculated as

$$LS(\overline{tr}, \mathcal{A}) = \sum_{i=1}^{|\overline{tr}|} LS(tr_i.\mathcal{A}, \overline{tr}_i.\mathcal{A}). \quad (3)$$

where $tr_i.\mathcal{A}$ indicates the i-th location of the trajectory tr with respect to the attribute $\mathcal{A}$. Here, $\mathcal{A}$ could denote x-coordinate, y-coordinate, or time.

Definition 4. The total loss with respect to an attribute $\mathcal{A}$ in an anonymized dataset $\overline{T}$ can be computed as

$$LS(\overline{T}, \mathcal{A}) = \sum_{\overline{tr} \in \overline{T}}^{|\overline{T}|} LS(\overline{tr}, \mathcal{A}) \quad (4)$$

3.2 Privacy Model

3.2.1 Adversary Model

In our work, we consider coordinates and the time of queries both to be quasi-identifiers, as they can be linked to other databases and compromise the privacy of users. We also assume that no uniquely identifiable information is released while publishing the dataset. However, the adversary may:

- already know about part of the released trajectory for an individual and attempt to identify the rest of the trajectory. For instance, the adversary is aware of the workplace of an individual and attempts to identify his or her home address.
- already know the whole trajectory that an individual has traveled, but try to access other information released while publishing the dataset by identifying the user in the dataset. For instance, the published dataset may also include the type of services provided to users and if the adversary can identify a user by its trajectory, it can also know the services provided to that user.

To this end, our aim is to protect users against the adversary's attempt to access sensitive information that may compromise user privacy.

3.2.2 Privacy Metric

In this paper, we use a well-known metric called k-anonymity [28] to ensure the privacy of users. The k-anonymity in our dataset implies that a given trajectory in the original dataset can at best be linked to $k-1$ other trajectories in the anonymized dataset. Definition 5 formally defines the k-anonymity in the context of dataset.

Definition 5. *k-anonymous dataset:* A trajectory dataset $\overline{T}$ is a k-anonymization of a trajectory dataset T if for every trajectory in the anonymized dataset $\overline{T}$, there are at least $k-1$ other trajectories with exactly the same set of points, and there is a one to one mapping relation between the trajectories in $\overline{T}$ and T.

3.2.3 Spatial Utility Metric

The k-anonymity metric ensures that the users are k-anonymous, implying that they cannot be identified from at least $k-1$ other users in the anonymized published dataset. To achieve k-anonymity, significant loss of information incurs during the generalization process of different algorithms. However, there is no unified metric to measure how much information has actually been lost to achieve k-anonymity. The metric explained in Section 3.1.2 is a viable option for this purpose and can be used to measure the amount of information lost based on different algorithms. However, there are two major drawbacks associated with the metric: (I) It is only applicable when the algorithms are using DGH trees as their principle encoding scheme. (II) The DGH tree used in the algorithms must be identical to generate comparable results. These two drawbacks significantly limit the practicality of the information loss metric for the purpose of comparison among various developed approaches. To address this challenge, we propose to use

the average released area per location to assess and compare the anonymization schemes.

Any anonymization approach aims to maximize utility while preserving the privacy of users. Utility in generalization techniques refers to the area released for locations in the dataset. Consider a location in the dataset T with coordinates $< x_1, y_1, t_1 >$ and an arbitrary generalization function $\mathcal{F} : T \rightarrow \overline{T}$. After the anonymization process, $< x_1, y_1, t_1 >$ is generalized with a number of other locations $< x_2, y_2, t_2 >$,..., $< x_a, y_a, t_a >$ in the dataset and an area S would be released representing these locations. For instance, if generalization returns the minimum rectangle surrounding the locations, the generalized area is given by:

$$S = (\max_i\{x_i\} - \min_i\{x_i\}) \times (\max_i\{y_i\} - \min_i\{y_i\}). \quad (5)$$

Once the anonymization is conducted, assume that n_1 locations are generalized to area S_1, n_2 locations are generalized to area S_2,..., n_b locations are generalized to area S_b. In this case, the average released area per location can be calculated as

$$(\sum_{i=1}^{b} n_i \times S_i)/(\sum_{i=1}^{b} n_i), \quad (6)$$

in which no location belongs to more than one area. Average released area per location helps to understand how efficiently the data has been generalized and how much loss of spatial utility has occurred by the generalization. Having k-anonymous locations, a smaller released area per location indicates a higher spatial utility of data while preserving the privacy of users.

3.3 Problem Formulation

The problem we seek to answer in this paper is formally presented in Problem 1 as follows.

Problem 1. *Given a trajectory dataset T, a privacy requirement k, quasi-identifiers x-coordinate, y-coordinate, and time, how to generate an anonymized dataset $\overline{T}$ which achieves the k-anonymity privacy metric and minimizes the total loss with respect to all quasi-identifiers, which can be explicitly formulated as*

$$Minimize\{LS(\overline{T}, x) + LS(\overline{T}, y) + LS(\overline{T}, t)\}. \quad (7)$$

4 MLA

In this section, we present our proposed framework, MLA, for anonymization of spatiotemporal datasets.

4.1 Overview of the MLA Framework

Fig. 2 demonstrates the overview of our proposed framework. The original dataset and the value of k are the inputs of the framework, and the output is the anonymized dataset preserving the privacy of users. The MLA framework consists of three mechanisms working together to anonymize spatiotemporal datasets, i.e., clustering, alignment, and generalization. A short description of each mechanism is provided as follows.

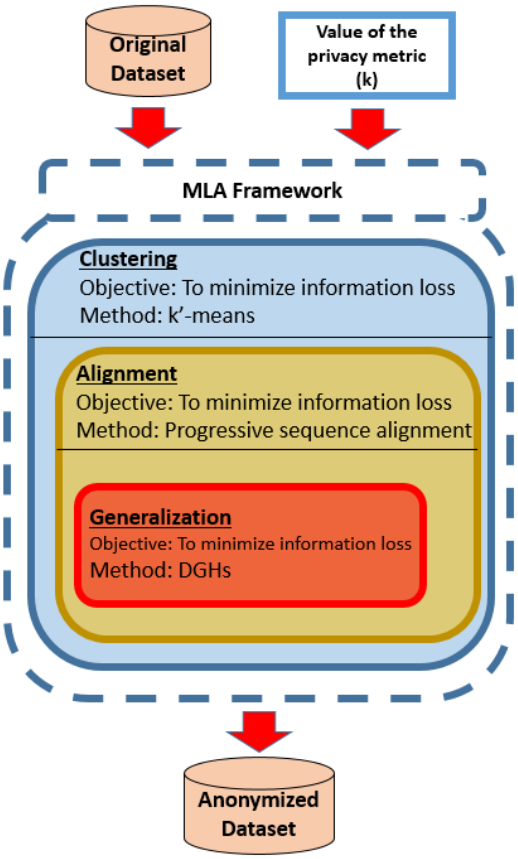


Figure 2: Overview of our proposed MLA framework.

- **Clustering:** At the highest level of the MLA framework, clustering is applied to seek for the most suitable grouping of trajectories that minimizes information loss. We propose to use k'-means clustering algorithm and a variation of it for overly sensitive datasets. Moreover, to have a baseline for comparison purposes, we develop a heuristic approach to cluster datasets. Our proposed clustering approaches are elaborated in Section 4.3.
- **Alignment:** For a given trajectory cluster, we propose to use progressive sequence alignment to find the arrangement of trajectories that results in the minimum information loss. Our approach for the alignment of trajectories is explained in Section 4.2.
- **Generalization:** At the heart of MLA framework resides the generalization approach. The generalization process is conducted based on DGHs explained in Section 3.1.

