

Computation and communication efficient approach for federated learning based urban sensing applications against inference attacks

Ayshika Kapoor, Dheeraj Kumar *

Department of Electronics & Communications Engineering, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, 247667, India

ARTICLE INFO

Keywords:

Urban sensing
Federated learning
Spatial-temporal entropy
Secure multiparty computation
Privacy
Kullback–Leibler divergence

ABSTRACT

Federated learning based participatory sensing has gained much attention lately for the vital task of urban sensing due to privacy and security issues in conventional machine learning. However, inference attacks by the honest-but-curious application server or a malicious adversary can leak the personal attributes of the participants, such as their home and workplace locations, routines, and habits. Approaches proposed in the literature to prevent such information leakage, such as secure multi-party computation and homomorphic encryption, are infeasible for urban sensing applications owing to high communication and computation costs due to multiple rounds of communication between the user and the server. Moreover, for effective modeling of urban sensing phenomenon, the application model needs to be updated frequently — every few minutes or hours, resulting in periodic data-intensive updates by the participants, which severely strains the already limited resources of their mobile devices. This paper proposes a novel low-cost privacy-preserving framework for enhanced protection against the inference of participants' personal and private attributes from the data leaked through inference attacks. We propose a novel approach of *strategically* leaking selected location traces by providing computation and communication-light direct (local) model updates, whereas the rest of the model updates (when the user is at sensitive locations) are provided using secure multi-party computation. We propose two new methods based on spatiotemporal entropy and Kullback–Leibler divergence for automatically deciding which model updates need to be sent through secure multi-party computation and which can be sent directly. The proposed approach significantly reduces the computation and communication overhead for participants compared to the fully secure multi-party computation protocols. It provides enhanced protection against the deduction of personal attributes from inferred location traces compared to the direct model updates by confusing the application server or malicious adversary while inferring personal attributes from location traces. Numerical experiments on the popular Geolife GPS trajectories dataset validate our proposed approach by reducing the computation and communication requirements by the participants significantly and, at the same time, enhancing privacy by decreasing the number of inferred sensitive and private locations of participants.

1. Introduction

The recent advancements in technology have seen substantial progress in mobile phone capabilities, *artificial intelligence* (AI), *internet of things* (IoT), telecommunication services, and edge computing that have interconnected the cyber and physical environment in urban areas. Around 7.26 billion people (91.54% of the world population) own a mobile phone today, of which 6.64

* Corresponding author.

E-mail addresses: ayshika_k@ece.iitr.ac.in (A. Kapoor), dheeraj.kumar@ece.iitr.ac.in (D. Kumar).

<https://doi.org/10.1016/j.pmcj.2024.101875>

Received 20 February 2023; Received in revised form 21 December 2023; Accepted 4 January 2024

Available online 9 January 2024

1574-1192/Â© 2024 Elsevier B.V. All rights reserved.

billion are smartphones [1]. The spatio-temporal data generated from ubiquitous smartphones, equipped with embedded sensors like GPS, microphone, camera, ambient light, proximity, accelerometer, etc., provide unlimited opportunities for various urban sensing tasks such as environment monitoring, healthcare, and intelligent transportation. The applications of urban sensing through residents' smartphones include environmental monitoring and pollution mapping [2–5], urban traffic, road condition, and travel time estimation [6–10], noise mapping [11–16], and air quality measurement [5].

An essential characteristic of these urban sensing applications is the highly personal nature of users' data, such as location and time. Apart from establishing a secure communication link between server and user, many provisions for privacy and security need to be made for scenarios like information leakage, misuse of users' private data, and eavesdropping. Even with a perfectly secure communication channel and anonymization, the sensed data shared with the central server for application model training can be used to infer personal attributes of the user, e.g., participants' home and workplace locations, routines, and habits [17]. This information leakage also holds true for anonymous contributions, where location traces may be analyzed to infer participants' identity, daily routine, and frequently traveled routes using publicly available information [18]. Applications that use a mobile phone's microphone to record audio can behave as smart spies and may be used by adversaries to determine the context. Therefore, the major privacy concern of the urban sensing participants is the potential leak of personal information *unintentionally* while providing sensor data. Without strong privacy guarantees, many users are unlikely to participate in the process, thus defeating the whole purpose of large-scale urban sensing.

The recent advancements in smartphones' hardware configuration and software capabilities have paved the way to shift the burden of data processing from a centralized server to distributed mobile phone units. To this end, *federated learning* (FL) proposed by McMahan et al. of Google [19] seems a potential solution for urban sensing applications. FL is a novel paradigm that allows collaborative training of machine learning models in a distributive manner over data sensed, collected and stored locally on users' smartphones [20]. In FL, the application server transmits the global model to the participating users, who update it using their local model trained using local measurements. The locally updated model is transmitted back to the central application server without transmission of the raw data. The application server aggregates the updated local models received from the participating users to generate a global model. This process is repeated periodically to get an up-to-date application model.

Although users' local data never leaves the device, FL alone is not sufficient to provide privacy to users' sensitive data [21–23]. Recent works have shown that local models trained on users' smartphones are vulnerable to model inversion attacks [24,25], or inference attacks [26–29]. The model inversion attack reconstructs the local data using model gradients, whereas, in the inference attack, the adversary tries to infer the user's sensitive data from model updates. This raises serious privacy concerns for users participating in FL. To prevent such information leakage, a few strategies such as *secure multiparty computation* (SMC), *differential privacy* (DP), and *homomorphic encryption* (HE) have been proposed. SMC allows distributed parties or users to collaboratively compute a function without revealing their private inputs to other clients. This technique ensures that clients learn only the final cumulative model weight and not the individual inputs. DP provides statistical privacy guarantees by injecting random noises into model parameters, thus compromising model accuracy. With HE, the application server can aggregate encrypted local model updates without decrypting them. The resulting output, when decrypted, provides the same results as that produced when the operations are performed on the unencrypted data. This ensures that no local model update is revealed to the server during aggregation. SMC and HE both provide strong privacy guarantees at no expense of accuracy loss. However, they have significant communication and computation costs due to multiple rounds of communication between the user and server. Moreover, not all application models can be trained using HE due to implementation limitations [30–32]. For effective modeling of urban sensing phenomenon, the data needs to be recorded periodically, every few minutes or hours, resulting in periodic data-intensive updates by the participants [14,33–35]. However, almost all SMC protocols face high computation and communication overheads, making them unfavorable for urban sensing applications [36–38]. Moreover, the participants in urban sensing may have limited resources, such as poor network connection, processing, and battery resources. Applying the existing SMC schemes will result in high resource utilization for participants, which they might not have. Therefore, a low-cost privacy-preserving SMC method against inference attacks is the need of the hour.

For a participant, the most dreadful scenario would be that a malicious adversary can identify personal attributes such as home and workplace location, identity, daily routines, and frequently traveled routes using inferred location traces and the information available in the public domain. As a reference scenario, if a participant has a weekly appointment with a doctor, an adversary could potentially infer this detail from his/her location traces and publicly available information, such as Google maps or yellow pages, revealing sensitive health-related information about the participant. However, it is not essential that all the places visited by an individual are significant locations, as some of them may only be infrequent visits and hence will not reveal any significant information about the user, even if they are leaked or inferred by the honest-but-curious application server or a malicious adversary. Applying resource intensive inference attack preventing strategies such as SMC or HE for all application model updates might result in high communication and computational costs for the participants, draining the precious resources of their mobile devices. On the other hand, using direct model updates for all model updates would lead to potential privacy leaks. Both these scenarios might lead to the disenchantment of participants from these applications, leading to their failure; hence, a middle ground needs to be established for the success of these federated learning based urban sensing applications.

Many researchers have proposed clustering techniques such as *density-based spatial clustering of applications with noise* (DBSCAN) [39], *EDBSCAN* [40], *SNN+* [41] or implemented de-anonymization attack called as *mobility Markov chain* (MMC) [42], or *user hidden Markov model* [43] to identify sensitive locations from spatial-temporal data. To address this legitimate concern of participants, this paper proposes a novel approach of *strategically* leaking (making the application server aware of it) selected location traces by providing computation and communication-light direct model updates (without utilizing SMC). The rest of the

model updates (when the user is at sensitive locations) are provided using SMC. The objective is to *confuse* the application server or malicious adversary while inferring personal attributes mentioned above from location traces. The proposed approach significantly reduces computation and communication overhead for participants while providing enhanced protection against the deduction of personal attributes from inferred location traces. We propose two new methods based on *spatiotemporal entropy* and *Kullback–Leibler* (KL) divergence for *automatically*, deciding which model updates need to be sent through SMC and which can be sent directly. The objective of this work is to design a low-cost, privacy-preserving framework for effective modeling of urban sensing phenomenon using federated learning. More specifically, our work aims to:

1. Find a balance between computation and communication cost on the mobile device of crowd-sourcing participants and preserve their privacy by prohibiting inference of sensitive and private locations.
 - (a) Develop an adaptive transmission strategy, which is a combination of direct model updates for reducing computation and communication costs and SMC for preventing leakage of private and sensitive information of participants.
2. Designing data-driven approaches for automatic detection of private and sensitive locations, which may vary with time, so that appropriate adaptive transmission strategy can be adopted.
3. Provide theoretical proof of the effectiveness of the proposed approach by establishing the relationship between parameters controlling communication and computation cost of the strategy proposed in this paper and the achievable privacy.
4. Experimentally validate the computation and communication cost and privacy leakage while employing the adaptive transmission strategy for various datasets.

The remainder of this paper is organized as follows. In Section 2, we review the relevant related work. In Section 3, we discuss the approaches for automatic detection of non-sensitive locations. We illustrate the details of our proposed scheme in Section 4, discussing spatiotemporal entropy and histogram distance measurement using KL divergence for a federated learning framework. In Section 5 we analyze the privacy and security of proposed schemes and state the conditions for privacy guarantees. Section 6 presents the experimental validation of the proposed methodology and elucidates the obtained results. In Section 7 we state few limitations of our work. Finally, Section 8 concludes this paper and highlights possible directions for future work.

2. Related work

Several works have addressed various privacy and security issues of clients' sensitive data in urban sensing applications. There are a lot of existing works on the privacy preservation of location traces for FL. Most of them rely on SMC or HE, leading to significant communication and computation costs, making it unaffordable for applications requiring frequent model updates, while DP compromises model accuracy. The first DP based FL was proposed by McMahan et al. [44], which evaluated *federated stochastic gradient descent* (FedSGD) and *federated averaging* (FedAvg) algorithms proposed by McMahan et al. in [45] to provide user-level privacy protection for recurrent language models. Bonawitz et al. [46] proposed a secure aggregation protocol based on cryptographic primitives that can handle worst-case user dropouts. However, the computational cost is quadratic for users and cubic for the central aggregator. Kang et al. [47] introduced a novel framework based on the concept of global DP and developed variances of artificial noise terms on the server and client side. However, a better privacy guarantee results in a lower convergence level. Yang et al. [48] proposed a *local differential privacy* (LDP) for scenarios where several vehicles are connected to a cloud server. This mechanism reverses the uploaded gradients, injects LDP noise, and integrates it with the FedSGD algorithm to prevent privacy leakage of gradients. Truex et al. [49] proposed a secure protocol that combines DP and SMC. However, each iteration requires at least four communication rounds between the central server and clients. Fereidooni et al. [50] provide a secure aggregation protocol that requires two communication rounds at each iteration. It is based on the encryption and decryption of global and local models. The encryption techniques used can be HE or SMC. Choi et al. [51] proposed a low-complexity protocol that reduces communication and computational resources, which integrates cryptographic tools of secret sharing and encryption. Their key idea is to reduce complete star topology with a random subset of clients and apply secret sharing or encryption techniques only to a subset of clients. This approach requires at least three rounds of communication between the client and server at each iteration. Bell et al. [52] presented a secure aggregation scheme that provides security against a semi-malicious aggregation server and a few malicious participants by achieving polylogarithmic communication and computation per participant. Kadhe et al. [53] proposed a multi-secret sharing scheme based on *fast Fourier transform* (FFT). It offers security against adaptive adversaries and can handle user dropouts. So et al. [54] proposed a secure aggregation protocol *turbo aggregate* that employs a multi-group circular strategy. The scheme leverages additive secret sharing and adds aggregation redundancy via *Lagrange coding* to enable robustness against dropped or delayed users. It involves one communication round per participant. Niu et al. [55] proposed a spatial data aggregation scheme for privacy protection in *mobile crowdsensing* (MCS) networks. A *Paillier* HE system is integrated with an aggregate signature algorithm to protect the location privacy of mobile clients. This protocol includes two communication rounds at each iteration. The communication and computation overhead of these approaches makes them unfeasible for urban sensing applications, which require frequent model updates from clients having limited network resources [14,33–35]. Therefore, a low-cost method to preserve privacy for urban sensing applications is needed.

