Differentially Private Geospatial streams publish for VANETs

Zhe Ding1,2*, Binyong Li1, Jun Zhao1, Lei Shi1
1School of Cybersecurity, Chengdu University of Information Technology, Xuefu Road, Chengdu 610225, China
2School of Information and software engineering, University of Electronic Science and Technology of China, Jianshe Road, Chengdu, 611731, China

*dingzhe@cuit.edu.cn

Abstract. With the development of vehicular Ad hoc Networks (VANETs) technology, application in VANETs has becomes an important role in human life. VANETs are the result of IoT (Internet of Things) development. There are two kinds of communication nodes in VANETs, as shown in Fig1. The first kind node is Roadside Units (RSU) which is fixed on both sides of the road as a transit point for communication. The second kind node is vehicle’s on-board unit (OBU) which is fixed in vehicle and utilizes RSU to communicate with servers in VANETs. Servers in VANETs can provide services for a user by analyzing location information uploaded by the user. Location information of users contains sensitive information. As a consequence, it has become the hot spot in academe that how to protect sensitive information in location information cannot be leaked[1].

Traditional privacy protection for VANETs technology is based on data anonymization such as k-anonymity, l-diversity and so on [2]. But study found that attacker with stronger background information can steal sensitive information by analyzing data proceed by traditional privacy schemes [3]. As a consequence, we make used of differential privacy technology to guarantee security of sensitive information in this paper.

Keywords: One-sided differential privacy, VANETs, Geospatial stream, Time windows, Roadside Units

1. Introduction

With development of vehicular Ad hoc Networks (VANETs) technology, application in VANETs has becomes an important role in human life. VANETs are the result of IoT (Internet of Things) development. There are two kinds of communication nodes in VANETs, as shown in Fig1. The first kind node is Roadside Units (RSU) which is fixed on both sides of the road as a transit point for communication. The second kind node is vehicle’s on-board unit (OBU) which is fixed in vehicle and utilizes RSU to communicate with servers in VANETs. Servers in VANETs can provide services for a user by analyzing location information uploaded by the user. Location information of users contains sensitive information. As a consequence, it has become the hot spot in academe that how to protect sensitive information in location information cannot be leaked[1].

Traditional privacy protection for VANETs technology is based on data anonymization such as k-anonymity, l-diversity and so on [2]. But study found that attacker with stronger background information can steal sensitive information by analyzing data proceed by traditional privacy schemes [3]. As a consequence, we make used of differential privacy technology to guarantee security of sensitive information in this paper.
The technology based on differential privacy has been used in many fields [4]. For example, apple has applied technology based on differential privacy to provide services and guarantee security of user privacy [5]. Differential privacy injects noise to original data in order that attacker cannot steals sensitive information in original data [3-4].

![Fig. 1 communication system model of VANENTs](image)

In practice, the location information of not all vehicles in VANENTs belongs to sensitive information for users. The traditional differential privacy technology assumes that all the location information contains sensitive information for users. As a consequence, although the traditional differential privacy can guarantee security of published result, utility of published result has been decreased in VANENTs [6]. So one-sided differential privacy is applied to guarantee security of published result in this paper.

In this paper, based on one-sided differential privacy, we proposed a novel algorithm to publish geospatial stream for VANENTs, denoted as PGSV. This novel algorithm can publish numbers of vehicles in different RSU areas and guarantee security of sensitive information in published result.

2. Related work
The concept of differential privacy was proposed by Dwork in 2006 [3]. Newsday most algorithms of publishing static geospatial data makes used of differential privacy as theoretical support of privacy protection. Those algorithms have applied or improved differential privacy mechanism to achieve privacy protection. The designing scheme of those algorithms has derived from space division perspectives and optimized data structure to increase utility of published result [7]. Cormode introduced nonuniform noise to differential privacy mechanism to divide space. During dividing space, Cormode applied a tree structure to process space data and preset height of the tree structure. In low level of the tree structure, Quad-tree is applied to divide space. And in high level of the tree, KD-tree is applied to divide space [8]. The efficiency of this algorithm is better than the method which only apply KD-tree to divide space. Based on spatial distribution, Qardaji proposed to uniform division method based fitness and then introduced this division method to publishing high dimensional data with differential privacy [9].

There are two kinds of algorithms of publishing stream, which is based on differential privacy. The first kind algorithms are publishing limited stream. The second kind algorithms are publishing unlimited stream. For publishing unlimited stream, the aim of privacy protection made data in the stream undistinguishable. Mir make used of overview of data in unlimited stream to protect privacy in the stream [10]. Bolot proposed the concept of privacy attenuation that the important of privacy is decreasing with time lapses. And then Bolot introduced privacy attenuation to publish unlimited stream [11]. For publishing limited stream, Kellaris proposed a concept of $\omega-\text{privacy}$. Based on differential privacy, Kellaris make used of this concept to transform problem of publishing unlimited stream to the problem of publishing limited stream [12].