Note that these mechanisms are not independent of each, and they all work together with the objective to minimize the incurred information loss. The information loss incurred as a result of an arbitrary clustering of trajectories can only be known if the alignment and the generalization are applied in each cluster.

4.2 Alignment

The process of alignment is defined as finding the best match between two trajectories in order to minimize the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964658, IEEE Transactions on Knowledge and Data Engineering

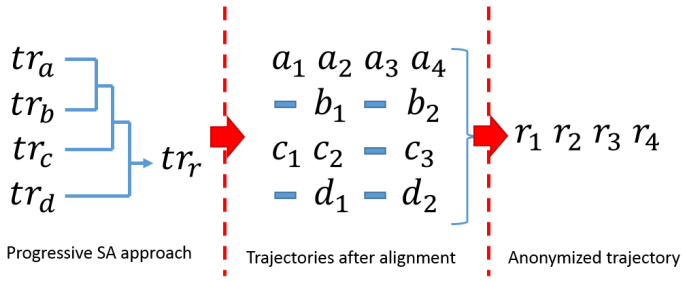


Figure 3: An overview of progressive SA for alignment of four trajectories and generating the anonymized trajectory.

overall cost of generalization and suppression. The process of alignment between two trajectories has been studied in different domains mostly referred to as sequence alignment (SA). In this paper, we adopt a multiple SA technique called progressive SA [29] for anonymization of spatiotemporal trajectories.

4.2.1 Progressive Sequence Alignment

The progressive SA is commonly used for SA of a set of protein sequences. Progressive SA is a greedy approach for multiple SA. As a part of the algorithm, pairwise alignment of the trajectories is required. We use dynamic SA for this purpose. Dynamic SA is based on dynamic programming and commonly used in DNA SA [30], [31]. Fig. 3 illustrates an example of how the progressive SA works for four hypothetical sequences $tr_a = \{a_1, a_2, a_3, a_4\}$, $tr_b = \{b_1, b_2\}$, $tr_c = \{c_1, c_2, c_3\}$ and $tr_d = \{d_1, d_2\}$ to generate the resultant aligned trajectory $tr_r = \{r_1, r_2, r_3, r_4\}$. The longest path tr_a is chosen as the basis and it is aligned with a randomly chosen trajectory tr_b. The pairwise alignment process is implemented using dynamic SA. Then, the resultant trajectory is aligned with a third trajectory. The process continues until all trajectories are aligned. Instead of choosing the trajectories randomly during the progressive SA, the algorithm can choose the trajectory resulting in the lowest loss during the alignment. In Fig. 3, the way trajectory elements are located with respect to the longest path is referred to as the structure of the shorter path, and also, the spaces indicate the suppression operation during the alignment.

The dynamic SA algorithm is formally represented in Algorithm 1. Dynamic SA is based on dividing the problem of finding the best SA to subproblems and storing the solutions of subproblems in a table or matrix referred to as $SAmatrix$ in the pseudocode. The objective is to achieve the minimal cost for SA. As before, the cost of alignment refers to the loss incurred during the alignment for different attributes of the sequence, which are x-coordinate, y-coordinate, and the time of the query.

The algorithm starts by creating a $(m+1)\times(n+1)$ matrix ($SAmatrix$), where m and n denote the length of the trajectories. The matrix will be used to store the minimum cost of each cell of the grid. Moreover, a list called $code$ stores how cells have been reached. Cell $[j+1, i+1]$ can be reached from three cells $[j, i+1]$, $[j+1, i]$, $[j, i]$. Each path corresponds to one of the subproblems explained. After finding all values

Algorithm 1: DynamicSA(tr_1, tr_2, H_x, H_y, H_t).

Required variables: $tr_1 = \{p_1, , p_2, ..., p_m\}$, $tr_2 = \{q_1, , q_2, ..., q_n\}$, H_x, H_y, H_t
$SAmatrix \leftarrow$ np.zeros($[m+1, n+1]$)
for i *in range*(m) **do**
  $Loss \leftarrow LS(p_i.x, rt(H_x)) + LS(p_i.y, rt(H_x)) + LS(p_i.t, rt(H_t))$
  $SAmatrix[i+1, 0] \leftarrow SAmatrix[i, 0] + Loss$
end
for i *in range*(n) **do**
  $Loss \leftarrow LS(q_i.x, rt(H_x)) + LS(q_i.y, rt(H_x)) + LS(q_i.t, rt(H_t))$
  $SAmatrix[0, i+1] \leftarrow SAmatrix[0, i] + Loss$
end
$options \leftarrow$ np.zeros(3)
$code \leftarrow$ list()
for i *in range*(m) **do**
  for j *in range*(n) **do**
    $Loss \leftarrow$ loss incurred by generalizing p_i and q_j
    $options[0] \leftarrow SAmatrix[i, j] + Loss$
    $Loss \leftarrow$ loss incurred by suppressing q_j
    $options[1] \leftarrow SAmatrix[i+1, j] + Loss$
    $Loss \leftarrow$ loss incurred by suppressing p_i
    $options[2] \leftarrow SAmatrix[i, j+1] + Loss$
    $BestOption \leftarrow$ np.argmin $(options)$
    $SAmatrix[i+1, j+1] \leftarrow options[BestOption]$
    $code$.append(index of option with minimum value)
  end
end
$TotLoss \leftarrow SAmatrix[m, n]$
$GenTraj \leftarrow$ trace back the $code$ to generate the aligned trajectory
$ShoTrajStr \leftarrow$ trace back the $code$ to find out structure of shorter trajectory while alignment
Return $GenTraj, ShoTrajStr, TotLoss$

of the matrix and tracing back the list $code$, the outputs of the algorithm are the value of cell $[m, n]$ indicating the minimum value of the total loss ($TotLoss$) required for the dynamic SA, the aligned trajectory ($GenTraj$), and the structure of the shorter path compared to the longer path as $ShoTrajStr$.

4.3 Clustering

Clustering can be seen as a search for hidden patterns that may exist in datasets. In simple words, it refers to grouping data entries in disjointed clusters so that the members of each cluster are very similar to each other. Clustering techniques are applied in many application areas, such as data analysis and pattern recognition. There are three clustering approaches considered in this work, i.e., heuristic, k'-means, and iterative k'-means. The latter two algorithms are our proposed approaches to significantly improve the utility of published spatiotemporal datasets. The heuristic algorithm is presented for the purpose of comparison. Each one of these approaches works independently and can be

Algorithm 2: HeuristicClustering($OriginalDataset$, k).

```
NumOfClus ← ⌈|T| / k⌉
T ← OriginalDataset
Let Clusters be a two-dimensional array storing the clusters and their corresponding trajectories
for c in range(0, NumOfClus) do
    Select a trajectory randomly from T and append it to cluster[c] while removing it from T
    for i in range(1, k) do
        for j in range(1, |T|) do
            Add the trajectory to the cluster cluster[c]
            Align based on DynamicSA and store the information loss
            Remove the trajectory from cluster[c]
        end
        Append the trajectory resulting in the minimum loss to cluster[c] and remove it from the dataset
    end
end
(T̄, Loss) ←GenerateAnonymizedDataset(cluster, OriginalDataset)
Return (T̄, Loss)
```

Algorithm 3: GenerateAnonymizedDataset($cluster$, $OriginalDataset$).

```
Let the TotalLoss store the total loss incurred by applying progressive SA
Let T̄ be an empty set that will store the anonymized dataset
for i in range(0, len(cluster)) do
    Apply progressive SA on trajectories in cluster[i]
    Add the incurred loss to TotalLoss
    Append the generated trajectory to T̄
end
Return (T̄, TotalLoss)
```

embedded in the MLA framework to cluster trajectories. A short description of these algorithms is provided as follows.

