Measuring location privacy in V2X communication systems with accumulated information

Zhendong Ma, Frank Kargl, and Michael Weber
Institute of Media Informatics, Ulm University, Germany
{zhendong.ma|frank.kargl|michael.weber}@uni-ulm.de

***Abstract*—Vehicle-to-vehicle/vehicle-to-infrastructure (V2X) communication systems are envisioned to greatly improve road safety, traffic efficiency, and driver convenience. However, many V2X applications rely on continuous and detailed location information, which raises location privacy concerns. A multitude of privacy-protection mechanisms have been proposed in recent years. However, few efforts have been made to develop privacy metrics, which can provide a rigorous way to assess the privacy risk, evaluate the effectiveness of a given mechanism, and exploit the full possibilities of protection methods in V2X systems. Therefore, in this paper we present a *trip-based location privacy metric* for measuring user location privacy in V2X systems. The most distinguishable aspect of the metric is to take into account the *accumulated information*, which is the privacy-related information acquired by an adversary for an extended period of time, e.g., days or weeks. We develop methods to model and process the accumulated information, and reflect the impact on the privacy level in the metric. We further prove the viability and correctness of the metric by various case studies. Our simulations find out that under certain conditions, accumulated information can significantly decrease the level of user location privacy. The metric and our findings in this paper give some valuable insights into location privacy, which can contribute to the development of effective privacy-protection mechanisms for the users of V2X systems.**

I. Introduction

The emerging vehicle-to-vehicle/vehicle-to-infrastructure (V2X) communication systems enable a new way of cooperation among vehicles, traffic operators, and service providers. Based on Dedicated Short Range Communications (DSRC) technology, vehicles can communicate among each others and with the entities in the back-end system via Roadside Units (RSU). It is envisioned that V2X communication systems can significantly improve road safety, traffic efficiency, and driver convenience. Example V2X applications include collision warning, floating car data, and location-based services. If deployed, such systems will be one of the biggest realizations of Mobile Ad Hoc Networks (MANET).

However, many V2X applications rely on continuous and detailed location information of the vehicles. Vehicles are personal devices. Locations of a vehicle reveals the movements and activities of its driver and passengers. Sending and disseminating location information of the users of V2X systems has the potential to infringe the users' location privacy. The location privacy issue in V2X communication systems has been identified and a multitude of privacy-protection mechanisms have been proposed in recent years, e.g., in [1]–[4].

To evaluate the effectiveness of these mechanisms, a metric for measuring the level of user location privacy is crucial and indispensable. For example, we need a metric which can tell us that the user privacy level has been increased by 20% after applying one of the protection mechanisms. However, so far the main focus on the topic is to devise privacy-protection mechanisms, very few metrics exist for measuring user location privacy in V2X systems in a rigorous way. Hence, the usefulness of privacy-protection mechanisms cannot be strictly evaluated and compared and the trustworthiness of V2X systems cannot be assessed. Furthermore, the range of possible protection methods cannot be fully exploited.

In our previous work [5], we introduced a *trip-based location privacy metric* to measure the level of location privacy of individual users in V2X systems. Based on the observation that the uncertainty of a potential adversary and the user privacy level are indeed two sides of the same coin, the metric measures the level of location privacy as the linkability of location information to the individuals who generate it. The uncertainty in the information is quantified into entropy. Our previous work assumes that the information available to the adversary is limited to a short period of time. To stay realistic, it is reasonable to assume that an adversary will do its best to decrease the uncertainty of the obtained information. Therefore, the adversary is likely to take the maximum available information into account. In particular, the adversary will try to utilize the accumulated information, which is privacy-related information acquired by capturing communications from running V2X systems for an extended period of time, e.g., days or weeks. Hence, the

978-1-4244-5113-5/09/$25.00 ©2009 IEEE

assumption of a limited time period is over-simplified from the real world.

To reflect the true underlying privacy value in V2X communication systems, the metric must take into account the impact of accumulated information on privacy level. Intuitively, the more information an adversary has, the more it can draw conclusions with less uncertainties. However, the impact of accumulated information on location privacy has not been investigated up to now. In this paper, we address this issue by extending the current location privacy metric to take into account accumulated information. As a result, the metric can more accurately reflect users' privacy value in V2X communication systems. Specifically, in this paper we

- develop a method to model the accumulated information,
- design approaches to process, propagate, and utilize the accumulated information, and reflect the effect in the metric,
- prove the viability and correctness of the metric by means of various case studies.

In the following, Section II gives the background information on the basics of the trip-based location privacy metric. Section III describes the method to model the accumulated information. Section IV introduces two approaches to process the accumulated information and reflects it in the metric. Section V evaluates the metric by case studies. Section VI discusses the related work, followed by the conclusion in Section VII.

II. Metric fundamentals

This section gives the necessary background information on the trip-based location privacy metric introduced in [5].