Many researchers have introduced inference attacks specifically for spatiotemporal data to discover the user's whereabouts. Aryal et al. [41] proposed an unsupervised density based spatiotemporal clustering algorithm *SNN+* and showed that it could be used to discover useful information and interesting patterns from geo-located data. The technique extends *shared nearest neighbor* (SNN) clustering algorithm by defining the spatiotemporal distance between a pair of events and can automatically determine the number

of clusters in the spatiotemporal domain. However, the algorithm does not support the clustering of large spatiotemporal data sets. The *enhanced dynamic density based clustering algorithm* (EDDBSCAN) proposed by Angmo et al. [40] is an improved DBSCAN clustering algorithm. The method can adaptively set minimum point parameter based on merge sort and silhouette analysis. This technique can be used to conduct inference attacks by identifying users' sensitive locations and behavior patterns from density variations. Gambs et al. [42] focused on a de-anonymization attack to re-identify users behind a set of mobility traces. From the traces, *mobility Markov chain* (MMC) can be implemented by an adversary that models the mobility behavior of a user. Similarly, Kang et al. [56] presented a supervised graph-based post-processing algorithm that clusters and automatically extracts significant places from geo-coordinates while inferring transportation modes.

To the best of our knowledge, we are the first to apply the concept of strategic exposure of user sensed data for achieving enhanced user privacy by confusing the malicious adversary to identify personal attributes using inferred location traces and the information available in the public domain. The proposed scheme is validated by performing inference attacks. Moreover, the proposed approach significantly reduces computation and communication overhead for participants, which is discussed next.

3. Approaches for automatically detecting non-sensitive location-time for strategic leaking

For most of the urban sensing systems [2–4,6–8,11–16], the data and the application model are both spatiotemporal in nature, as these applications aim to model the dynamic spatiotemporal variations of the monitored phenomenon. Participants' smartphones record geographical location, i.e. latitude and longitude, using an inbuilt GPS sensor or is provided by the network service provider through triangulation. To achieve effective modeling, it is necessary to frequently update the application model. From the participants' perspective, some location-time may be sensitive, while others might be non-sensitive. Sensitive location-time refers to the data that, when combined with the information available in the public domain, may reveal an individual's private or confidential activities, habits, or personal details. This comprises data points that, if disclosed, might jeopardize someone's privacy, safety, or security. Sensitive location-time data examples include a person's home address, workplace location, regular visits to medical institutions, daily routine, and participation at sensitive events. In contrast, non-sensitive location-time data pertains to information that, when considered independently, lacks any personal or sensitive context. These individual data points pose no risk to an individual's privacy or security. For instance, examples of non-sensitive location-time data may include visits to public places such as parks, museums, or restaurants, which hold no association with an individual's identity or personal life.

The periodically recorded data leads to regular data-intensive spatiotemporal model updates by the participants. Due to the highly personal nature of this data, ensuring significant privacy protection becomes essential. However, nearly all SMC protocols encounter considerable computation and communication overhead, making them unsuitable for urban sensing applications. Furthermore, participants in urban sensing often contend with limited resources, including poor network connections, processing capabilities, and battery resources. Any individual will not agree to participate in contributing to the urban sensing applications until his sensitive information is being protected from leakage or being inferred. This paper presents a novel scheme in which the participants strategically share selected model updates directly with the application server (without SMC) and the rest with SMC so that the application server or malicious adversary cannot infer personal attributes from the inferred location traces, while significantly reducing communication and computation cost for the participants. Employing the proposed schemes, the sensitive location-time data can be identified through a decrease in the relative change in spatio-temporal entropy or an increase in the relative change in KL divergence distance. For this, first, we need to partition spatiotemporal data into space–time bins from which personal locations and routines can be inferred.

3.1. Spatiotemporal data binning

An urban sensing task s_n is a collection of sequenced spatiotemporal records $s_n = (l_1, t_1) \rightarrow (l_2, t_2) \rightarrow (l_3, t_3) \rightarrow \dots \rightarrow (l_n, t_n)$, where l denotes the recorded GPS coordinate and t indicate the observation time [57]. The GPS coordinates are expressed as latitude and longitude pairs, i.e., (l_x, l_y) . Thus, the sensed data can be denoted as a three-dimensional tuple, (l_x, l_y, t_i) , where $i \in \mathbb{I}$ is the indexing variable. Each participant's location and time data can be mapped to a three-dimensional spatiotemporal space that is uniformly binned into several *space–time cubes* (STC). Specifically, we partition the geographical region into $i \times j$ grids, where i refers to partition along the x -axis or latitude, and j refers to partition along the y -axis or longitude. The z -axis or time is divided into k equal intervals (say, 24 partitions, one each for an hour of the day). Each recorded data can then be assigned to one of the STC according to its latitude, longitude and time of observation as shown in Fig. 1. Each STC is considered to be a *symbol* or *bin* while calculating the spatiotemporal entropy or KL divergence, as discussed next.

In urban sensing, the data is recorded at regular intervals of time, implying that an individual's destination location and the travel route that he took while arriving will both be recorded. In this paper, we are considering inter-day routine, i.e., someone visiting a similar or nearby location at a similar hour of the day for several days, which might help the adversary to predict a user's whereabouts at a particular time of day in the future. This further helps to protect a user's sensitive location-time data such as home or workplace location, daily routine, and regular visits to a doctor.

3.2. Spatiotemporal entropy based sensitive location-time determination

The concept of *entropy* is utilized in many areas of engineering and science to measure the heterogeneity (diversity) of observations. The classic definition of entropy was introduced in 1948 by Shannon in his seminal work on information theory.

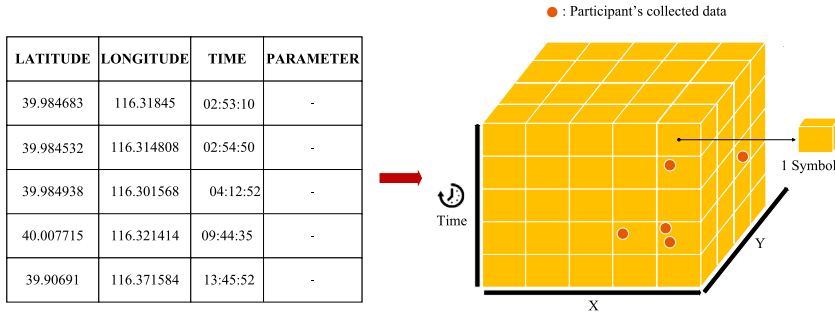


Fig. 1. Structure of space-time cube and spatiotemporal data partition of participant's recorded data.

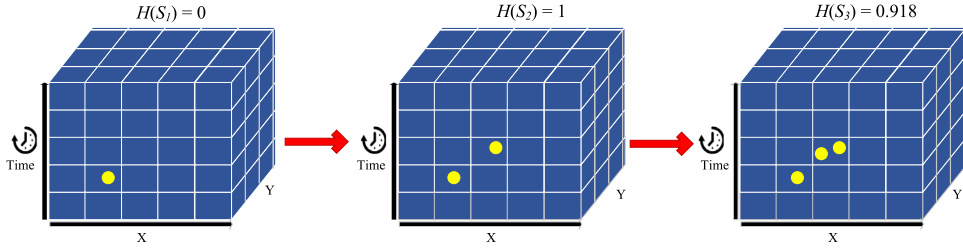


Fig. 2. An example of change in spatial-temporal entropy as a function of time.

Formally, Shannon's entropy is simply the amount of information in a variable. The term information is associated with the concept of uncertainty, i.e. greater the uncertainty in observing a random variable, the greater the information it contains. Let X be a discrete random variable that takes value x_i in a set of N outcomes with probability $P(x_i)$, where $i \in \{1, 2, \dots, N\}$. The Shannon's entropy of X is then defined as [58]:

$$\text{Entropy}(X) = \sum_{i=1}^N P(x_i) \log \left(\frac{1}{P(x_i)} \right) \quad (1)$$

For spatiotemporal analysis, we consider each STC as a separate symbol to transform continuous spatiotemporal data points into discrete random variables. Thus, the three-dimensional spatiotemporal space can be binned into a total of $i \times j \times k$ symbols. Each observation of the participant, i.e., latitude, longitude, and time of occurrence, is mapped to an STC symbol. The spatial-temporal entropy for the measurements till time n , i.e., s_n can then be calculated as:

$$\text{Entropy}(s_n) = H(s_n) = \sum_{m=1}^{i \times j \times k} P(Y_m) \log \left(\frac{1}{P(Y_m)} \right) \quad (2)$$

where, Y_m refers to the STC symbols and $P(Y_m)$ is the probability of occurrence of that STC symbol. Please note that $H(s_n)$ is a function of time n , and is set to zero at the beginning of the experiment and keeps on changing as new measurements are made periodically. Also, the STC is not refreshed at the end of each day; the data points recorded will be mapped onto the same STC the following day.

The expression for $H(s_n)$ indicates that when the user is at an unexplored location-time (where he/she usually does not go), Y_m will map to a different STC symbol, leading to an *increase* in the entropy $H(s_n)$. However, if the user is at the same/nearby location for a long time (sensitive information for inferring home/office location, routine, habits, etc.), Y_m will be a map to the same STC symbol, making the uncertainty in Y small, leading to *decreasing* entropy ($H(s_n)$). This concept is illustrated by an example in Fig. 2. Let the first recorded spatio-temporal data be mapped to symbol '7', resulting in $H(s_1) = 0$ at $n = 1$. When the second data point maps to symbol '13', the entropy increases to $H(s_2) = 1$ due to increased randomness in data. Now, let us say the third record maps to symbol '13' again, which results in a decrease in entropy from 1 at $n = 2$ to $H(s_3) = 0.918$. From this discussion, we can conclude that if the spatio-temporal entropy increases, the new measurement is from a non-sensitive location and can be shared with the application server (without SMC); however, if the new data point leads to a decrease in entropy, the location where the measurement was taken is personal and sensitive and should be transmitted through SMC only.

3.3. Inferring location-time sensitivity through KL divergence

The spatial-temporal data mapped onto various STC bins can be represented using histograms. The ideal scenario would be one where the application server or malicious adversary cannot infer sensitive information from the user location data. That would be

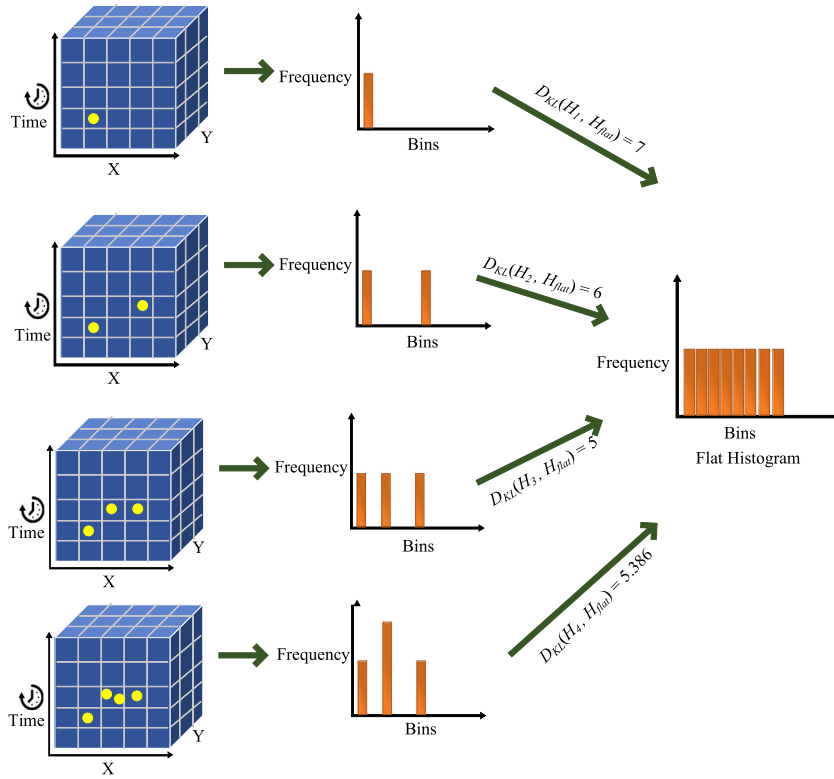


Fig. 3. An illustration showing KL divergence distance between histogram of a participants STCs and a flat histogram.