- The heuristic algorithm is a widely used scheme in the literature [3], [8]. This algorithm is often applied for optimizing different objective functions, however, with a similar structure. We have used this approach as a benchmark to compare our proposed algorithms.
- The k'-means algorithm is our proposed scheme for clustering the spatiotemporal trajectories, which can significantly improve the utility of published spatiotemporal datasets. The algorithm provides robust performance, but some of the users may not achieve k-anonymity due to the possibility that some of the clusters may include less than k trajectories.
- The iterative k'-means algorithm is a variation of the k'-means algorithm, we have proposed to address the privacy issue for overly sensitive datasets. This approach guarantees privacy requirements for all users with the cost of higher information loss compared with the k'-means algorithm.

4.3.1 Heuristic Approach

The heuristic approach for clustering spatiotemporal trajectory datasets is detailed in Algorithm 2 and its helper function in Algorithm 3. The intuition behind the heuristic algorithm is to form the clusters by sequentially adding the most suitable trajectory that minimizes the total loss incurred by generalization and suppression for x-coordinate, y-coordinate, and the time of query, given their DGHs H_x, H_y, H_t.

The algorithm starts by calculating the number of clusters that need to be generated and making a duplicate of the dataset called T. Moreover, a two-dimensional list is created, which holds the trajectory IDs for each cluster. For each cluster (i.e., cluster c), the algorithm appends a random trajectory from T. This trajectory is removed for T and would be the first member of the cluster c. Then, given the privacy requirement k, $k-1$ other members of the cluster are chosen in a greedy approach. For every remaining trajectory in the dataset, the algorithm calculates the information loss incurred by applying dynamic alignment and determine the trajectory that results in the minimum loss. The chosen trajectory will be added to the cluster and removed from the dataset. The process continues until all members of the cluster are chosen. After clustering the trajectories, the helper function GenerateAnonymizedDataset is called in order to generate the anonymized dataset ($\overline{T}$) and the total incurred loss.

The helper function (GenerateAnonymizedDataset) takes the original dataset and the two-dimensional list of clusters as inputs. The target of the algorithm is to find the total loss and anonymize the dataset. The algorithm starts by initializing the total loss to zero and creating an empty list ($\overline{T}$) to hold the generated anonymized dataset. Then, for each cluster, the progressive SA is applied to calculate the incurred loss in addition to the generalized trajectory. In the next step, the total loss is accumulated, and the generalized trajectory is appended to the anonymized dataset $\overline{T}$. Eventually, the anonymized dataset and the overall information loss happened due to alignment are returned.

4.3.2 k'-means Clustering Approach

k'-means algorithm [32] is an attractive clustering algorithm currently used in many applications, especially in data analysis and pattern recognition [33]. The main advantage of the k'-means algorithm is simplicity and fast execution.

The algorithm aims to partition the input dataset into k' clusters. The only inputs to the algorithm are the number of clusters k' and the dataset. Clusters are represented by adaptively-changing cluster centres. The initial values of the cluster centres are chosen randomly. In each stage, the algorithm computes the Euclidean distance of data from the centroids and partition them based on the nearest centroid to each data. More formally, representing the set of all centroids by $C = \{c_1, c_,..., c_{k'}\}$, each point in the dataset, denoted by x, is assigned to a centroid that has the shortest Euclidean distance to the point. This can be written as

$$\underset{c_i \in C}{\operatorname{argmin}}\, dist(x, c_i)^2, \tag{9}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964658, IEEE Transactions on Knowledge and Data Engineering

$$\text{Total loss} = \underbrace{\sum_{i=1}^{|T|}(LS(tr_i.x, RT(H_x)) + LS(tr_i.y, RT(H_y)) + LS(tr_i.t, RT(H_t)))}_{\text{A}} - \underbrace{(\sum_{i=1}^{|cluster||cluster[i]|}\sum_{j=1}(LS(h_j.x, RT(H_x)) + LS(h_j.y, RT(H_y)) + LS(h_j.t, RT(H_t))))}_{\text{B}}. \quad (8)$$

where the function $dist(.)$ returns the Euclidean distance between two points. Denoting the set of assigned data to the i-th cluster by S_i, new centroids are calculated in the second stage via

$$c_i = \frac{1}{|S_i|}\sum_{x_i \in S_i} x_i. \quad (10)$$

The algorithm continues the same process until the values of centroids no longer change. The k'-means algorithm is guaranteed to converge [34].

In the rest of this section, we first present a Lemma followed by explaining how the k'-means algorithm can be applied to trajectory datasets to reinforce the privacy preservation of users.

Lemma 2. *The total loss incurred by generalizing node$_i$ and node$_j$ with respect to $H_{\mathcal{A}}$ can be calculated as*

$$LS(node_i, node_j) = |LS(node_i, RT(H_{\mathcal{A}})) - LS(node_j, RT(H_{\mathcal{A}}))|. \quad (11)$$

Example 3. Lemma 2 provides an alternative way to calculate the information loss by generalizing node$_i$ and node$_j$ in a given DGH. For instance, based on Lemma 2, the information loss incurred by generalizing node '10' to '1' in Fig. 1 (x-coordinate DGH), can be calculated as $|(\log_2 8 - \log_2 2) - (\log_2 8 - \log_2 4)| = 1$ bit.

Lemma 2 indicates that the loss incurred by generalizing two nodes is equal to the difference between losses incurred by their suppression. As before, for any clustering outcome of data, assume that $cluster$ is a two-dimensional list, in which the j-th element of the list returns the IDs of the trajectories in the j-th cluster. Moreover, we denote the j-th cluster head after generalization and suppression for all trajectories as h_j. Therefore, the total loss can be written as

$$\begin{aligned}\text{Total loss} &= LS(\overline{T}, x) + LS(\overline{T}, y) + LS(\overline{T}, t) \\ &= \sum_{j=0}^{k-1}\sum_{tr \in cluster[j]}(LS(h_j.x, tr.x) \\ &\quad + LS(h_j.y, tr.y) + LS(h_j.t, tr.t)).\end{aligned} \quad (12)$$

As explained in (7), the objective of clustering algorithms is to minimize this equation. Therefore, using Lemma 2 the equation (12) can be written as

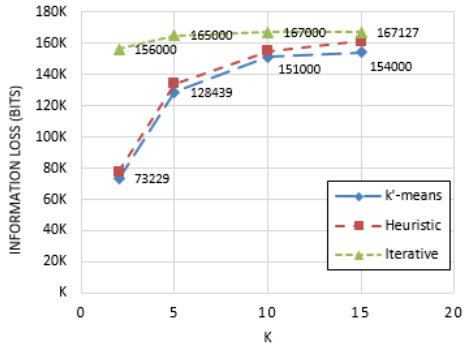
$$\begin{aligned}&\text{Total loss} = \quad (13)\\ &\sum_{j=0}^{k-1}\sum_{tr \in cluster[j]}(|LS(h_j.x, RT(H_x)) - LS(tr.x, RT(H_x)| \\ &\quad + |LS(h_j.y, RT(H_y)) - LS(tr.y, RT(H_y)| \\ &\quad + |LS(h_j.t, RT(H_t)) - LS(tr.t, RT(H_t))|. \quad (14)\end{aligned}$$

Rearranging (13), the objective equation can be found by minimizing total loss formulated in (8). This can be done by maximizing part B and minimizing part A. Since the cluster heads are generated based on the clustering algorithm, they cannot be used as part of the optimization process. Therefore, we aim at minimizing part A in (8).