3. Preliminaries
In this section, the foundational concept used in this paper is introduced.

Definition 1 (Neighbor datasets) [3]. Let $D$ and $D'$ be the two datasets. If the difference between $D$ and $D'$ is only one record, $D$ and $D'$ are two neighbor datasets.
Definition 2 (Differential Privacy) [3] Let $A$ is a randomized algorithm. $D$ and $D'$ are two neighbor datasets. $O$ is output domain of $A$. If inequation $\Pr[A(D) \in O] \leq e^\varepsilon \Pr[A(D') \in O]$ is satisfied, $A$ satisfies $\varepsilon$–differential privacy.

Definition 3 (Global sensitivity) [3]. Let $q$ is a query function $q:D \rightarrow \mathbb{R}^d$. For any two neighbor datasets $D$ and $D'$, the global sensitivity of $q$, denoted as $\Delta q$, is defined as $\Delta q = \max_{B,D} \|q(D) - q(D')\|_1$

Theorem 1 (Exponential Mechanism) [15] Let $u(D,o)$ is a utility function for dataset $D$ and output $o$. If the possibility of $o$ outputted by randomized mechanism $A$ satisfies $\Pr(o) \propto \exp\left(-\frac{m(u(D,o))}{2\Delta u}\right)$, $A$ satisfies $\varepsilon$–differential privacy.

Definition 3 (Policy Function) [6] A policy function $P:T \rightarrow \{0,1\}$. When location record $r \in T$ belong to sensitive information, $P(r) = 0$. When location record $r \in T$ not belong to sensitive information, $P(r) = 1$.

Definition 4 (stream prefix) [13] Let geospatial data stream $F$ be a sequence of geospatial datasets $(D_1,D_2,\cdots)$. $F[i] = D_i$. The stream prefix at time stamp $t$, denoted as $F_t$, is defined as $F_t = (D_1,D_2,\cdots,D_t)$.

Definition 5 (ω– One-sided $P$-neighbor) Let $F_t$ and $F'_t$ be two prefixes and $P$ is a policy function. If at least one condition is satisfied in following two conditions, $F_t$ is $\omega$– One-sided $P$-neighbor of $F'_t$, that $F'_t \in N_P(F_t)$ . (1) there is only one timestamp $i \leq t, P(F_t[i]) = 0$ and $F_t[i] \neq F'_t[i]$ or (2) there are two timestamps $i, i'$ with $i < i'$ and $i - i + 1 \leq \omega$ , $P(F_t[i]) = P(F_t[i']) = 0$ $F_t[i] \neq F'_t[i]$ and $F_t[i'] \neq F'_t[i']$.

Definition 6 ( $(P,\omega,\varepsilon)$ – One-Sided differential privacy) Let $A$ is a randomized mechanism. The input of $A$ is a stream prefix with time windows size $\omega$. And $O$ is the set of all possible output of $A$. If for arbitrary timestamp $t$ in the time window, formula $\Pr[A(F_t) \in O] \leq e^\varepsilon \Pr[A(F'_t) \in O]$ is held, $A$ satisfied $\omega$– One-sided $P$-neighbor $F_t$ and $F'_t$. If $A$ can compose of $A_1,A_2,\cdots,A_m$ and $A_i$ guarantee independent $(P,\omega_i,\varepsilon_i)$– One-Sided differential privacy, A guarantee $(P,\omega,\varepsilon)$ – One-Sided differential privacy $\Sigma_{i=1}^{t} \varepsilon_i \leq \varepsilon$.

Theorem 2 [6] Any $\varepsilon$– differential privacy mechanism also guarantees $(P,\omega,\varepsilon)$ – One-Sided differential privacy.

Definition 7 (Laplace Mechanism) [3] For dataset $D$, $f$ is a function $f:D \rightarrow \mathbb{R}^d$. $K$ is a randomized mechanism. If equation $K(D) = f(D) + Lap(\Delta f/\varepsilon)$ is satisfied, randomized mechanism $K$ satisfies $\varepsilon$– differential privacy.