the case when this histogram is flat, i.e., all the STC bins are equally likely, and the user can be at any location at any time. Hence, we use the distance or dissimilarity of the users' actual histogram of space-time data with a flat histogram to check whether the user is at a sensitive private location. In this paper, we use KL divergence as a distance measure between users' actual histogram of space-time data and a flat histogram. The KL divergence finds non-symmetric distance or dis(similarity) measures between two n -dimensional probability distributions $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$ and $\beta = (\beta_1, \beta_2, \dots, \beta_n)$ by generalizing the classical Shannon's entropy and is given as [59]:

$$D_{KL}(\alpha, \beta) = \sum_{i=1}^n \alpha_i \log \left(\frac{\alpha_i}{\beta_i} \right) - \alpha_i + \beta_i. \quad (3)$$

Let the histogram generated after mapping each space-time point measured by a user to a STC symbol be represented by $H = (H_1, H_2, \dots, H_n)$ and the flat histogram is represented by $H_{flat} = [1, 1, \dots, 1]_{1 \times n}$, where $n = i \times j \times k$. The distance measure can be computed as:

$$D_{KL}(H_1, H_{flat}) = \sum_{n=1}^{i \times j \times k} H_{1_n} \log \left(\frac{H_{1_n}}{H_{flat_n}} \right) - H_{1_n} + H_{flat_n} \quad (4)$$

The above expression indicates that when the participant's measured space-time at time instant t_n maps to a different symbol than the one mapped at past instances, it will decrease the distance between its STC histogram and the flat histogram, and the current measurement can be shared without SMC. However, when a participant is at the same location for a prolonged time, it will map the sensed data to the same STC symbol, resulting in an increase in distance between histograms, indicating increased dissimilarity among them. For such measurements, SMC is needed to ensure privacy. This concept is illustrated using a sample example in Fig. 3. Let us say the first measured space-time point maps to symbol '1', resulting in $D_{KL}(H_1, H_{flat}) = 7$. Then, the second and third point maps to symbols '5' and '3' respectively, resulting in $D_{KL}(H_2, H_{flat}) = 6$ and $D_{KL}(H_3, H_{flat}) = 5$, i.e. distance decreases as randomness in data increases. Now, the fourth point maps to symbol '3' again, which results in an increase in KL-divergence distance by 0.386 (from 5 to 5.386), indicating symbol '3' is a private location as it is frequent and a malicious adversary could infer sensitive information from it if leaked.

From the aforementioned discussion, it can be inferred that when a new measurement is taken by a participant's mobile device, it updates the STC histogram of the location-time transmitted through direct model updates (likely to be inferred), and if the new STC histogram is *closer* to the flat histogram (the distance between updated STC histogram and a flat histogram *decreases*), then that measurement is transmitted through direct model update. On the other hand, if the new measurement is at a sensitive location time,

it is likely to increase the weight of an already bulky STC bin, thus increasing its distance or dissimilarity from a flat histogram, and should be communicated through SMC only.

4. Strategic location exposure framework for protection against inference attack

This section presents two methods for secure federated learning that does not compromise accuracy and reduces the communication cost of participants in urban sensing, as discussed below.

4.1. Spatiotemporal entropy based

The spatiotemporal entropy is calculated for each participant's sensed data using Eq. (2). When new data is recorded at time instant t_n , the relative change in entropy can be calculated using:

$$\Delta_{H(s_n)} = \frac{H(s_n) - H(s_{n-1})}{H(s_{n-1})} \quad (5)$$

where, $H(s_n)$ and $H(s_{n-1})$ represents entropy at time instants t_n and t_{n-1} respectively. $\Delta_{H(s_n)} > 0$ indicates increased uncertainty or randomness in data samples, implying non-sensitive location-time. An advisory will not be able to use this location to infer participants' personal attributes such as home or office location or daily routine. However, when $\Delta_{H(s_n)} \leq 0$, the sensed data pertains to the user's sensitive location-time, thus requiring enhanced privacy.

4.2. KL divergence distance based

In this approach, when new data is recorded at time instant t_n , the relative change in KL-divergence distance between the STC histogram and the flat histogram is calculated using:

$$\Delta_{D_{KL}(H_n, H_{flat})} = \frac{D_{KL}(H_{t_n}, H_{flat}) - D_{KL}(H_{t_{n-1}}, H_{flat})}{D_{KL}(H_{t_{n-1}}, H_{flat})}, \quad (6)$$

where, $D_{KL}(H_{t_n}, H_{flat})$ and $D_{KL}(H_{t_{n-1}}, H_{flat})$ represents KL divergence or distance measure between mapped and flat histogram at time instants t_n and t_{n-1} respectively. $\Delta_{D_{KL}(H_n, H_{flat})} < 0$ indicates increased randomness or uncertainty in the space-time data due to the current measurement. Whereas, $\Delta_{D_{KL}(H_n, H_{flat})} \geq 0$ indicate increased distance from the flat histogram, implying more concentrated space-time points which can be used to infer private information. Algorithms 1 and 2 shows the pseudo-codes of the two proposed schemes. Fig. 4 is a schematic of the proposed scheme for urban sensing applications, which reduces the communication and computation cost immensely and preserves participants' sensitive data against inference attacks. The proposed protocol employs two user defined thresholds $\delta_1 > 0$ and $\delta_2 < 0$ to fine-tune the following two competing requirements of an urban sensing system:

1. User privacy, and
2. Computation and communication costs.

The participating users will directly transmit (without any SMC) the measurements to the application server if $\Delta_{H(s_n)} > \delta_1$ or $\Delta_{D_{KL}(H_n, H_{flat})} < \delta_2$. This exposure will not help the application server or a malicious adversary infer personal and sensitive locations such as home or office or routine and habits of the participant. In fact, it will confuse the adversary by increasing randomness in the exposed location traces. Additionally, it will save significant computation and communication resources for resource-constrained users. The rest of the measurements are transmitted through SMC to avoid leakage of sensitive location-time data.

Algorithm 1: Spatiotemporal entropy based

Input: (latitude, longitude, time, parameter)
Output: Transmission of input for direct model based or SMC approach

- 1: Map input to the 3-D space-time cube (STC)
- 2: Calculate the spatio-temporal entropy, $H(s_n)$
- 3: **if** $H(s_{n-1}) \neq 0$ **then**
- 4: Calculate relative change in entropy, $\Delta_{H(s_n)}$
- 5: **end if**
- 6: **if** $\Delta_{H(s_n)} > \delta_1$ **then**
- 7: Transmit input via direct model-based approach
- 8: **else**
- 9: Transmit input via SMC approach
- 10: **end if**

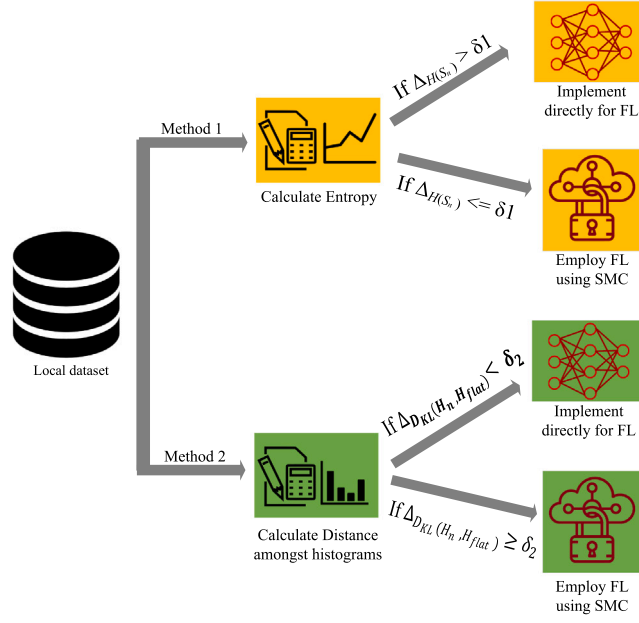


Fig. 4. Proposed scheme for strategic exposure of participants' location data for enhanced protection against inference attack in federated learning.

Algorithm 2: KL divergence distance based

Input: (latitude, longitude, time, parameter)

Output: Transmission of input for direct model based or SMC approach

- 1: Map input to the 3-D space-time cube (STC)
 - 2: Calculate the KL divergence between STC and flat histogram, $D_{KL}(H_1, H_{flat})$
 - 3: **if** $D_{KL}(H_{t_{n-1}}, H_{flat}) \neq 0$ **then**
 - 4: Calculate relative change in KL divergence, $\Delta_{D_{KL}}(H_n, H_{flat})$
 - 5: **end if**
 - 6: **if** $\Delta_{D_{KL}}(H_n, H_{flat}) < \delta_2$ **then**
 - 7: Transmit input via direct model-based approach
 - 8: **else**
 - 9: Transmit input via SMC approach
 - 10: **end if**
-

5. Privacy and security analysis

In this section, we theoretically analyze and prove the security of our proposed schemes. Our FL setting consists of one central cloud server and N participants. The threat model is considered where participants and server are honest-but-curious. The server and participants will honestly follow the protocol and not modify their model architectures to fit their attack better, nor will they provide malicious model updates. However, the server and participants are expected to be curious and try to infer useful information from other users.

An urban human mobility model uses mathematical or physical models to represent human movement's fundamental characteristics and features. Numerous urban mobility models have been proposed in the literature to understand and predict human mobility with high accuracy. A few of the popular models for human mobility are the random waypoint mobility model [60], power-law based [61–64], exponential distribution based [65,66], and gravity based [67] distributions.

Considering any mobility model, let the 3-dimensional spatio-temporal density of a user's space-time measurements be given by $f_{Latitude, Longitude, Time}(lat, long, time)$ where,

$$\begin{aligned} lat_{min} &\leq lat \leq lat_{max}, \\ long_{min} &\leq long \leq long_{max}, \\ 0 &\leq time \leq 24, \end{aligned} \tag{7}$$

where, lat_{min} , lat_{max} , $long_{min}$, and $long_{max}$ are the minimum and maximum values of the latitude and longitude traveled by the user, respectively. In the context of urban mobility modeling, it is commonly observed that the density function $f(\cdot)$ exhibits higher values

in proximity to residential areas during nighttime, near office locations during the daytime, and intermediate values during periods associated with market or leisure activities. Conversely, the function assumes smaller values for space-time values corresponding to infrequently visited or rarely explored locations.

Most state-of-the-art methods for sensitive location inference from leaked location traces use clustering of space-time traces to analyze user trajectories and extract meaningful locations [68,69]. The majority of the approaches are based on grid-based clustering [70], partition-based clustering [71], hierarchical clustering [72], or density-based clustering [40,41,73]. Among all, density-based clustering is the most widely used method since it can detect clusters of any shape. The majority of the density-based clustering algorithms proposed in the literature are modified versions of the popular DBSCAN algorithm, which detects clusters based on two user-defined parameters: $minPoints$ and ϵ . For the DBSCAN algorithm to identify a significant location from space-time traces, there must be more than $minPoints$ data points within a sphere of radius ϵ centred around any spatio-temporal data point. Hence, for DBSCAN to be able to detect a sensitive location from a user location traces, the following equation must be satisfied:

$$\iiint_{\substack{\|(lat, long, time) \\ -(x, y, t)\| \leq \epsilon}} f(.) \geq minPoints \quad (8)$$

where, $f(.)$ is the 3-dimensional spatio-temporal density of a user's space-time measurements, which can be a function of travel distance, travel speed or home location and (x, y, t) represents any space-time point within a sphere centered at $(lat, long, time)$, and having a radius of ϵ .

Conversely, to avoid inferring sensitive locations from leaked space-time traces, we want to expose only those data points that meet the following criteria:

$$\iiint_{\substack{\|(lat, long, time) \\ -(x, y, t)\| \leq \epsilon}} f(.) < minPoints. \quad (9)$$

Hence, the remaining points, i.e.

$$\iiint_{\substack{\|(lat, long, time) \\ -(x, y, t)\| \leq \epsilon}} f(.) - minPoints$$

should be transmitted through secure multi-party computation. The following theorems prove the security guarantees of the schemes proposed in this paper. With the given conditions met, it ensures that our schemes protect users' sensitive location-time information and can successfully avert inference attacks.