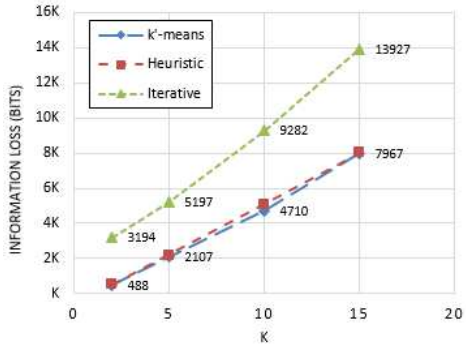
Part A in the equation (8) refers to finding the total distance of each trajectory from DGH root of the attributes. Therefore, for each trajectory, a three-dimensional vector $< d_x, d_y, d_t >$ is constructed, where d_x, d_y, d_t store the loss incurred by generalizing the x-coordinate, y-coordinate, and time, respectively. Having distances of all points from the roots, we cluster the trajectories using the k'-means algorithm. The algorithm clusters trajectories with a similar loss from the root in the same group. This process is particularly important as trajectory datasets usually include trajectories as short as one query to trajectories with hundreds of queries.

A major drawback of the k'-means algorithm is clustering the trajectories without any constraint on the minimum number of trajectories that needs to be in each cluster. Therefore, the algorithm might result in some of the clusters containing less than k trajectories that violates the k-anonymity of trajectories. If the data is not extremely sensitive such as the data used in the military, it is usually acceptable to have a few trajectories below the k-anonymity criterion. As it will be demonstrated in Section 5, the number of trajectories not achieving k-anonymity is close to or below 20% of the trajectories based on the value of k chosen for the privacy. To amend the naive k'-means algorithm for sensitive applications, we propose to use a variation of k'-means algorithm, which we call it iterative k'-means. The idea relies on running the k'-means algorithm iteratively to ensure that all clusters will achieve k-anonymity. Therefore, after each iteration of the k'-means algorithm, the clusters including at least k trajectories are disbanded, and the trajectories are put back into the pool for the next iteration of the k'-means algorithm. This process continues until all clusters have at least k members. Algorithm 4 represents the

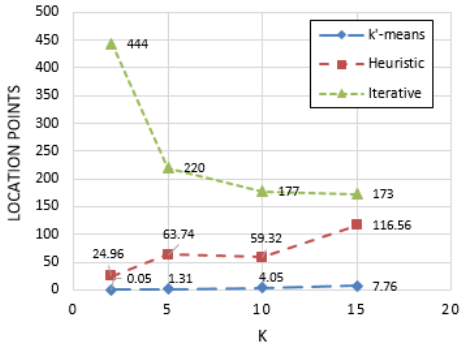
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964658, IEEE Transactions on Knowledge and Data Engineering



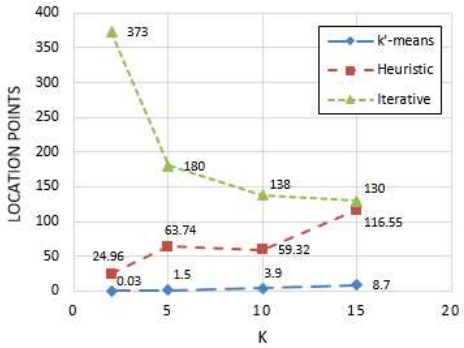
(a) Total information loss



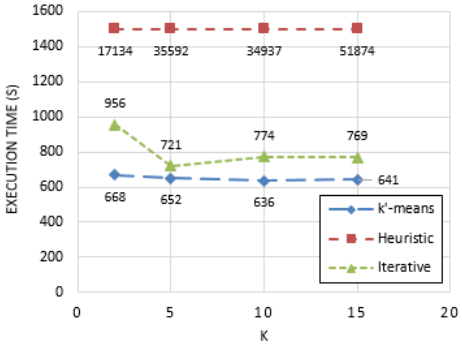
(b) Average information loss per cluster



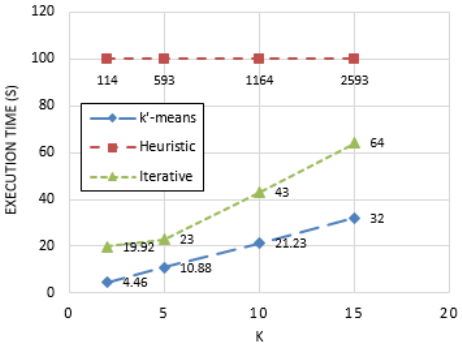
(c) Average increase in length of trajectories



(d) Average increase in length of trajectories per cluster



(e) Total execution time (the heuristic algorithm's results are shown as a flat line with the values written below the line)



(f) Average execution time per cluster (the heuristic algorithm's results are shown as a flat line with the values written below the line)

Figure 4: Performance evaluation of MLA with different values of k.

Algorithm 4: Pseudocode of iterative k'-means algorithm.

```
while true do
    run k'-means algorithm on dataset
     (#clusters = ⌊ #data trajectories / k ⌋)
    remove trajectories that belong to clusters with at
     least k members from the dataset
    if #len(dataset) < 2 * k then
        cluster the remaining trajectories together
        break;
    end
end
```

pseudocode of the iterative k'-means.

5 Experiments

In our experiments, we use the data collected by Geolife project [35]–[37], T-Drive dataset [38], [39], and Gowalla dataset [40]. For Geolife and T-Drive datasets, which include the GPS trajectories of mobile users, and taxi drivers in Beijing (China), we have considered a $1km \times 1\,km$ central part of the Beijing map with the resolution of $0.01km \times 0.01km$. For the Gowalla dataset, we have chosen the users over the map of New York City with the same resolution as the Geolife and T-Drive. The detailed statistics

Table 1: Statistics of datasets used in our experiments.

Dataset	Geolife	T-Drive	Gowalla
Total number of samples	47581	27916	138957
Number of trajectories	13561	301	7115
Average number of samples per trajectory	3.5	92.74	19.53

on the datasets are given in Table 1. Various location privacy requirements (k) of the users are investigated for values 2, 5, 10, and 15. The experiments were performed on a PC with a 3.40 GHz Core-i7 Intel processor, 64-bit Windows 7 operating system, and an 8.00 GB of RAM. The Python programming language was used to implement the algorithms.

We compare our work to prior methods as follows:

- Many of prior approaches for the anonymization of spatiotemporal trajectory datasets use a greedy or so-called heuristic approach to anonymize datasets. In Section 4.3.1, we explained and adopted this approach based on our system model. We use the heuristic approach in Section 5.1 as a baseline for comparison.
- The full comparison of the MLA framework with the recent work in [6] is provided in Section 5.3. The

results are verified on both of the T-Drive and Geolife datasets to ensure reliability.
- As MLA and the proposed algorithm in [6] seek to fulfill different objectives, we have further evaluated the two frameworks based on random clustering. Doing so shifts the focus to the alignment of trajectories in each cluster. The results are verified on both of the T-Drive and Geolife datasets (Section 5.3).
- We also compare our alignment approach with the widely used static algorithm in [3] (Section 5.3).