Definition 8 (One-Sided differential privacy Laplace, denoted as OSDPLAPACE) [6] Let $D$ be a database, $D_{\omega} = \{r \in D | P(r) = 1\}$ is a subset of $D$. All the elements in $D_{\omega}$ do not belong to sensitive information. OSDPLAPACE can obtain a result of histogram query with b bins based on $D_{\omega}$ by adding b dimensional vector. Each element in the vector is drawn from $Lap_-(\lambda)$. For $\lambda = \frac{1}{\varepsilon}$ OSDPLAPACE satisfies $(P,\varepsilon)$ – One-Sided differential privacy. $f_{Lap_-(x,\lambda)}$ is pdf of $Lap_-(\lambda)$.

$$f_{Lap_-(x,\lambda)} = \begin{cases} \frac{1}{\lambda} \exp \left( \frac{x}{\lambda} \right) & x \leq 0 \\ 0 & otherwise \end{cases}$$

4. Problem model and Proposed method

In this paper, a novel algorithm is proposed to publish geospatial stream for VANETs, denoted as PGSV. PGSV is based on one-side differential privacy and time windows. And PGSV can publish numbers of vehicles outfitting OBU in scope of all the RSU and guarantee security of sensitive information in published result.

Two level partition strategy is applied to divide two-dimension geographic space. The allocation of privacy budget for PGSV refers to [13] and the detail as follows. The scopes of each RSU are not overlapped with each other. According to parallel property of differential privacy [3], although PGSV
allocates privacy budget ε to scope of each RSU, the total privacy budget of PGSV is also equal to ε. For scope of each RSU, privacy budget ε is divided to two part, that ε_1 and ε_2 (ε_1 + ε_2 = ε). ε_1 is used to count the number of vehicles in the scope and equally distributed among time windows whose size is equal to ω. ε_2 is used to divide the second level based the scope of RSU and count the number of vehicles outfitting OBU in the second level. ε_2 is divided to 3ε_2/4 and 3ε_2/4. 3ε_2/4 is allocated to protect similarity estimation. And 3ε_2/4 is allocated to count the number of vehicles outfitting OBU is the second level. The detail of PGSV is shown as follow.

(1) Two level partition strategy is applied to divide two-dimension geographic space.

For VANENTs, a lot of RSU are fixed in both sides of the road. And each RSU is corresponding to a scope. And the RSU can communicate with vehicles outfitting OBU in the scope. For two level partition strategy, the first level partition is that geographic space is divided by the scopes of RSU. In the time stamp i in time window, based on the first level partition, the second level partition is that the granularity of each scope of RSU is calculated by \( \frac{N_i \cdot \epsilon}{c} \) (1 ≤ i ≤ ω), where N_i is total count of vehicles outfitting OBU in time stamp i and in the scope of RSU. ε is total privacy budget. c is a constant. ω is the size of time window. PGSV is based on the time windows to publish geospatial stream. As a result, in each time stamp of time window, the granularity in the second level partition maybe different. For instance, there are two RSU in geospatial space. According to two level partition strategy, geospatial space is divided to two partition, as is shown in the first level of Fig 2. According to the number of vehicles outfitting OBU in each partition, each partition is divided in three subspaces, as is shown in the second level of Fig2.

![Diagram](https://via.placeholder.com/150)

**Fig. 2 Example of two-level partition strategy**

(2) calculating separately the number of vehicles outfitting OBU in a scope of RSU and in each subspace of the second level in the scope and applying One-side differential privacy to guarantee security of the result.

Let N is the number of vehicles outfitting OBU in scope of RSU or a subspace of the second level in the scope. \( N_{ns} \) is the number of vehicles outfitting OBS whose information not belong to sensitive information in this scope or the subspace of this scope. And f is a query about the number of vehicles outfitting OBU in the scope. According to the concept of global sensitivity, global sensitivity of f is equal to 1. The noise count of vehicles outfitting OBU in the scope is shown as Formula (2).

\[
f(N) = (N + z) \cdot \left( 1 - \frac{N_{ns}}{N} \right) + \left( N_{ns} + z \right)
\]  

(2)
\[ z \sim \text{Lap} \left( \frac{1}{\eta} \right) \quad z' \sim \text{Lap}_- \left( \frac{1}{\eta} \right) \]  

(3)

\[ z \] is a random variable drawn from Laplace distribution whose parameter is equal to \( \frac{1}{\eta} \). And \( z' \) is also random variable drawn from \( \text{Lap}_- \left( \frac{1}{\eta} \right) \).

When PSGV makes used of Formula (2) to obtain noise count of scope of each RSU, \( \eta \) is equal to \( \frac{\varepsilon_1}{\omega} \). When PSGV applies Formula (2) to obtain noise count of subspace of scope of each RSU, \( \eta \) is equal to \( \frac{3\varepsilon_2}{4\omega} \). Privacy budget of PSGV is equally distributed among time window whose size is equal to \( \omega \). As a result, privacy budget of obtaining noise count must divides \( \omega \).