Theorem 1. For a given user, let at time instance $n - 1$, the distribution of space-time points in various space-time cubes (STC) be α_m ($1 \leq m \leq i \times j \times k$), and at the next time instant n , the new space-time point maps to the STC q . Using scheme 1 based on spatiotemporal entropy, the new data point at time n will be transmitted directly (without SMC) when the following condition is met:

Proof. For a point to be transmitted directly the relative change in entropy, $\Delta_{H(S_n)} = \frac{H(s_n) - H(s_{n-1})}{H(s_{n-1})} > \delta_1$ is calculated using Eq. (5). Where,

$$H(S_{n-1}) = \sum_{\substack{m=1 \\ m \neq q}}^{i \times j \times k} \frac{\alpha_m}{n-1} \log \left(\frac{n-1}{\alpha_m} \right) + \frac{\alpha_q}{n-1} \log \left(\frac{n-1}{\alpha_q} \right) \quad (10)$$

$$H(S_n) = \sum_{\substack{m=1 \\ m \neq q}}^{i \times j \times k} \frac{\alpha_m}{n} \log \left(\frac{n}{\alpha_m} \right) + \frac{\alpha_q + 1}{n} \log \left(\frac{n}{\alpha_q + 1} \right) \quad (11)$$

By substituting the values, we obtain the following condition on α_q for the new data point at time n to be transmitted directly, without using SMC:

$$\alpha_q = \iiint_{\substack{\|(lat, long, time) \\ -(x, y, t)\| \leq \epsilon}} f(.) \quad (12)$$

$$\leq ne^{-\delta_1 n \sum_{\substack{m=1 \\ m \neq q}}^{i \times j \times k} \frac{\alpha_m}{n} \log \left(\frac{n}{\alpha_m} \right)} < minPoints \quad (13)$$

where, $\delta_1 \geq 0$. The value of δ_1 is determined by the user's desired level of privacy and available resources.

The detailed proof of Theorem 1 is given in Appendix A: Section A of the supplementary material. Therefore, if Eqs. (12) and (13) hold true, it indicates increased randomness in data, and the adversary will not be able to infer participants' personal attributes such as home or office location or daily routine. However, if the above condition fails, then the point needs to be transmitted through SMC.

Theorem 2. For a given user, let at time instance $n - 1$, the distribution of space-time points in various space-time cubes (STC) be α_m ($1 \leq m \leq i \times j \times k$), and at the next time instant n , the new space-time point maps to the STC q . Using scheme 2 based on KL divergence, the new data point at time n will be transmitted directly (without SMC) when the following condition is met:

Proof. For a point to be transmitted directly the relative change in KL-divergence distance between the STC histogram and the flat histogram,

$$\Delta_{D_{KL}(H_n, H_{flat})} = \frac{D_{KL}(H_n, H_{flat}) - D_{KL}(H_{n-1}, H_{flat})}{D_{KL}(H_{n-1}, H_{flat})} < \delta_2 \text{ is calculated using Eq. (6). Where,}$$

$$D_{KL}(H_{n-1}, H_{flat}) = \sum_{\substack{m=1 \\ m \neq q}}^{i \times j \times k} \frac{\alpha_m}{n-1} \log \left(\frac{\frac{\alpha_m}{n-1}}{\frac{1}{u}} \right) - \frac{\alpha_m}{n-1} + \frac{1}{u} + \frac{\alpha_q}{n-1} \log \left(\frac{\frac{\alpha_q}{n-1}}{\frac{1}{u}} \right) - \frac{\alpha_q}{n-1} + \frac{1}{u} \quad (14)$$

$$D_{KL}(H_n, H_{flat}) = \sum_{\substack{m=1 \\ m \neq q}}^{i \times j \times k} \frac{\alpha_m}{n} \log \left(\frac{\frac{\alpha_m}{n}}{\frac{1}{u}} \right) - \frac{\alpha_m}{n} + \frac{1}{u} + \frac{\alpha_q + 1}{n} \log \left(\frac{\frac{\alpha_q + 1}{n}}{\frac{1}{u}} \right) - \frac{\alpha_q + 1}{n} + \frac{1}{u} \quad (15)$$

where, $u = i \times j \times k$. By substituting the values, we obtain the following condition on α_q for the new datapoint at time n to be transmitted directly, without using SMC:

$$\alpha_q = \iiint_{\substack{\|(lat, long, time) \\ -(x, y, t)\| \leq \epsilon}} f(\cdot) \quad (16)$$

$$\leq \frac{n}{u} e^{\delta_2(n-1)} \left(\sum_{\substack{m=1 \\ m \neq q}}^{i \times j \times k} \frac{\alpha_m}{n} \log \left(\frac{\alpha_m}{n} \right) - \frac{\alpha_m}{n} \right) < minPoints \quad (17)$$

where, value of $\delta_2 \leq 0$ and depends upon the resource availability and the level of privacy that a user requires.

The detailed proof of Theorem 2 is given in Appendix A: Section B of the supplementary material. The validation of the Eqs. (16) and (17) shows increased data randomness, preventing adversaries from inferring personal attributes such as home or office location and daily routines. The finding demonstrates the ability of the proposed approach in protecting the privacy and safeguarding sensitive user information from malicious adversaries. However, if the above condition fails, then the current data point needs to be transmitted through SMC.

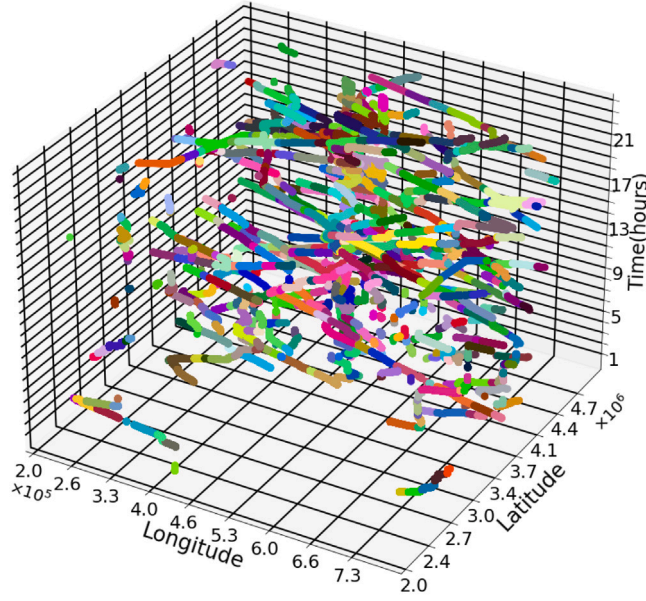
6. Experimental evaluation

In urban sensing systems, the data is spatiotemporal in nature. Each recorded data consists of location (latitude, longitude), time (time-stamp) information, and various other physical parameters for effective modeling of various urban phenomena. Moreover, for effective real-time modeling of the phenomenon of interest, the data need to be recorded at frequent intervals. To validate the proposed schemes, we conducted numerical experiments on the *geolife GPS trajectories* dataset [74–76]. The dataset includes GPS trajectories of 182 users in Beijing, China, from April 2007 to August 2012. The dataset comprises a total of 17,621 trajectories recorded by GPS phones and GPS loggers on different individuals. Each trajectory is represented by a sequence of data points consisting of latitude, longitude and time-stamp information. Approximately 91% of trajectories are recorded every 1~5 s that, includes users' home and work locations and outdoor movements such as sports activities and entertainment. Given the nature of the trajectories in the dataset, which includes both sensitive and non-sensitive user data, and since the approaches proposed in this paper aim to avoid leakage of sensitive user locations, this dataset is suitable for evaluating the proposed protocol. On the basis of data recorded by users, we have categorized them into four categories, viz., *small contribution*, *moderate contribution*, *large contribution* and *enormous contribution*. The small contribution users have <1000 GPS data logs, and moderate contribution ones have the number of data points in the range [1000, 10000). Similarly, large contribution users have more than 10000, but less than 100000 GPS points, and users having enormous contribution have >100000 raw GPS data logs. A preliminary analysis of the geolife dataset reveals that the number of users belonging to small, moderate, large and enormous contribution categories is 13.18%, 22.5%, 35.16% and 29.12% respectively of the total number of the users in the dataset.

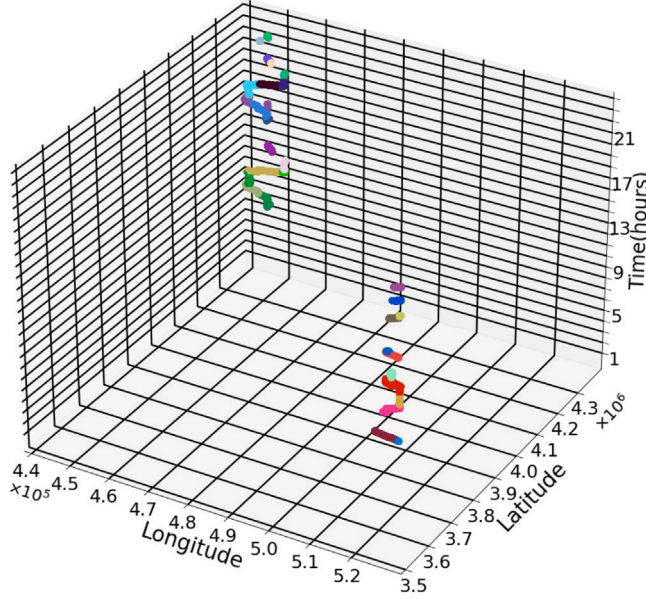
As a pre-processing step, we convert the latitude and longitude coordinates to *universal transverse mercator* (UTM) coordinates to map onto a 10×10 2-dimensional grid. We also convert the time-stamp corresponding to each data point into the hour of the day, i.e., one of the integer values $\{0, 1, 2, \dots, 23\}$. For spatiotemporal data binning, we map each user's space-time data onto one of the $10 \times 10 \times 24$ STC symbols. The experiments were performed on a workstation with a 2.9 GHz Xeon(R) Gold Intel processor, a 64-bit Windows 11 Pro operating system, and 256 GB of RAM. The proposed algorithm was implemented in Python 3 programming language.

6.1. Statistical analysis of effectiveness of the proposed scheme

In this section, we statistically demonstrate the number of contributions having positive and negative values of the relative change in spatiotemporal entropy and KL divergence distance for two categories of users, one having enormous amount of contribution, and one having moderate amount of data points. Fig. 5 shows spatiotemporal data binning for two users, one belonging to the enormous category and the other to the moderate category. The user in Fig. 5(a) has a massive 935,000 spatiotemporal measurements, whereas the one in Fig. 5(b) has only 9700 data points mapped to $10 \times 10 \times 24$ STC symbols shown by different colors. Here, different color denotes different STC symbols assigned to various data points. The data points having the same color are the one lying within the same STC and are considered as a single symbol for entropy and KL divergence calculation.



(a) Participant with enormous amount of spatiotemporal data traces
(~ 935,000 datapoints)

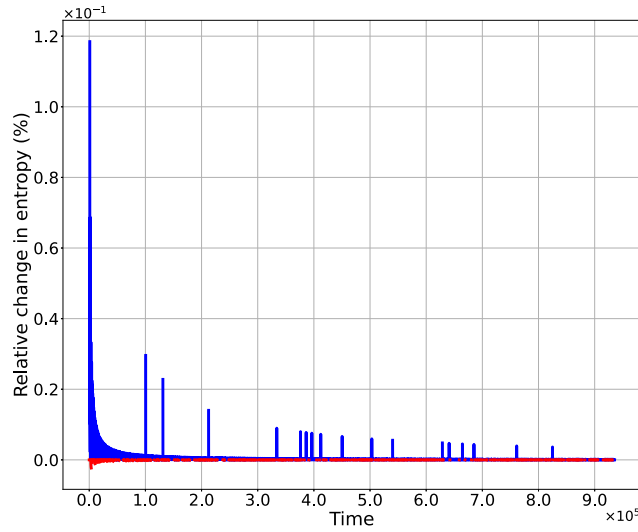


(b) Participant with moderate amount of spatiotemporal data traces
(~ 9,700 datapoints)

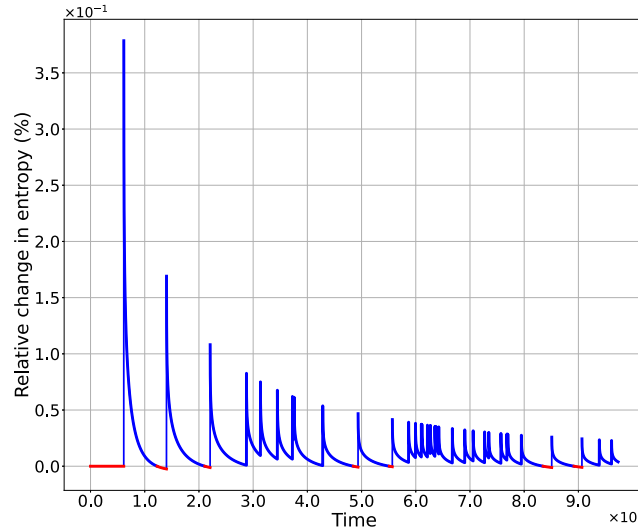
Fig. 5. Space time cube (STC) binning of two sample participants corresponding to enormous and moderate amount of collected data. Different color represents different STC symbol being assigned to data points.