5.1 Performance Evaluation

Fig. 4 presents the performance evaluation of MLA predicated on three clustering approaches developed in this paper. The algorithms have been investigated from three aspects: information loss, increase in trajectory length, and execution time. In all graphs, x-axis indicates k-anonymity requirement for the dataset. The total information loss and average information loss per cluster of algorithms are considered in Figs. 4a and 4b, respectively. Information loss, shown in the y-axis, indicates the total loss incurred while applying generalization and suppression on x-coordinate, y-coordinate, and the time of the query. The maximum possible incurred information loss for the whole dataset by suppressing all trajectories is 474572 bits. This value is the upper bound on all anonymization algorithms. Note that this constant changes for different datasets. The main existing trend in Figs. 4a and 4b is that by increasing the value of k, the total incurred loss increases. This outcome meets our expectation as increasing the value of k indicates having larger cluster sets, which results in the alignment of a higher number of trajectories in each cluster, and thereby, a higher total loss by the alignment. Among our proposed algorithms, k'-means algorithm provides the best performance as it corresponds to minimum lost bits incurred by the generalization and suppression.

The amount of information that k'-means algorithm preserves is higher than that of the heuristic approach, in which the most suitable trajectories are chosen to minimize the information. This trend can be seen for both of the total information loss of the dataset and the average information loss of dataset per cluster for different k values. Such a trade-off exists, because some clusters contain a small number of trajectories not satisfying the k-anonymity requirement. The loss of privacy by k'-means algorithm is further analyzed in Fig. 5 which will be explained later in this section. The iterative k'-means algorithm is constructed on top of the k'-means algorithm to ensure that all the trajectories satisfy the required privacy requirement. This is particularly important for sensitive applications, in which there are strict requirements for privacy preservation. The cost of having higher privacy for the iterative k'-means algorithm is a larger loss of information.

Figs. 4c and 4d present the average increase in the length of trajectories for the whole dataset and per cluster. Due to the alignment process, shorter trajectories may need to be aligned with longer trajectories, which result in an increase in the length of trajectories in the anonymized released dataset. The best performance among the algorithms is yielded by the k'-means algorithm with the lowest increase in the lengths of trajectories. Compared to other two approaches, the heuristic strategy performs better than the iterative k'-means with a smaller k, but as the k value increases, the average increase in trajectory length converges due to large cluster size. Figs. 4e and 4f compare the total and average per cluster execution time of the different algorithms. Note that since the heuristic algorithm requires a significantly higher amount of time to run, it is shown on top of the graphs as a flat line with the corresponding values shown below it. The execution time of the k'-means and iterative k'-means algorithms are significantly lower than that of the heuristic algorithm and as expected the iterative k'-means consumes slightly more execution time as it has additional steps to ensure the k-anonymity of all trajectories.

5.2 Detailed Analysis of k'-means Algorithm

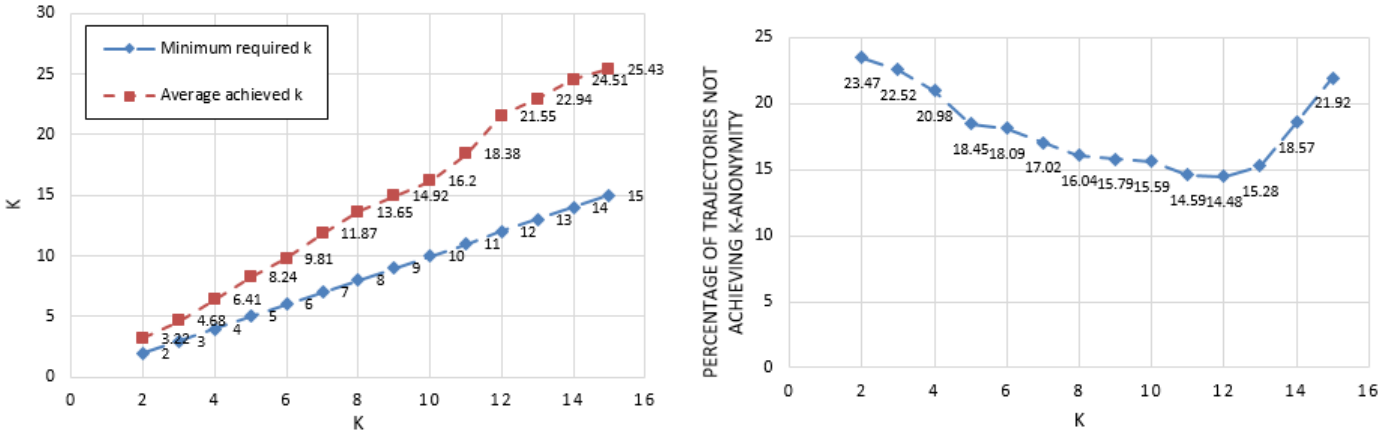
Overall, the detailed k'-means algorithm's results in satisfactory performance in terms of information loss, execution time, and the average increase in the length of trajectories. Moreover, the complexity of k'-means algorithm is of an order of the number of data entries for large datasets, whereas the order of the heuristic algorithm is proportional to the square of this number. Therefore, the k'-means algorithm has several significant advantages compared to the heuristic approach. Hence, if it is acceptable for the datasets to have a few trajectories below the k-anonymity requirement, then, it is more beneficial to use the k'-means algorithm instead of the heuristic or the iterative k'-means algorithm. This is usually true for datasets not entailing classified information. Therefore, we further analyze the performance of this algorithm in the remaining of this section and compare it to the state-of-art algorithms recently proposed. Also, note that in the rest of this paper when MLA is mentioned, the k'-means algorithms is adopted for clustering by default.

Fig. 5 provides two graphs showing the details of the performance yielded by the k'-means algorithm. The first graph indicates the average value of k achieved while applying the k'-means algorithm, and the second graph shows the percentage of trajectories that did not achieve the k-anonymity in the anonymization process with different values of k. In Fig. 5(a), it is evident that despite some of the trajectories losing their k-anonymity during the anonymization, the average value of anonymity achieved is above the minimum requirement. The value of the average gets even better as the value of k increases. Fig. 5(b) shows the percentage of the trajectories not achieving the minimum required k-anonymity. This value is below 20% on average, which means that over 80% of the trajectories are guaranteed to at least have k-anonymity. The reason causing the uneven curves in the figure is because the number of clusters is divisible by k, which results in an additional cluster distorting the curves.

5.3 Comparison

We compare MLA with the static algorithm proposed in [3], and recently published anonymization approach in [6]. The idea behind the static alignment algorithm in [3] is that two trajectories are matched element by element without any shifts or spaces. In more details, the static algorithm attempts to match two sequences based on the same index.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964658, IEEE Transactions on Knowledge and Data Engineering



(a) Average value of k achieved by applying the k'-means algorithm

(b) Percentage of users not satisfying k-anonymity requirement by applying the k'-means algorithm

Figure 5: Detailed performance evaluation of the k'-means algorithm.