(3) Dynamic changing of the granularity of the second level

PSGV can publish geospatial stream for VANETs. As a consequence, during time window, the number of vehicles outfitting OBU is different among each time stamp. As a result, according to the changing of the number of the vehicles, the granularity of scope of each RSU has been changed. Let \( N(i) = \{N_1, N_2, \cdots, N_k\} (1 \leq k \leq \omega) \) is the set of total count of vehicles with outfitting OBU from time stamp 1 to time stamp \( k \) in a scope of the \( i \)th RSU. And \( N^i(i) = \{N^i_1, N^i_2, \cdots, N^i_m\} \) is the total count of subspace of the scope in time stamp \( j \). \( N_{i+k+1} \) is total count of vehicles with outfitting OBU in the scope in time stamp \( k+1 \). The detail of changing is shown as follows:

1) If \( \exists \forall N^i_j, N^i_{j+1}, \cdots, N^i_{j+k} \in N, N_j = N_{j+1} = \cdots = N_{j+k} = N_{i+1} \), PSGV make used Exponential Mechanism to guarantee security of total count of vehicles outfitting OBU in the scope. According to Exponential Mechanism, Pearson Correlation is applied to measure similarity between \( N_{j+i} (1 \leq l \leq k) \) and \( N_{i+1} \), as shown in Equation (4).

\[
\cos (N_{j+i}, N_{i+1}) = \frac{\sum_{n=1}^{m} (N^i_n - \bar{N}) (N_{i+1} - \bar{N})}{\sqrt{\left( \sum_{n=1}^{m} (N^i_n - \bar{N})^2 \right) \left( \sum_{n=1}^{m} (N_{i+1} - \bar{N})^2 \right)}} \tag{4}
\]

\[
\bar{N} = \frac{1}{k} \sum_{d=1}^{k} N^i_d \tag{5}
\]

\[
\bar{N} = \frac{1}{k} \sum_{d=1}^{k} N^i_d \tag{6}
\]

Pearson Correlation is less than 1. As a result, the global sensitivity of Pearson Correlation is equal to 1. As a consequence, when \( N_j, N_{j+1}, \cdots, N_{j+k} \in N, N_j = N_{j+1} = \cdots = N_{j+k} = N_{i+1} \), we select the noise count of an element from the set of \( \{N^i_j, N^i_{j+1}, \cdots, N^i_{j+k}\} \) to replace the noise count of \( N_{i+1} \) with probability proportional \( \exp \left( -\frac{\varepsilon \cos (N_{j+i}, N_{i+1})}{2} \right) \) to be published.

2) If \( \exists \forall N^i_j, N^i_{j+1}, \cdots, N^i_{j+k} \in N, N_j = N_{j+1} = \cdots = N_{j+k} = N_{i+1} \) is not satisfied, PSGV make used of Equation (2) to obtain noise count of \( N_{i+1} \) and publish.

5. Analyzing privacy

PSGV firstly applies Laplace Mechanism to obtain the total noise count of vehicles outfitting OBU in a scope of RSU and applies OSDDPLACE to obtain the total noise count of vehicles outfitting OBU which belong to non-sensitive information. According to those noise counts, PSGV has obtained final noise count for a scope of RSU which can be used to publish. According to Theorem 2, the noise count satisfies \( (P, \alpha, \epsilon) \) – One-Sided differential privacy. For subspaces in the second level, PSGV make used of Exponential Mechanism to guarantee security of count of vehicles and satisfies \( (P, \alpha, \epsilon) \) – One-Sided differential privacy. As a result, PSGV is satisfies \( (P, \alpha, \epsilon) \) – One-Sided differential privacy.
6. Performance evaluation
In order to verify efficiency of PSGV, we apply two public datasets to evaluate performance of PSGV. The first dataset is T-drive dataset which record 10000 taxis in Beijing from February 2th to February 8th in 2008[14]. And the second dataset is T-Rome dataset which record trajectories of 300 taxis in Rome [15]. And we random select 1/5 taxis in both datasets as sensitive information [6]. Algorithms proposed in [7] and [16] are applied as contrast algorithms, denoted as KM and DGP respectively. We make used of four query types to evaluate efficiency of PSGV and Mean Relative Error, denoted as MRE, is used to evaluate experiment results. The detail of those four query types is described in [7]. And the result is shown in Fig3.