6.1.1. Relative change in spatiotemporal entropy

After mapping the sensed data point onto STC symbols, entropy ($H(s_n)$) for each new datapoint arrival (as a function of time t_n) is calculated using Eq. (2). Then, the relative change in entropy, $\Delta_{H(S_n)}$ is evaluated using Eq. (5). Fig. 6 plots the relative change in entropy, $\Delta_{H(S_n)}$ of the two participants (whose GPS traces are shown in Fig. 5). It can be noticed that for the user with the enormous amount of recorded data (Fig. 5(a)), $\Delta_{H(S_n)}$ dips below 0 only for a small portion of contributions and is positive for most of the measurements. Statistically, only 25.65% of the measurements has $\Delta_{H(S_n)} < 0$, and for the remaining 74.35%, we observe a positive relative change in spatiotemporal entropy. For the user with the moderate amount of data (Fig. 5(b)), the relative change in entropy, $\Delta_{H(S_n)}$ is shown in Fig. 6(b), the number of data points having a positive change in spatiotemporal entropy is even higher and stands



(a) Participant with enormous amount of spatiotemporal data traces



(b) Participant with moderate amount of spatiotemporal data traces

Fig. 6. Relative change in entropy $\Delta_{H_{S_n}}$ of participants corresponding to enormous and moderate amount of collected data as a function of time. $\Delta_{H_{S_n}} > 0$ shown by blue and $\Delta_{H_{S_n}} < 0$ shown by red.

at 92.07%, whereas only 7.92% of the contributions has a negative relative change in entropy. The user with an enormous amount of data traces will have more sensitive and routine locations included in their data traces as compared to the user with a moderate amount of data. Moreover, when the user has a smaller amount of data or scattered data points, a higher percentage of points will account to positive $\Delta_{H_{S_n}}$, as the majority of the points will help to attain a uniform distribution in STCs due to the smaller amount of data. An analysis of the contribution of all the participants reveals that for almost three-fourth of data points, $\Delta_{H_{S_n}} > 0$, and they can be transmitted directly, while the remaining one-fourth requiring secure multi-party computation for the distributed model training in the FL framework as shown in Fig. 7. A point to note: the data points transmitted through direct model updates need not necessarily be non-frequent location-time data points but could be from sensitive locations as long as the number of data points in most of the STC bins is almost equal so that the density-based approaches to detect sensitive location and routines become ineffective. Thus, a higher percentage of points will result in positive $\Delta_{H_{S_n}}$ for all categories of users.

6.1.2. Relative change in KL divergence

We computed generalized KL divergence $D_{KL}(H_{t_n}, H_{flat})$ for each new datapoint arrival (as a function of time t_n) as given in Eq. (4). Then, a relative change in KL-divergence distance between the STC histogram and the flat histogram, $\Delta_{D_{KL}(H_n, H_{flat})}$ is

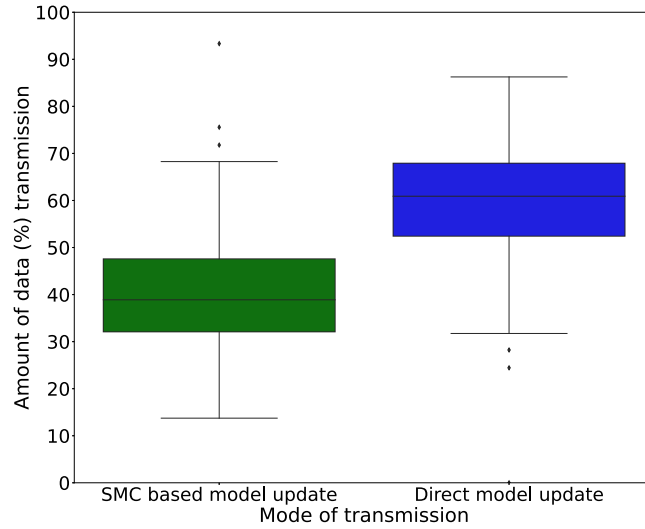


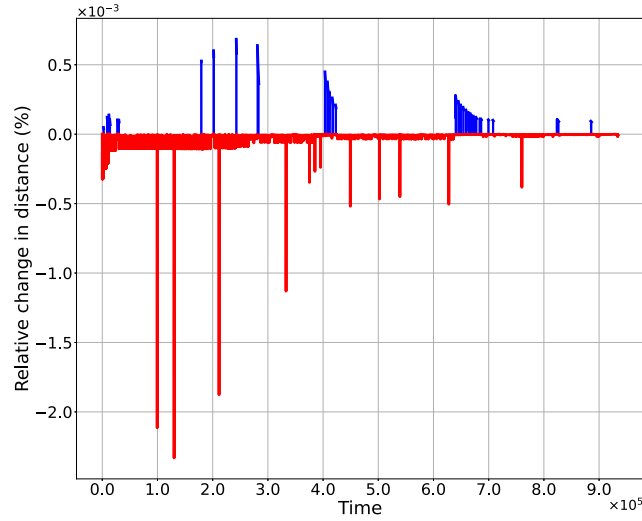
Fig. 7. Amount of data transmission through direct and secure multi-party computation by implementing proposed scheme 1 (based on spatiotemporal entropy).

evaluated for all 182 users using Eq. (6). Fig. 8 plots $\Delta_{D_{KL}(H_n, H_{flat})}$ of the two participants (Negative $\Delta_{D_{KL}(H_n, H_{flat})}$ shown by red and positive $\Delta_{D_{KL}(H_n, H_{flat})}$ shown by blue), whose contributions are shown in Fig. 5. It can be observed that for the user with the enormous amount of contributed data (Fig. 5(a)), only for very few time instances ($\sim 3.2\%$ of the total data points), $\Delta_{D_{KL}(H_n, H_{flat})}$ is positive, whereas, for a majority (96.8%) of the contributions, the STC histogram of the participant become flatter with the addition of a new data point. Similarly, for a user with moderate contributions (Fig. 5(b)), $\Delta_{D_{KL}(H_n, H_{flat})}$ is positive for just about 2% of the contributions, whereas almost 98% contributions has a negative relative change in KL divergence distance implying a flatter STC histogram. The user with an enormous amount of data has more sensitive as well as non-sensitive locations, and therefore, to make the histogram flatter, a higher percentage of points will require SMC as compared to a user with a moderate amount of data. The aggregated results from all 182 participants are shown in Fig. 9, which illustrates that around 60% of the data can be transmitted directly since $\Delta_{D_{KL}(H_n, H_{flat})} < 0$, with the remaining 40% requiring secure multi-party computation. All four categories of users consist of sensitive as well as non-sensitive location-time data, and therefore, to make the histograms flatter, a majority of percentage of data results in negative $\Delta_{D_{KL}(H_n, H_{flat})}$ and can be transmitted through direct model updates.

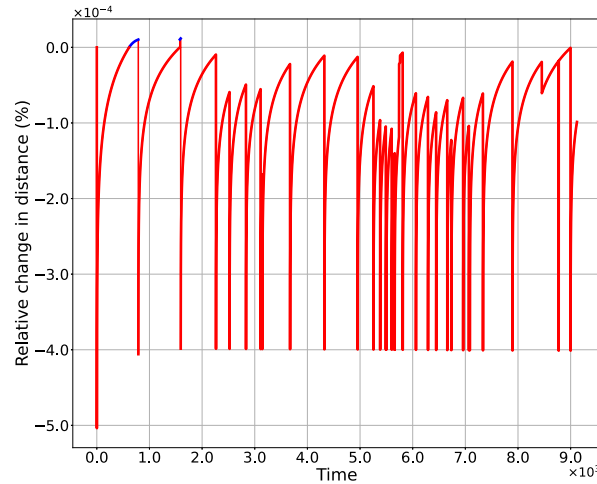
6.2. Protection against private and sensitive location inference

Many works [40,41,77,78] have employed clustering algorithms like DBSCAN and its variants on spatiotemporal data logs to infer a participant's sensitive locations such as home or workplace, posing privacy risks to participants. DBSCAN is a clustering algorithm that clusters data points based on their proximity and density characteristics by setting two parameters: radius (ϵ) and minimum points ($minPoints$). The ϵ refers to the maximum distance required for data points to be considered neighbors, while $minPoints$ denotes the minimum number of data points needed within the ϵ neighborhood of a given data point. The clustered data is analyzed to identify the participants' behavior. Other Markov chain based techniques [42,43] launch de-anonymization attacks to infer the participant's mobility traces prior to clustering based sensitive location detection. To demonstrate the effectiveness of our proposed scheme, we launch inference attacks by employing clustering based sensitive location detection algorithms such as SNN+ [41], DBSCAN and EDBSCAN [40] on raw data of participants (inferred location traces in case of FL model update without using SMC) as well as on data strategically leaked to the application server in the proposed scheme. Other techniques, such as MMC [42] and UHMM [43] are supervised learning algorithms and hence, are not comparable for urban sensing tasks for which training data is not available.

Table 1 shows the effectiveness of the proposed spatiotemporal entropy and KL divergence based strategic location leaking approach in protecting the participants against the inference of their private and sensitive locations. We have shown the results for four categories of users, viz., participants having *small* (User ID 87), *moderate* (user ID 45), *large* (user ID 11), and *enormous* (User ID 119) amount of spatiotemporal data contribution. The second column of Table 1 displays (by different colors) the clusters obtained by applying the DBSCAN clustering algorithm to the entire raw data of respective users, which is obtained by the adversary using inference or model inversion attack. The DBSCAN parameters used for clustering of four categories of users are $\epsilon = 0.1$ and $minPoints = 5\%, 2\%, 1\%$ and 0.5% of total GPS logs recorded by a participant, respectively. The values of ϵ and $minPoints$ are chosen such that only a large group of compact points are declared as clusters, and the small ones are declared noise points (indicating only sensitive locations such as home or workplace are detected). Various colors represent different clusters (except black, which represents noise points), which are then analyzed to infer home/office location and routines based on the time profile of each cluster. The third and fifth column shows the 3-D scatterplot of the strategically leaked points using the two approaches proposed



(a) Participant with enormous amount of spatiotemporal data traces



(b) Participant with moderate amount of spatiotemporal data traces

Fig. 8. $\Delta_{D_{KL}(H_n, H_{flat})}$ of participants corresponding to enormous and moderate amount of collected data as a function of time. Negative values shown by red and positive values shown by blue.

in the paper, viz., spatiotemporal entropy based, and KL divergence distance based, respectively. Applying the DBSCAN clustering algorithm using the same parameters (as applied on raw data) on these leaked data logs results in clusters shown in the fourth and sixth column of [Table 1](#) respectively. As earlier, different colors represent different clusters, and the noise points are shown in black. After applying our proposed techniques, it can be observed from [Table 1](#) that the number of detected clusters (likely candidates for sensitive and private locations) significantly reduced, e.g., by 66.67%, 80%, 80% and 50% for users' belonging to *small* (User ID 87), *moderate* (User ID 45), *large* (User ID 11) and *enormous* (User ID 119) category respectively. [Table 1](#) shows that the raw data can be easily used to infer sensitive and private locations and mobility patterns from various clusters detected from the raw data shown in the second column. Whereas, the data points which are intentionally leaked to the server in the proposed approach, contributing to an increase in spatiotemporal entropy or decrease in the KL divergence distance, are scattered, making it difficult for a clustering algorithm to find clusters; hence, an adversary may not be able to infer sensitive locations from these traces. The threshold for relative change in entropy, $\Delta_{H(S_n)}$ and for relative change in distance, $\Delta_{D_{KL}(H_n, H_{flat})}$ depends on distribution of data points in spatiotemporal bins. Participants contributing a small amount of raw data have scattered and fewer points in a bin than participants contributing large or enormous amounts of raw data. This implies that the threshold value for the relative change in entropy, $\Delta_{H(S_n)}$ and the relative change in distance, $\Delta_{D_{KL}(H_n, H_{flat})}$ should decrease as the data distribution becomes more compact in spatiotemporal bins, and vice versa. For consistency of the experiments reported in this paper, we have set both $\Delta_{H(S_n)}$ and $\Delta_{D_{KL}(H_n, H_{flat})}$ to be 0.