Therefore, each element of the first sequence tr_1 is aligned with an element having the same index in the other input trajectory tr_2. Based on our evaluation, the total incurred information loss is reduced by 7.2% by using the proposed progressive SA algorithm. It must be noted that the dataset includes trajectories as large as hundreds of queries and as small as a single query from the location-based service provider. Therefore, matching these length-variant trajectories would impose a substantial information loss even for the best possible match of the sequences.

Fig. 6 indicates the comparison result between our proposed anonymization technique and the recent generalization method proposed in [6]. The authors in [6] attempted to minimize the incurred loss of the anonymization by sorting out the spatiotemporal locations in the time domain and applying a heuristic approach for generalization. They also used a heuristic approach for clustering trajectories. Note that any anonymization approach aims to maximize utility while preserving the privacy of users. Utility in generalization techniques refers to the area released for locations in the dataset. Therefore, to have a fair comparison, we compare our work with the approach proposed in [6] based on the average released area for locations. The metric is thoroughly explained in Section 3. It can be seen from the figure that our proposed algorithm can significantly increase the spatial utility of the generalization approach. In other words, the anonymized dataset has on average smaller released area per location while preserving the privacy of users. To further compare alignment approaches, in Fig. 6, we applied random clustering to group the trajectories, and then, used the alignment approach in our proposed work and the previous work to generate anonymized trajectories. As can be seen in the figure, our alignment approach outperforms the previous work by a higher spatial utility of anonymized dataset.

5.4 Discussion

As can be seen in Fig. 6, the MLA framework has significantly improved the spatial utility of data while achieving k-anonymity for the entries of datasets. A major reason for such an improvement is that MLA considers all three dimensions of time, x-coordinate, and y-coordinate together. Such consideration helps to minimize the overall cost and not just the utility in time or spatial domain. For instance, the Geolife datasets consists of sampling time interval of 177 seconds with the average distance interval of 623 meters, whereas the T-Drive dataset has the average sampling interval of $1-5$ seconds and $5-10$ meters of sampling distance interval. Therefore, the two datasets enatail a highly different sparsity characteristic in time and spatial domain. However, as can be seen in Figs. 6 and 7, the MLA algorithm considers all three dimensions, and can significantly improve the utility in the process of anonymization.

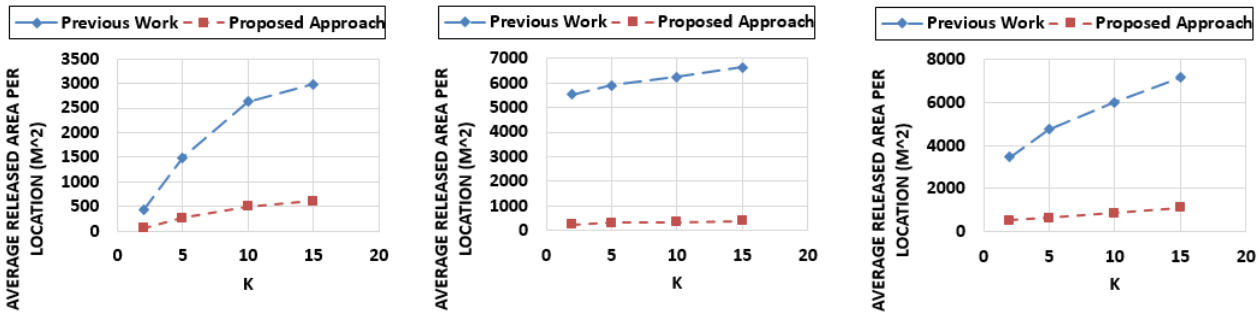
In essence, the performance improvement in our proposed model is predicated on both the clustering and alignment of trajectories. In terms of the alignment, progressive SA has resulted in significant improvement of the alignment process. Such an impact can be seen in Fig. 6, where we apply random clustering, and therefore, the focus is on the alignment. As the figure suggests, utilizing a multiple SA technique such as progressive SA used in MLA provides major improvements to the utility of the anonymized datasets.

For clustering, as finding the optimal anonymization of the datasets is proven to be NP-hard, most of the literature has focused on following heuristic approaches to cluster the trajectories. We adopted such a heuristic approach for the system model of our paper and presented the results in Fig. 4. Note that in the heuristic approach used in the figure, we are applying progressive SA alignment; therefore, the results show the improved version of the previously existing algorithms. As it was revealed in Fig. 4, the k'-means algorithm can outperform such heuristic approaches in addition to having a much lower implementation complexity and processing time.

6 Applications

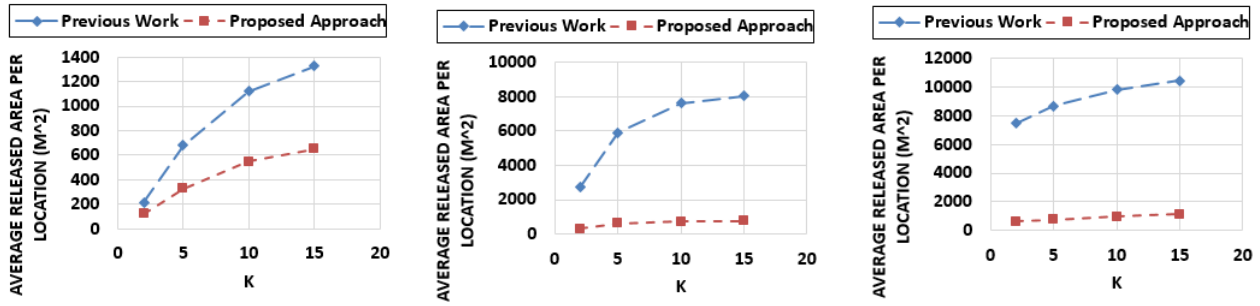
In this section, we introduce several applications that we believe our work has the most impact on.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964658, IEEE Transactions on Knowledge and Data Engineering



(a) Geolife dataset (b) T-Drive dataset (c) Gowalla dataset

Figure 6: Comparison of MLA with the previous work proposed in [6].



(a) Geolife dataset (b) T-Drive dataset (c) Gowalla dataset

Figure 7: Comparison of MLA with the previous work proposed in [6], applying random clustering.

6.1 Location-Based Data

As the framework for anonymization presented in this paper considers location trajectories, one of the main applications of the framework is the privacy of location-based data. The use of location-based applications is more prevalent than any time before. Governments attempt to analyze the infrastructure using the location data and researchers use these data to investigate human behavior. Research has verified that even simple analytics on these published trajectory data would yield serious risk of users' privacy and even be capable of identifying users of location-based applications. [41]. Therefore, applying anonymization techniques such as the one we have developed in this paper is necessary to preserve the privacy of the users.

6.2 Medical Records

The recent advances in medical information technology have enabled the collection of a detailed description of patients and their medical status [42]. Such data is usually stored in electronic medical record systems [43]–[45]. Similar to spatiotemporal trajectories, many of the medical records need to be published by agencies and organizations. Unfortunately, research has shown that solely relying on de-identification is insufficient to protect users' privacy, as the medical records from multiple databases can be linked together to identify individual patients [3]. Therefore, there is an urgent need for viable algorithms to anonymize the medical data. The problem of anonymization in spatiotemporal trajectories is very similar to anonymization in longitudinal electronic medical records. This can be easily justified by the similar way, in which these data are stored. Assume a patient who has referred to medics several times in his or her lifetime. Each time the records of the patient are stored in a longitudinal dataset, in which the age and the diagnosed disease record are registered. These longitudinal records can be seen as a trajectory for the patient, and our proposed algorithms in this paper can be applied to anonymize a dataset of such longitudinal electronic medical records.