![Fig. 3 The result of Experiment](image)

There are two experiments to verify efficiency of PSGV for two datasets. In the first experiment, privacy budget is equal to 1. The result of the experiment is shown in sub-graph (b) and (c) of Fig3. And in the second experiment, privacy budget is equal to 0.5. The result is shown in sub-graph (a) and (d) of Fig 3. The results of all experiments show that MRE of PSGV is better than the other algorithms. The reason is that the granularity is changing with the changing of the number count of vehicles outfitting OBU. And the granularity can have decreasing the influence of noise to total count. And PSGV make used of One-Sided differential privacy to guarantee security of the count. And the noise inputted for One-Sided differential privacy is less than for differential privacy. So, the efficiency is better than the other algorithms.

7. Conclusion
In this paper, a novel algorithm is proposed to publish geospatial stream, denoted as PGSV. PGSV make used of one-sided differential privacy and traditional differential privacy to guarantee security of sensitive information. PGSV firstly applies two level partition strategies to divide two-dimension geographic space. Based on partition strategy, PGSV can obtain the noise count in the subspace. And the noise count has strong security and high utility for user’s privacy. Two public datasets are applied to verify the efficiency of PGSV. The result shows that MRE of PGSV is better than the other contrast algorithms.

Acknowledgments
This work was supported by Sichuan Science and Technology Program (No.2019YFG0425); The open Project of Network and Data security Key Laboratory of Sichuan Province (No. NDS2018-1); the National Natural Science Foundation of China (Grant No. 71701026)

References

[1] Zhaojun Lu, Gang Qu and zhenglin Liu 2018 A survey on recent advances in vehicular network security, trust, and privacy J. IEEE Transactions on Intelligent Transportation Systems vol 20 pp 760-776

[2] Zhu H, Lu R and Shen X 2009 Security in service-oriented vehicular networks J. IEEE Wireless Communications vol 16 pp 16-22

[3] Dwork C 2006 Differential Privacy J. lecture notes in computer science vol 2 pp 1-12

[4] Zhen H, Chiou B C, Tsou Y T, Kuo S Y and Wang P C 2020 Association Rule Mining with Differential Privacy C. 50th Annual IEEE/IFIP Int. Conf. on Dependable Systems and Networks Workshops (Valencia) pp 47-54

[5] Apple 2017 Learning with privacy at scale Machine learning Journal vol 1 pp 1-25

[6] Kotsogiannis I, Doudalis S and Haney S 2020 One-sided Differential Privacy C. IEEE 36th Int. Conf. on Data Engineering (Dallas) pp 493-504

[7] Yiwen N, Liusheng H and Zongfeng L 2016 Geospatial streams publish with differential privacy C. Int. Conf. on Collaborative Computing, Networking Applications and Worksharing (Washington) pp 152-164

[8] Cormode G, Procopiuc C and Srivastava D 2012 Differentially private spatial decompositions C. The 28th IEEE Int Conf on Data Engineering (Washington) pp 20-31

[9] Qardajiw W and Li N 2012 Recursive partitioning and summarization: a practical framework for differentially private data publishing C. Proc. of the 7th ACM Symp. to Information, Computer and Communication Security (New York) pp 38-49

[10] Mir D, Muthukrishnan S and Nikolov A 2011 Pan-private algorithms via statistics on sketches C. Proc. of the thirtieth ACM SIGMOD-SIGACT-SICART symp. on Principles of database system (Washington) pp 37-48

[11] Bo lot J, Jawaz N and Muthukrishnan S 2013 Private decayed predicate sums on streams C. Int. Conf. on Display Technology (San Francisco) pp 284-295

[12] Kellaris G, Papadopoulos S and Xiao X 2014 Differentially Private event sequences over infinite streams C.The 40th Int. Conf. on Very Large Data Bases (Washington) vol 7 pp 1155-1166

[13] Yiwen N, Liusheng H and Zongfeng L 2016 Geospatial Streams Publish with Differential Privacy C.International Conference on Collaborative Computing, Networking (Washington) pp 152-164

[14] Amici R, Bonola M and Bracciale L 2014 Performance Assessment of an Epidemic Protocol in VANET Using Real Traces J. Procedia Computer Science vol 40 pp 92-99

[15] Yuan N J, zheng y and Zhang C. 2010 T-drive: driving directions based on taxi trajectories C. ACM SIGSPATIAL GIS 2010 ACM SIGSPATIAL Intl conf on advances in geographic information systems (San Jose) pp 99-108

[16] Jun W, Shubo L and Yongkai L I 2016 Differentially Private Spatial Decompositions for Geospatial Point Data J. China Communications vol 13 pp 97-107