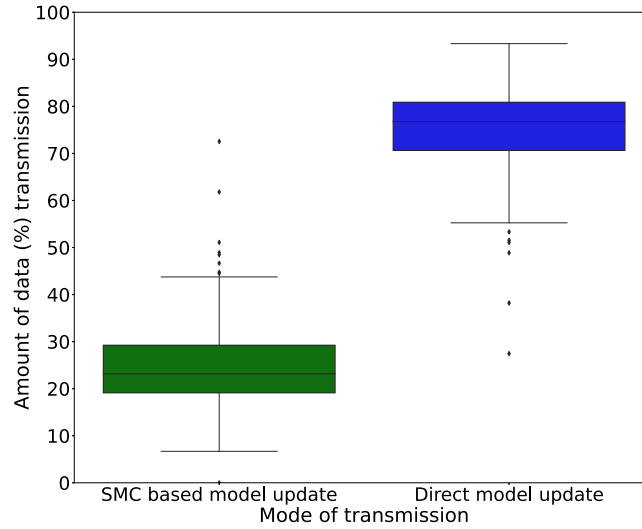


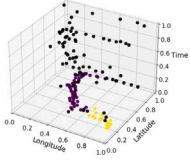
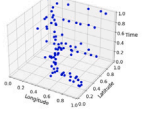
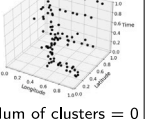
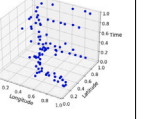
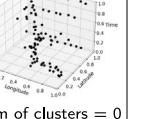
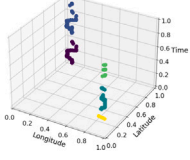
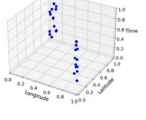
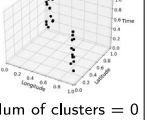
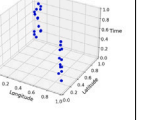
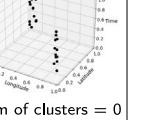
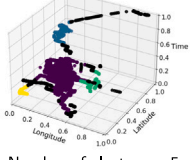
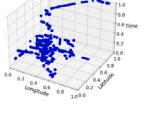
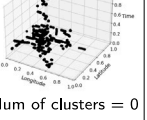
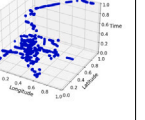
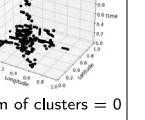
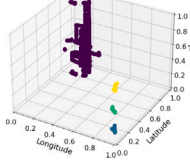
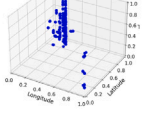
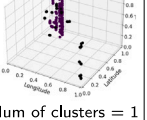
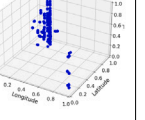
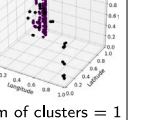
Fig. 9. Amount of data transmission through direct and secure multi-party computation by implementing proposed scheme 2 (based on KL divergence).

A similar computation has been done using SNN+ and EDDSCAN algorithms for the same set of data. However, SNN+ cannot be applied to users of the *enormous* category as the algorithm does not support clustering of huge spatiotemporal data sets [41]. Table 2 depicts the silhouette score, Davies–Bouldin score and the number of clusters for launching inference attacks on data transmission via traditional (raw data) and proposed schemes. The silhouette score is a metric that determines the quality of clusters in a dataset by computing the cohesiveness of points inside clusters and the separation between clusters. The score ranges from -1 to 1 , where a high value closer to 1 indicates that the clusters are well-defined and well-separated from each other [79]. The Davies–Bouldin score is a metric for assessing the quality of clusters generated using a clustering technique. It computes the average similarity between each cluster and its most similar cluster while considering distances between cluster centres. A lower Davies–Bouldin score implies more distinct and well-defined clusters [80]. A determination of zero clusters signifies that all data points have been categorized as noise points; hence, Silhouette and Davies–Bouldin scores cannot be computed. DBSCAN based approach detects 3, 5, 5, and 4 potential sensitive locations for *small* (User ID 87), *moderate* (User ID 45), *large* (User ID 11) and *enormous* (User ID 119) category user respectively with Davies–Bouldin scores in the range 0.12 and 0.48 and Silhouette score between 0.63 and 0.9 indicating dense clusters, which are strong contenders for participants home and office location. However, applying spatiotemporal entropy and KL divergence based selective sharing approaches proposed in this paper, DBSCAN is not able to detect any cluster for users 87 (small contribution), 45 (moderate contribution), and 11 (large contribution). Only one cluster is detected for user 119, having enormous contribution with a higher Davies–Bouldin score and lower Silhouette score as compared to the clusters obtained from raw data, indicating sparse clusters, which may not be strong contenders for sensitive locations such as home, office, or party. Similar results were obtained using SNN+ and EDDSCAN inference attack approaches on strategically leaked data according to spatiotemporal entropy or KL divergence distance criteria compared to the raw/inferred data. In SNN+, the number of detected clusters was reduced by 50% , 93.1% and 75% for users belonging to *small* (User ID 87), *moderate* (User ID 45), and *large* (User ID 11) category respectively. The number of detected clusters in EDDSCAN decreased by 50% , 50% , 66.67% and 33.33% for users' belonging to *small* (User ID 87), *moderate* (User ID 45), *large* (User ID 11), and *enormous* (User ID 119) category respectively. Moreover, the fewer detected clusters have higher Davies–Bouldin score and lower Silhouette score, indicating poor quality of the detected clusters, which may not be attributed to private and sensitive locations of the users with high confidence.

To demonstrate the effectiveness of the proposed approach in terms of reduced computation and communication cost for participating users, we performed a comparison of the number of users transmitting their locally updated model through SMC to guarantee maximum privacy as a function of time. In the conventional approach, for maximum privacy guarantee, all model updates are transmitted through SMC, whereas for the framework proposed in this paper, only a few updates need to be transmitted through SMC as shown in Fig. 10. This experiment is performed on the users who contributed at least one spatiotemporal data between 11 AM and 2 PM. We assume that the application model is updated every 5 min by all the participants, which are then aggregated by the application server to generate the time-varying global model. The number of participants during the duration of the experiment varied between 70 and 90. To achieve maximum privacy through the traditional approach, the local model updates by every user need to be performed through communication and computationally expensive SMC protocols every 5 min (model update frequency), as shown by tall blue bars for each 5-min cycle. The two approaches proposed in this paper significantly reduce the need for SMC to communicate local model updates by the participating users to the application server without any compromise on privacy. The much shorter orange and green bars in Fig. 10 show a significant reduction in the number of users requiring SMC for spatiotemporal entropy and KL divergence distance based approaches, respectively. For the entire duration of the experiment, the number of users who needed to transmit their local model updates by SMC decreased by 57.33% and 61.23% for spatiotemporal entropy and KL divergence distance based approaches, respectively.

Table 1

Effectiveness of the proposed approaches in nullifying sensitive location detection using DBSCAN clustering post inference attack. Different colors in columns 2, 4, and 6 indicate extracted clusters, except black, which indicate noise points.

User ID	Raw/inferred spatiotemporal data and clusters indicating sensitive locations	Strategically leaked spatiotemporal data and clusters indicating sensitive locations			
		Spatiotemporal entropy based		KL divergence distance based	
		Strategically leaked spatiotemporal data	clusters indicating sensitive locations	Strategically leaked spatiotemporal data	clusters indicating sensitive locations
87	 Number of clusters = 3	 Num of clusters = 0	 Num of clusters = 0	 Num of clusters = 0	 Num of clusters = 0
45	 Number of clusters = 5	 Num of clusters = 0	 Num of clusters = 0	 Num of clusters = 0	 Num of clusters = 0
11	 Number of clusters = 5	 Num of clusters = 0	 Num of clusters = 0	 Num of clusters = 0	 Num of clusters = 0
119	 Number of clusters = 4	 Num of clusters = 1	 Num of clusters = 1	 Num of clusters = 1	 Num of clusters = 1

Several SMC approaches have been proposed in the literature for communicating local model updates, of which the most prominent are the one proposed by Bonawitz et al. [46], Choi et al. [51], and Turbo-Aggregate [54] as discussed in Section 2. However, all these approaches have high communication costs for participating users, as given in Table 3. Specifically, SMC approach proposed by Bonawitz et al. [46] and Choi et al. [51] have time complexity of $O(n^2)$ and $O(n\sqrt{n\log n})$ respectively. Similarly, Turbo-Aggregate [54] has a computational complexity of $O(n\log n)$. Here, n is the total number of participants participating in the federated learning task. Of these three state-of-the-art approaches, Turbo-Aggregate [54] has the best time complexity and can handle the maximum number of user dropouts. Hence, we use Turbo-Aggregate in conjunction with the two proposed approaches in this paper to reduce computation and communication costs while providing maximum privacy protection of federated learning based urban sensing. The last two rows of Table 3 list the communication cost per user for spatiotemporal entropy and KL divergence distance based approaches, respectively, where, $0 < n_1 < n$ and $0 < n_2 < n$ local model updates are transmitted through Turbo-Aggregate SMC, whereas the remaining $n - n_1$ and $n - n_2$ local model updates are directly transmitted to the application server. Please note that the value of n_1 and n_2 would depend on the distribution of spatiotemporal points among different STC bins and the thresholds δ_1 and δ_2 . For the experiment shown in Fig. 10, a comparison of the number of communication rounds per user for various SMC algorithms and the approaches proposed in this paper is shown in Fig. 11. The number of SMC rounds per user for the approach proposed by Bonawitz et al. [46] is maximum and varies between 77 and 92 during the entire course of the experiment. Turbo Aggregate and

Table 2

Comparison of launching inference attacks on raw data versus data generated through proposed schemes in terms of silhouette score, Davies–Bouldin score and number of clusters.

Inference attack algorithm: DBSCAN				
User ID	Data transmission approach	Measured parameters		
		Number of clusters	Davies–Bouldin score	Silhouette score
87	Raw data	3	0.485	0.628
	Spatiotemporal entropy based	0	–	–
	KL divergence distance based	0	–	–
45	Raw data	5	0.120	0.905
	Spatiotemporal entropy based	0	–	–
	KL divergence distance based	0	–	–
11	Raw data	5	0.352	0.814
	Spatiotemporal entropy based	0	–	–
	KL divergence distance based	0	–	–
119	Raw data	4	0.22	0.734
	Spatiotemporal entropy based	1	0.676	0.687
	KL divergence distance based	1	0.676	0.687
Inference attack algorithm: SNN+				
User ID	Data transmission approach	Measured parameters		
		Number of clusters	Davies–Bouldin score	Silhouette score
87	Raw data	2	0.665	0.627
	Spatiotemporal entropy based	1	0.959	0.514
	KL divergence distance based	1	0.961	0.512
45	Raw data	29	0.105	0.950
	Spatiotemporal entropy based	2	0.271	0.823
	KL divergence distance based	2	0.272	0.822
11	Raw data	8	0.210	0.80
	Spatiotemporal entropy based	2	0.556	0.697
	KL divergence distance based	2	0.556	0.697
119	Experiment could not be performed due to high computational complexity of SNN+ and large amount of data for this user.			
Inference attack algorithm: EDDSCAN				
User ID	Data transmission approach	Measured parameters		
		Number of clusters	Davies–Bouldin score	Silhouette score
87	Raw data	4	0.422	0.649
	Spatiotemporal entropy based	2	0.499	0.595
	KL divergence distance based	2	0.499	0.596
45	Raw data	4	0.302	0.895
	Spatiotemporal entropy based	2	0.386	0.583
	KL divergence distance based	2	0.385	0.564
11	Raw data	3	0.319	0.815
	Spatiotemporal entropy based	1	0.591	0.542
	KL divergence distance based	1	0.590	0.542
119	Raw data	3	0.154	0.802
	Spatiotemporal entropy based	2	0.569	0.743
	KL divergence distance based	2	0.572	0.743

Choi et al. [51] are more efficient than Bonawitz et al. [46] but are not data adaptive. Of these two, Turbo Aggregate is more efficient and has approximately 20%–25% fewer communication rounds per user. The spatiotemporal entropy and KL divergence distance based approaches proposed in this paper are data adaptive and significantly reduce communication rounds per user compared to Turbo Aggregate. Over the duration of the experiment, the average number of communication rounds per user is just 8.9 and 8.5 for spatiotemporal entropy and KL divergence distance based along with Turbo Aggregate, respectively, as compared to 17.6 for only Turbo Aggregate, a reduction of more than 50%.