6.3 Web Analytics

Another important application of the framework developed in this paper is web analytics. Web analytics refers to analyzing online traces of users. Web analytics has become a competitive advantage for many companies due to the amount of detailed information that can be extracted from the data. Therefore, protecting the trajectories that the users explored on the Internet has become a major challenge for researchers. The similarity between spatiotemporal trajectories and web analytics can be well explained by the following example. For instance, Geoscience Australia is constantly recording and publishing the site logs users make on their website. The site log filename is composed of a four-digit station identifier, followed by a two-digit month and a two-digit year, e.g., ALIC0414 is the site log for the Alice Springs GNSS site that was updated in April 2014 [2]. Such a trajectory of logins to the website is analogous to a spatiotemporal trajectory with three attributes. Therefore, the framework developed in this paper can be used to anonymize the online traces of users before publishing web browsing data.

7 Conclusion

In this paper, we have proposed a framework to preserve the privacy of users while publishing the spatiotemporal trajectories. The proposed approach is based on an efficient alignment technique termed as progressive sequence alignment in addition to a machine learning clustering approach that aims at minimizing the incurred loss in the anonymization process. We also devised a variation of k'-means algorithm for guaranteeing the k-anonymity in overly sensitive datasets. The experimental results on real-life GPS datasets indicate the superior spatial utility performance of our proposed framework compared with the previous works.

References

[1] S. Shaham, M. Ding, B. Liu, Z. Lin, and J. Li, "Machine learning aided anonymization of spatiotemporal trajectory datasets," *arXiv preprint arXiv:1902.08934*, 2019.

[2] A. Government, "New australian government data sharing and release legislation," 2018.

[3] A. Tamersoy, G. Loukides, M. E. Nergiz, Y. Saygin, and B. Malin, "Anonymization of longitudinal electronic medical records," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 3, pp. 413–423, 2012.

[4] F. Xu, Z. Tu, Y. Li, P. Zhang, X. Fu, and D. Jin, "Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 1241–1250.

[5] Y. Dong and D. Pi, "Novel privacy-preserving algorithm based on frequent path for trajectory data publishing," *Knowledge-Based Systems*, vol. 148, pp. 55–65, 2018.

[6] M. Gramaglia, M. Fiore, A. Tarable, and A. Banchs, "Towards privacy-preserving publishing of spatiotemporal trajectory data," *arXiv preprint arXiv:1701.02243*, 2017.

[7] M. Terrovitis, G. Poulis, N. Mamoulis, and S. Skiadopoulos, "Local suppression and splitting techniques for privacy preserving publication of trajectories," *IEEE Trans. Knowl. Data Eng*, vol. 29, no. 7, pp. 1466–1479, 2017.

[8] M. E. Nergiz, M. Atzori, and Y. Saygin, "Towards trajectory anonymization: a generalization-based approach," in *Proc. of the SIGSPATIAL ACM GIS*. ACM, 2008, pp. 52–61.

[9] S. Gurung, D. Lin, W. Jiang, A. Hurson, and R. Zhang, "Traffic information publication with privacy preservation," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 3, p. 44, 2014.

[10] R. Yarovoy, F. Bonchi, L. V. Lakshmanan, and W. H. Wang, "Anonymizing moving objects: How to hide a mob in a crowd?" in *Proc. of the 12th International Conference on Extending Database Technology: Advances in Database Technology*. ACM, 2009, pp. 72–83.

[11] B. Liu, W. Zhou, T. Zhu, L. Gao, and Y. Xiang, "Location privacy and its applications: A systematic study," *IEEE Access*, vol. 6, pp. 17 606–17 624, 2018.

[12] G. Poulis, G. Loukides, S. Skiadopoulos, and A. Gkoulalas-Divanis, "Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints," *Journal of biomedical informatics*, vol. 65, pp. 76–96, 2017.

[13] T. Takahashi and S. Miyakawa, "Cmoa: Continuous moving object anonymization," in *Proceedings of the 16th International Database Engineering & Applications Sysmposium*. ACM, 2012, pp. 81–90.

[14] X. Zhou and M. Qiu, "A k-anonymous full domain generalization algorithm based on heap sort," in *International Conference on Smart Computing and Communication*. Springer, 2018, pp. 446–459.

[15] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, 2005, pp. 49–60.

[16] S. Yaseen, S. M. A. Abbas, A. Anjum, T. Saba, A. Khan, S. U. R. Malik, N. Ahmad, B. Shahzad, and A. K. Bashir, "Improved generalization for secure data publishing," *IEEE Access*, vol. 6, pp. 27 156–27 165, 2018.

[17] G. Poulis, A. Gkoulalas-Divanis, G. Loukides, S. Skiadopoulos, and C. Tryfonopoulos, "Secreta: A tool for anonymizing relational, transaction and rt-datasets," in *Medical Data Privacy Handbook*. Springer, 2015, pp. 83–109.

[18] K. Sreedhar, M. Faruk, and B. Venkateswarlu, "A genetic tds and bug with pseudo-identifier for privacy preservation over incremental data sets," *Journal of Intelligent & Fuzzy Systems*, vol. 32, no. 4, pp. 2863–2873, 2017.

[19] M. E. Rana, M. Jayabalan, and M. A. Aasif, "Privacy preserving anonymization techniques for patient data: An overview," in *Third International Congress on Technology, Communication and Knowledge (ICTCK 2016)*, 2016.

[20] M. Jayabalan and M. E. Rana, "Anonymizing healthcare records: A study of privacy preserving data publishing techniques," *Advanced Science Letters*, vol. 24, no. 3, pp. 1694–1697, 2018.

[21] D. Narula, P. Kumar, and S. Upadhyaya, "Privacy preservation using various anonymity models," in *Cyber Security: Proceedings of CSI 2015*. Springer, 2018, pp. 119–130.

[22] J. Ding, "Trajectory mining, representation and privacy protection," in *Proceedings of the 2nd ACM SIGSPATIAL PhD Workshop*. ACM, 2015, p. 2.

[23] A. E. Cicek, M. E. Nergiz, and Y. Saygin, "Ensuring location diversity in privacy-preserving spatio-temporal data publishing," *The VLDB JournalThe International Journal on Very Large Data Bases*, vol. 23, no. 4, pp. 609–625, 2014.

[24] C. Romero-Tris and D. Megías, "Protecting privacy in trajectories with a user-centric approach," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 12, no. 6, p. 67, 2018.

[25] F. T. Brito, A. C. A. Neto, C. F. Costa, A. L. Mendonça, and J. C. Machado, "A distributed approach for privacy preservation in the publication of trajectory data," in *Proceedings of the 2nd Workshop on Privacy in Geographic Information Collection and Analysis*. ACM, 2015, p. 5.

[26] E. Naghizade, L. Kulik, and E. Tanin, "Protection of sensitive trajectory datasets through spatial and temporal exchange," in *Proc. of the 26th International Conference on Scientific and Statistical Database Management*. ACM, 2014, p. 40.

[27] K. Jiang, D. Shao, S. Bressan, T. Kister, and K.-L. Tan, "Publishing trajectories with differential privacy guarantees," in *Proc. of the 25th International Conference on Scientific and Statistical Database Management*. ACM, 2013, p. 12.