7. Limitations

In this paper, we have proposed an approach to protect private and sensitive location detection using clustering on the data inferred by the malicious server or an adversary. The proposed approach assumes the continuous spatiotemporal data being collected by the user and transmitted to the application server for urban sensing tasks. This assumption may encounter challenges in a

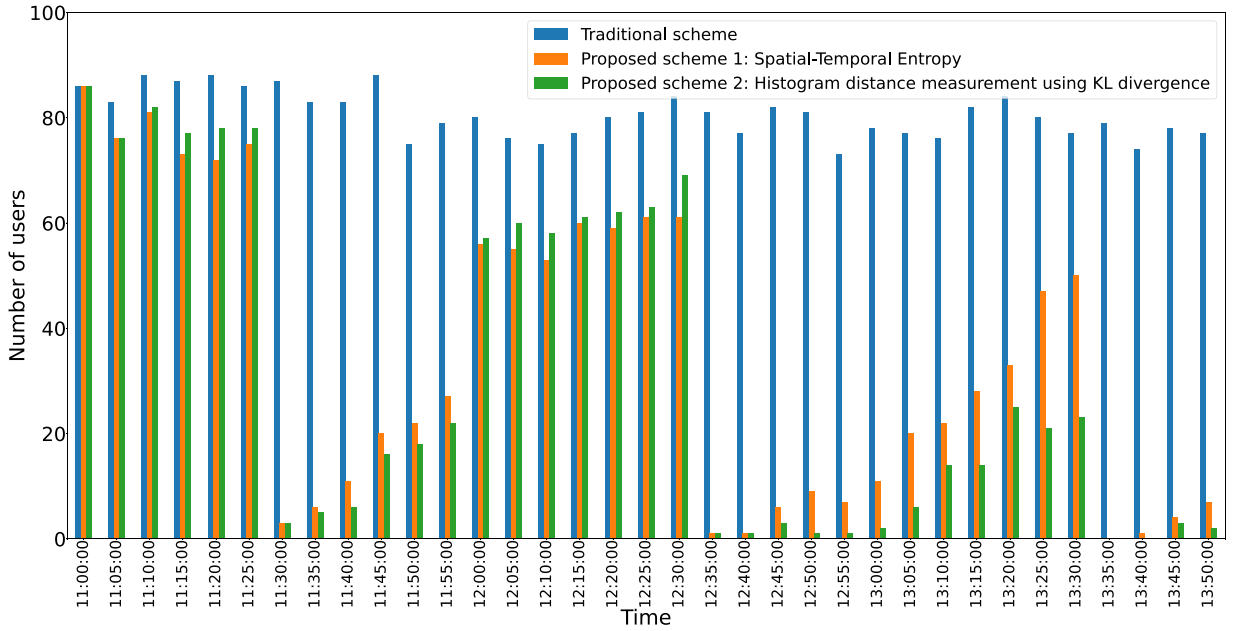


Fig. 10. Comparison of the number of participants engaging in SMC protocol for communicating local model updates for maximum privacy using conventional scheme and the approaches proposed in this paper.

Table 3

Communication cost per training iteration for various SMC algorithms and the approaches proposed in this paper.

Approach	Communication cost (per user)
Bonawitz et al. [46]	$O(n^2)$
Choi et al. [51]	$O(n\sqrt{n\log n})$
Turbo-Aggregate [54]	$O(n\log n)$
Spatial-Temporal Entropy + Turbo-Aggregate	$O(n_1\log n_1 + (n - n_1))$
Histogram distance using KL divergence + Turbo-Aggregate	$O(n_2\log n_2 + (n - n_2))$

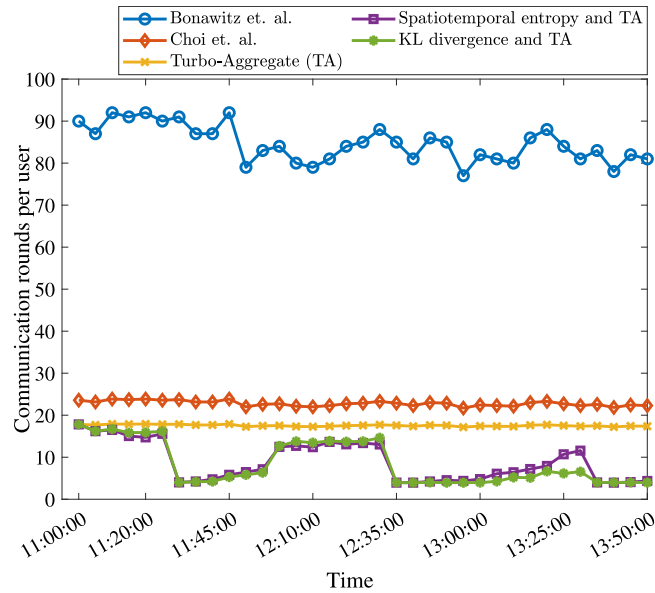


Fig. 11. Comparison of number of communication rounds per user for various SMC schemes and the two approaches proposed in this paper.

few real-world scenarios due to sensor failures in users' devices, data collection inconsistencies, and temporal gaps. Our work has not considered more advanced active inference attacks, such as generative adversarial network attacks. Further study and experimentation are needed to validate the effectiveness or limitation of the proposed approach against such attacks.

8. Conclusions and future work

This paper proposes and experimentally validates two data-adaptive approaches for preserving the privacy of participants at significantly lower communication and computation costs for federated learning based urban sensing applications. The proposed approaches strategically leaks selected location traces (through direct local model updates) to prevent a participant from inference attacks by confusing the application server or malicious adversary. The proposed schemes are based on spatiotemporal entropy and KL divergence distance that selectively leak those data points, which increases randomness in the data that can be inferred through inference attacks, making it difficult for the application server or malicious adversary to infer private and sensitive locations, routines and habits of participants. Numerical experiments on the Geolife trajectory dataset validate that the proposed schemes significantly save on participants' computation and communication costs, making them suitable for frequent model updates in dynamic urban sensing applications. Future work involves analyzing other approaches for inference attack mitigation, such as homomorphic encryption instead of SMC for urban sensing applications based on FL, in order to achieve a balance between privacy and computation/communication costs. We would also like to further explore and enhance the proposed approach by analyzing weekdays, weekends, holidays, and other events as separate entities, which could result in increased individual privacy protections and further reduce communication and computation overhead on participating mobile devices.

CRedit authorship contribution statement

Ayshika Kapoor: Conceptualization, Investigation, Methodology, Resources, Software, Writing – original draft. **Dheeraj Kumar:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

Authors declare no conflict of interest

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.pmcj.2024.101875>.

References

- [1] A. Turner, How many people have smartphones worldwide? 2022, <https://www.bankmycell.com/blog/how-many-phones-are-in-the-world>.
- [2] E. Kanjo, S. Benford, M. Paxton, A. Chamberlain, D.S. Fraser, D. Woodgate, D. Crellin, A. Woolard, MobGeoSen: Facilitating personal geosensor data collection and visualization using mobile phones, *Pers. Ubiquitous Comput.* 12 (8) (2008) 599–607.
- [3] S. Devarakonda, P. Sevusu, H. Liu, R. Liu, L. Iftode, B. Nath, Real-time air quality monitoring through mobile sensing in metropolitan areas, in: *ACM SIGKDD International Workshop on Urban Computing*, 2013.
- [4] W. Sun, Q. Li, C. Tham, Wireless deployed and participatory sensing system for environmental monitoring, in: *IEEE International Conference on Sensing, Communication, and Networking, SECON*, 2014, pp. 158–160.
- [5] D. Iskandaryan, F. Ramos, S. Trilles, Air quality prediction in smart cities using machine learning technologies based on sensor data: A review, *Appl. Sci.* 10 (7) (2020) 2401.
- [6] P. Mohan, V.N. Padmanabhan, R. Ramjee, V. Padmanabhan, TrafficSense: Rich Monitoring of Road and Traffic Conditions using Mobile Smartphones, *Tech. Rep. MSR-TR-2008-59*, 2008.
- [7] P. Mohan, V. Padmanabhan, R. Ramjee, Nericell: Rich monitoring of road and traffic conditions using mobile smartphones, in: *ACM Sensys*, 2008.
- [8] T. Das, P. Mohan, V.N. Padmanabhan, R. Ramjee, A. Sharma, PRISM: Platform for remote sensing using smartphones, in: *International Conference on Mobile Systems, Applications, and Services*, 2010, pp. 63–76.
- [9] F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira Jr., C. Ratti, Understanding individual mobility patterns from urban sensing data: A mobile phone trace example, *Transp. Res. C* 26 (2013) 301–313.
- [10] D. Shin, D. Aliaga, B. Tunçer, S.M. Arisana, S. Kim, D. Zünd, G. Schmitt, Urban sensing: Using smartphones for transportation mode classification, *Comput. Environ. Urban Syst.* 53 (2015) 76–86.
- [11] R.K. Rana, C.T. Chou, S.S. Kanhere, N. Bulusu, W. Hu, Ear-phone: An end-to-end participatory urban noise mapping system, in: *International Conference on Information Processing in Sensor Networks*, 2010, pp. 105–116.
- [12] N. Maisonneuve, M. Stevens, M.E. Niessen, L. Steels, NoiseTube: Measuring and mapping noise pollution with mobile phones, in: *Information Technologies in Environmental Engineering*, 2009, pp. 215–228.
- [13] E. Shim, D. Kim, H. Woo, Y. Cho, Designing a sustainable noise mapping system based on citizen scientists smartphone sensor data, *PLOS ONE* 11 (9) (2016) 1–7.
- [14] E. Kanjo, NoiseSPY: A real-time mobile phone platform for urban noise monitoring and mapping, *Mob. Netw. Appl.* 15 (4) (2010) 562–574.
- [15] I. Schweizer, R. Bärtil, A. Schulz, F. Probst, M. Mühlhäuser, NoiseMap - Real-time participatory noise maps, in: *International Workshop on Sensing Applications on Mobile Phones*, 2011, pp. 1–4.

- [16] J. Picaud, N. Fortin, E. Bocher, G. Petit, P. Aumond, G. Guillaume, An open-science crowdsourcing approach for producing community noise maps using smartphones, *Build. Environ.* 148 (2019) 20–33.
- [17] K. Shilton, Four billion little brothers? Privacy, mobile phones, and ubiquitous data collection, *Commun. ACM* 52 (11) (2009) 48–53.
- [18] M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, P. Boda, PEIR, the personal environmental impact report, as a platform for participatory sensing systems research, in: *International Conference on Mobile Systems, Applications, and Services*, 2009, pp. 55–68.
- [19] J. Konečný, H.B. McMahan, F.X. Yu, P. Richtárik, A.T. Suresh, D. Bacon, Federated learning: Strategies for improving communication efficiency, 2016, arXiv preprint arXiv:1610.05492.
- [20] J.C. Jiang, B. Kantarci, S. Oktug, T. Soyata, Federated learning in smart city sensing: Challenges and opportunities, *Sensors* 20 (21) (2020) 6230.
- [21] J. Zhang, Y. Liu, D. Wu, S. Lou, B. Chen, S. Yu, VPFL: A verifiable privacy-preserving federated learning scheme for edge computing systems, *Digit. Commun. Netw.* 9 (4) (2023) 981–989.
- [22] X. Cheng, B. He, G. Li, B. Cheng, A survey of crowdsensing and privacy protection in digital city, *IEEE Trans. Comput. Soc. Syst.* (2022).
- [23] S. Jain, S. Gupta, K. Sreelakshmi, J.J. Rodrigues, Fog computing in enabling 5G-driven emerging technologies for development of sustainable smart city infrastructures, *Cluster Comput.* (2022) 1–44.
- [24] Y. Huang, S. Gupta, Z. Song, K. Li, S. Arora, Evaluating gradient inversion attacks and defenses in federated learning, *Adv. Neural Inf. Process. Syst.* 34 (2021).
- [25] Z. Xiong, Z. Cai, D. Takabi, W. Li, Privacy threat and defense for federated learning with non-iid data in AIoT, *IEEE Trans. Ind. Inform.* 18 (2) (2021) 1310–1321.
- [26] X. Luo, Y. Wu, X. Xiao, B.C. Ooi, Feature inference attack on model predictions in vertical federated learning, in: *2021 IEEE 37th International Conference on Data Engineering, ICDE, IEEE*, 2021, pp. 181–192.
- [27] H. Hu, Z. Salic, L. Sun, G. Dobbie, X. Zhang, Source inference attacks in federated learning, in: *2021 IEEE International Conference on Data Mining, ICDM, IEEE*, 2021, pp. 1102–1107.
- [28] A. Pustozero, R. Mayer, Information leaks in federated learning, in: *Proceedings of the Network and Distributed System Security Symposium*, 2020.
- [29] L. Lyu, H. Yu, J. Zhao, Q. Yang, Threats to federated learning, in: *Federated Learning: Privacy and Incentive*, Springer, 2020, pp. 3–16.
- [30] E. Sothiawat, L. Zhen, Z. Li, C. Zhang, Partially encrypted multi-party computation for federated learning, in: *2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing, CCGrid, IEEE*, 2021, pp. 828–835.
- [31] G. Peralta, R.G. Cid-Fuentes, J. Bilbao, P.M. Crespo, Homomorphic encryption and network coding in IoT architectures: Advantages and future challenges, *Electronics* 8 (8) (2019) 827.
- [32] A. Acar, H. Aksu, A.S. Uluagac, M. Conti, A survey on homomorphic encryption schemes: Theory and implementation, *ACM Comput. Surv. (CSUR)* 51 (4) (2018) 1–35.
- [33] F. Calabrese, L. Ferrari, V.D. Blondel, Urban sensing using mobile phone network data: a survey of research, *ACM Comput. Surv. (CSUR)* 47 (2) (2014) 1–20.
- [34] A. Khan, S.K.A. Imon, S.K. Das, A novel localization and coverage framework for real-time participatory urban monitoring, *Pervasive Mob. Comput.* 23 (2015) 122–138.
- [35] N. Petrushevsky, M. Manzoni, A. Monti-Guarnieri, Fast urban land cover mapping exploiting sentinel-1 and sentinel-2 data, *Remote Sens.* 14 (1) (2021) 36.
- [36] Y. Zhang, C.P. Chen, Secure heterogeneous data deduplication via fog-assisted mobile crowdsensing in 5G-enabled IIoT, *IEEE Trans. Ind. Inform.* 18 (4) (2021) 2849–2857.
- [37] J. Tang, S. Fu, X. Liu, Y. Luo, M. Xu, Achieving privacy-preserving and lightweight truth discovery in mobile crowdsensing, *IEEE Trans. Knowl. Data Eng.* 34 (11) (2021) 5140–5153.
- [38] S. Jiang, J. Liu, M. Duan, L. Wang, Y. Fang, Secure and privacy-preserving report de-duplication in the fog-based vehicular crowdsensing system, in: *2018 IEEE Global Communications Conference, GLOBECOM, IEEE*, 2018, pp. 1–6.
- [39] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD '96, AAAI Press*, 1996, pp. 226–231.
- [40] R. Angmo, N. Aggarwal, V. Mangat, A. Lal, S. Kaur, An improved clustering approach for identifying significant locations from spatio-temporal data, *Wirel. Pers. Commun.* 121 (1) (2021) 985–1009.
- [41] A.M. Aryal, S. Wang, Discovery of patterns in spatio-temporal data using clustering techniques, in: *2017 2nd International Conference on Image, Vision and Computing, ICIVC, IEEE*, 2017, pp. 990–995.
- [42] S. Gambs, M.-O. Killijian, M.N. del Prado Cortez, De-anonymization attack on geolocated data, *J. Comput. System Sci.* 80 (8) (2014) 1597–1614.
- [43] R. Wang, M. Zhang, D. Feng, Y. Fu, Z. Chen, A de-anonymization attack on geo-located data considering spatio-temporal influences, in: *International Conference on Information and Communications Security, Springer*, 2015, pp. 478–484.
- [44] H.B. McMahan, D. Ramage, K. Talwar, L. Zhang, Learning differentially private recurrent language models, in: *International Conference on Learning Representations*, 2018, URL <https://openreview.net/forum?id=BJ0hFIZ0b>.
- [45] H.B. McMahan, E. Moore, D. Ramage, B.A. y Arcas, Federated learning of deep networks using model averaging, 2016, p. 2, arXiv preprint arXiv:1602.05629.
- [46] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H.B. McMahan, S. Patel, D. Ramage, A. Segal, K. Seth, Practical secure aggregation for privacy-preserving machine learning, in: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1175–1191.
- [47] K. Wei, J. Li, M. Ding, C. Ma, H.H. Yang, F. Farokhi, S. Jin, T.Q. Quek, H.V. Poor, Federated learning with differential privacy: Algorithms and performance analysis, *IEEE Trans. Inf. Forensics Secur.* 15 (2020) 3454–3469.
- [48] Y. Zhao, J. Zhao, M. Yang, T. Wang, N. Wang, L. Lyu, D. Niyato, K.-Y. Lam, Local differential privacy-based federated learning for internet of things, *IEEE Internet Things J.* 8 (11) (2020) 8836–8853.
- [49] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang, Y. Zhou, A hybrid approach to privacy-preserving federated learning, in: *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, 2019, pp. 1–11.
- [50] H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, H. Möllering, T.D. Nguyen, P. Rieger, A.-R. Sadeghi, T. Schneider, H. Yalame, et al., SAFElearn: Secure aggregation for private FEderated learning, in: *2021 IEEE Security and Privacy Workshops, SPW, IEEE*, 2021, pp. 56–62.
- [51] B. Choi, J.-y. Sohn, D.-J. Han, J. Moon, Communication-computation efficient secure aggregation for federated learning, 2020, arXiv preprint arXiv:2012.05433.
- [52] J.H. Bell, K.A. Bonawitz, A. Gascón, T. Lepoint, M. Raykova, Secure single-server aggregation with (poly) logarithmic overhead, in: *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020, pp. 1253–1269.
- [53] S. Kadhe, N. Rajaraman, O.O. Koyluoglu, K. Ramchandran, Fastsecagg: Scalable secure aggregation for privacy-preserving federated learning, 2020, arXiv preprint arXiv:2009.11248.
- [54] J. So, B. Güler, A.S. Avestimehr, Turbo-aggregate: Breaking the quadratic aggregation barrier in secure federated learning, *IEEE J. Sel. Areas Inf. Theory* 2 (1) (2021) 479–489.
- [55] K. Niu, C. Peng, W. Tan, Z. Zhou, Y. Xu, Verifiable location-encrypted spatial aggregation computing for mobile crowd sensing, *Secur. Commun. Netw.* 2021 (2021).

- [56] J.H. Kang, W. Welbourne, B. Stewart, G. Borriello, Extracting places from traces of locations, *ACM SIGMOBILE Mob. Comput. Commun. Rev.* 9 (3) (2005) 58–68.
- [57] S. Ji, Y. Zheng, T. Li, Urban sensing based on human mobility, in: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 1040–1051.
- [58] R.M. Gray, *Entropy and Information Theory*, Springer Science & Business Media, 2011.
- [59] Similarity measures and generalized divergences, in: *Nonnegative Matrix and Tensor Factorizations*, John Wiley & Sons, Ltd, 2009, pp. 81–129, <http://dx.doi.org/10.1002/9780470747278.ch2>, Ch. 2. arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470747278.ch2>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470747278.ch2>.
- [60] C. Bettstetter, G. Resta, P. Santi, The node distribution of the random waypoint mobility model for wireless ad hoc networks, *IEEE Trans. Mob. Comput.* 2 (3) (2003) 257–269.
- [61] D. Brockmann, L. Hufnagel, T. Geisel, The scaling laws of human travel, *Nature* 439 (7075) (2006) 462–465.
- [62] I. Rhee, M. Shin, S. Hong, K. Lee, S.J. Kim, S. Chong, On the levy-walk nature of human mobility, *IEEE/ACM Trans. Netw.* 19 (3) (2011) 630–643.
- [63] B. Jiang, J. Yin, S. Zhao, Characterizing the human mobility pattern in a large street network, *Phys. Rev. E* 80 (2) (2009) 021136.
- [64] X.-Y. Yan, X.-P. Han, B.-H. Wang, T. Zhou, Diversity of individual mobility patterns and emergence of aggregated scaling laws, *Sci. Rep.* 3 (1) (2013) 2678.
- [65] X. Liang, J. Zhao, L. Dong, K. Xu, Unraveling the origin of exponential law in intra-urban human mobility, *Sci. Rep.* 3 (1) (2013) 2983.
- [66] X. Liang, X. Zheng, W. Lv, T. Zhu, K. Xu, The scaling of human mobility by taxis is exponential, *Physica A* 391 (5) (2012) 2135–2144.
- [67] W.-S. Jung, F. Wang, H.E. Stanley, Gravity model in the Korean highway, *Europhys. Lett.* 81 (4) (2008) 48005.
- [68] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M.L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, et al., Semantic trajectories modeling and analysis, *ACM Comput. Surv.* 45 (4) (2013) 1–32.
- [69] S. Spaccapietra, C. Parent, Adding meaning to your steps (keynote paper), in: *Conceptual Modeling–ER 2011: 30th International Conference, ER 2011, Brussels, Belgium, October 31–November 3, 2011. Proceedings 30*, Springer, 2011, pp. 13–31.
- [70] H. Nagesh, S. Goil, A. Choudhary, Adaptive grids for clustering massive data sets, in: *Proceedings of the 2001 SIAM International Conference on Data Mining*, SIAM, 2001, pp. 1–17.
- [71] L.O. Alvares, V. Bogorny, B. Kuijpers, J.A.F. de Macedo, B. Moelans, A. Vaisman, A model for enriching trajectories with semantic geographical information, in: *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems*, 2007, pp. 1–8.
- [72] S. Guha, R. Rastogi, K. Shim, CURE: An efficient clustering algorithm for large databases, *ACM Sigmod Rec.* 27 (2) (1998) 73–84.
- [73] A. Ram, S. Jalal, A.S. Jalal, M. Kumar, A density based algorithm for discovering density varied clusters in large spatial databases, *Int. J. Comput. Appl.* 3 (6) (2010) 1–4.
- [74] Y. Zheng, L. Zhang, X. Xie, W.-Y. Ma, Mining interesting locations and travel sequences from GPS trajectories, in: *Proceedings of the 18th International Conference on World Wide Web*, 2009, pp. 791–800.
- [75] Y. Zheng, Q. Li, Y. Chen, X. Xie, W.-Y. Ma, Understanding mobility based on GPS data, in: *Proceedings of the 10th International Conference on Ubiquitous Computing*, 2008, pp. 312–321.
- [76] Y. Zheng, X. Xie, W.-Y. Ma, et al., GeoLife: A collaborative social networking service among user, location and trajectory, *IEEE Data Eng. Bull.* 33 (2) (2010) 32–39.
- [77] K. Drakonakis, P. Ilia, S. Ioannidis, J. Polakis, Please forget where I was last summer: The privacy risks of public location (meta) data, 2019, arXiv preprint [arXiv:1901.00897](https://arxiv.org/abs/1901.00897).
- [78] B. Hoh, M. Gruteser, H. Xiong, A. Alrabady, Enhancing security and privacy in traffic-monitoring systems, *IEEE Pervasive Comput.* 5 (4) (2006) 38–46.
- [79] M.C. Thrun, M.C. Thrun, Approaches to cluster analysis, in: *Projection-Based Clustering Through Self-Organization and Swarm Intelligence: Combining Cluster Analysis with the Visualization of High-Dimensional Data*, Springer, 2018, pp. 21–31.
- [80] M. Mughnyanti, S. Efendi, M. Zarlis, Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation, in: *IOP Conference Series: Materials Science and Engineering*, vol. 725, IOP Publishing, 2020, 012128.