[28] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[29] B. Chowdhury and G. Garai, "A review on multiple sequence alignment from the perspective of genetic algorithm," *Genomics*, 2017.

[30] X. Chen, C. Wang, S. Tang, C. Yu, and Q. Zou, "Cmsa: a heterogeneous cpu/gpu computing system for multiple similar rna/dna sequence alignment," *BMC bioinformatics*, vol. 18, no. 1, p. 315, 2017.

[31] Q. Le, F. Sievers, and D. G. Higgins, "Protein multiple sequence alignment benchmarking through secondary structure prediction," *Bioinformatics*, vol. 33, no. 9, pp. 1331–1337, 2017.

[32] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[33] S. K. Pal and P. P. Wang, *Genetic algorithms for pattern recognition*. CRC press, 2017.

[34] A. Fischer and D. Picard, "Convergence rates for smooth k-means change-point detection," *arXiv preprint arXiv:1802.07617*, 2018.

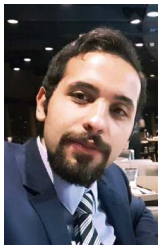
[35] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 791–800.

[36] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding mobility based on gps data," in *Proceedings of the 10th international conference on Ubiquitous computing*. ACM, 2008, pp. 312–321.

[37] Y. Zheng, X. Xie, and W.-Y. Ma, "Geolife: A collaborative social networking service among user, location and trajectory." *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32–39, 2010.

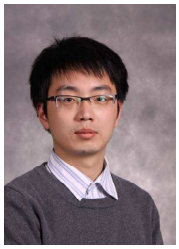
[38] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-drive: driving directions based on taxi trajectories," in *Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems*. ACM, 2010, pp. 99–108.

[39] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 316–324.
[40] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1082–1090.
[41] Y. Sun, M. Chen, L. Hu, Y. Qian, and M. M. Hassan, "Asa: Against statistical attacks for privacy-aware users in location based service," *Future Generation Computer Systems*, vol. 70, pp. 48–58, 2017.
[42] O. Asan, F. Cooper II, S. Nagavally, R. J. Walker, J. S. Williams, M. N. Ozieh, and L. E. Egede, "Preferences for health information technologies among us adults: Analysis of the health information national trends survey," *Journal of medical Internet research*, vol. 20, no. 10, p. e277, 2018.
[43] C. Gates, "Electronic medical record reminder to improve human papillomavirus vaccination rates among adolescents," 2018.
[44] F. Rahman, P. Karanikas, and T. D. Giles, "Systems and methods for creating contextualized summaries of patient notes from electronic medical record systems," Aug. 17 2017, uS Patent App. 15/430,401.
[45] J. Colleti Junior, A. B. d. Andrade, and W. B. d. Carvalho, "Evaluation of the use of electronic medical record systems in brazilian intensive care units," *Revista Brasileira de terapia intensiva*, vol. 30, no. 3, pp. 338–346, 2018.



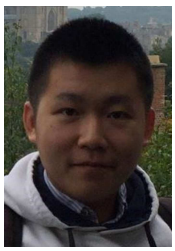
Sina Shaham received B.Eng (Hons) in Electrical and Electronic Engineering from the University of Manchester (with first class honors). He is currently an MPhil student at the University of Sydney. He has years of experience as a Data Scientist and Software Engineer in companies such as InDebted. His current research interests include applications of artificial intelligence in big data and privacy.

Ming Ding (M'12-SM'17) received the B.S. and M.S. degrees (with first class Hons.) in electronics engineering from Shanghai Jiao Tong University (SJTU), Shanghai, China, and the Doctor of Philosophy (Ph.D.) degree in signal and information processing from SJTU, in 2004, 2007, and 2011, respectively. From April 2007 to September 2014, he worked at Sharp Laboratories of China in Shanghai, China as a Researcher/Senior Researcher/Principal Researcher. He also served as the Algorithm Design Director and Programming Director for a system-level simulator of future telecommunication networks in Sharp Laboratories of China for more than 7 years. Currently, he is a senior research scientist at Data61, CSIRO, in Sydney, NSW, Australia. He has authored over 80 papers in IEEE journals and conferences, all in recognized venues, and about 20 3GPP standardization contributions, as well as a Springer book "Multi-point Cooperative Communication Systems: Theory and Applications". Also, he holds 16 US patents and co-invented another 100+ patents on 4G/5G technologies in CN, JP, EU, etc. Currently, he is an editor of IEEE Transactions on Wireless Communications. Besides, he is or has been Guest Editor/Co-Chair/Co-Tutor/TPC member of several IEEE top-tier journals/conferences, e.g., the IEEE Journal on Selected Areas in Communications, the IEEE Communications Magazine, and the IEEE Globecom Workshops, etc. He was the lead speaker of the industrial presentation on unmanned aerial vehicles in IEEE Globecom 2017, which was awarded as the Most Attended Industry Program in the conference. Also, he was awarded in 2017 as the Exemplary Reviewer for IEEE Transactions on Wireless Communications.

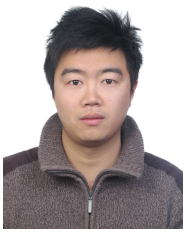


Bo Liu (M'10) received the B.Sc. degree from the Department of Computer Science and Technology from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2004 and then he received the and MEng. and PhD degrees from the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2007, and 2010 respectively. He is currently a Senior Lecturer with the University of Technology Sydney, Australia. His research interests include cybersecurity and privacy, wireless communications and networks.

Zihuai Lin received the Ph.D. degree in Electrical Engineering from Chalmers University of Technology, Sweden, in 2006. Prior to this he has held positions at Ericsson Research, Stockholm, Sweden. Following Ph.D. graduation, he worked as a Research Associate Professor at Aalborg University, Denmark and currently at the School of Electrical and Information Engineering, the University of Sydney, Australia. His research interests include source/channel/network coding, coded modulation, MIMO, OFDMA, SC-FDMA, radio resource management, cooperative communications, small-cell networks, 5G cellular systems, etc.



Shuping Dang (S'13–M'18) received B.Eng (Hons) in Electrical and Electronic Engineering from the University of Manchester (with first class honors) and B.Eng in Electrical Engineering and Automation from Beijing Jiaotong University in 2014 via a joint '2+2' dual-degree program. He also received D.Phil in Engineering Science from University of Oxford in 2018. Dr. Dang joined in the R&D Center, Huanan Communication Co., Ltd. after graduating from University of Oxford and is currently working as a Postdoctoral Fellow with the Computer, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology (KAUST). He serves as a reviewer for a number of key journals in communications and information science, including IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. His current research interests include artificial intelligence assisted communications, novel modulation schemes and cooperative communications.



Jun Li (M'09-SM'16) received Ph. D degree in Electronic Engineering from Shanghai Jiao Tong University, Shanghai, P. R. China in 2009. From January 2009 to June 2009, he worked in the Department of Research and Innovation, Alcatel Lucent Shanghai Bell as a Research Scientist. From June 2009 to April 2012, he was a Postdoctoral Fellow at the School of Electrical Engineering and Telecommunications, the University of New South Wales, Australia. From April 2012 to June 2015, he is a Research Fellow at the School of Electrical Engineering, the University of Sydney, Australia. From June 2015 to now, he is a Professor at the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests include network information theory, ultra-dense wireless networks, and mobile edge computing.