In V2X communication systems, each time a vehicle sends a message, it gives out its location information to the system. Although there are different levels of granularities, the location information in V2X systems can be categorized into three types, i.e., single locations, tracks, and trips. Location information only becomes privacy-relevant if it can be linked to identifiable individuals. Since for privacy concerns vehicles are very likely to use pseudonyms in communications [6], [7], information on single locations and tracks are less privacy-sensitive than the information on trips, which can be used to infer an individual's identity and activities. The first step to measure privacy is to capture the information on trips and individuals in an arbitrary defined area and time period. Hence the metric virtually takes a "snapshot" of the dynamic V2X systems.

The information captured in the snapshot is then modeled in a weighted tripartite graph, shown in Fig. 1. The graph contains three distinct sets of vertices, i.e., I, O, and D, which represent Individuals, Origins and Destinations of the trips. An adversary's knowledge on the linkabilitiy of an individual to a set of trips is expressed in probability distributions. The probabilities are used as the weights on the directed edges. For example, p_{jk} is a weight on an edge (v_j, v_k) between the vertices v_j and v_k.

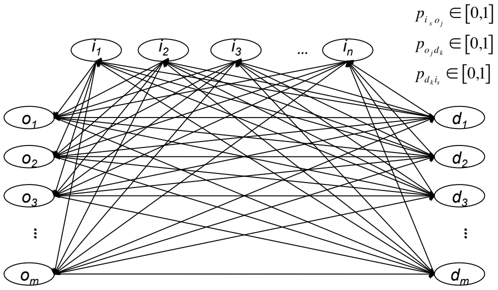

Fig. 1. Snapshot information modeled in weighted tripartite graph

For an individual to make a trip (e.g., $o_1 \longrightarrow d_1$), he or she must start from one of the origins, e.g., i_1 from o_1. If the trip from o_1 ends at one of the destinations, it must be possible to link i_1 to d_1 as well. Due to the uncertainty in the information, there can be many of such possible linkings among the vertices. A closed walk or a cycle starting from a vertex i_s and passing vertices $\{o_j, d_k\}$ in the graph has the semantics of i_s's probability p_{jk} to make a trip with origin o_j and destination d_k. By collecting all cycles connected to a particular individual in the graph, we can extract the probability distribution of the linkability of that individual to a set of trips. The probability distribution can be graphically expressed as a hub-and-spoke structure, shown in Fig. 2. The last spoke with probability p^c in the clock-wise order denotes the probability of an individual not making any trips, i.e., "staying at home".

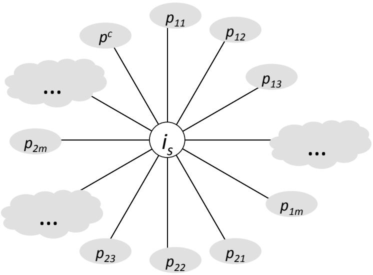

Fig. 2. Extracted probability distribution as hub-and-spoke

The normalized probabilities on each of the spokes are calculated as

$$\hat{p}_{jk} = \frac{p(i_s, o_j)p(o_j, d_k)p(d_k, i_s)}{\displaystyle\sum_{j=1}^{m}\sum_{k=1}^{m} p(i_s, o_j)p(o_j, d_k)p(d_k, i_s) \ + \ \hat{p}^c}$$

$$\hat{p}^c = 1 - \sum_{j=1}^{m} p(i_s, o_j)$$

with probabilities taken from the graph in Fig. 1. Applying Shannon's entropy [8], we quantify the uncertainty in the information about i_s in entropy as

$$H(i_s) = -(\sum_{j=1}^{m}\sum_{k=1}^{m} \hat{p}_{jk} log(\hat{p}_{jk}) \ + \ \hat{p}^c log(\hat{p}^c))$$

where the logarithm is taken to base 2 to have a unit of *bit*. $H(i_s)$ is used as a quantitative measure of i_s's level of location privacy. The privacy level is directly proportional to the value of entropy, i.e., the higher the entropy, the higher the privacy level, and vice versa. Entropy reaches its maximum if all trips are equally probable. For a snapshot with m^2 O/D pairs, the maximum entropy for each individuals in the snapshot is

$$H_{max} = log(m^2 + 1)$$

with 1 accounting for not making any trips [5].

III. Accumulated information

Using snapshots enables us to capture privacy-relevant information from V2X communication systems, which are continuous and dynamic in nature. However, privacy measurements based on a single snapshot only reflect the privacy values in a short period of time. It is reasonable to assume that a determined adversary will collect as much information as possible over a long period of time to work for its advantage. Intuitively, information accumulated over time should help to reveal more facts about the individuals and their vehicle movements.

To reflect this more realistic assumption on the adversary, instead of one snapshot, we extend the metric to include consecutive snapshots. Thus the metric yields measurements on "multiple snapshots". In a single snapshot, the information needed for measuring each individual can be represented by a hub-and-spoke structure shown in Fig. 2. When more snapshots are added to the metric, we can imagine that the information related to an individual i becomes a sequence of hub-and-spoke structures ordered in time as shown in Fig. 3. Notice that only one individual is shown in Fig. 3. But we can imagine that for each of the individuals captured in the snapshots, we can extract the information and build a similar sequence of hub-and-spoke structures. For simplicity in formulations, we will only consider *one* individual i in the rest of the paper. The same formulas and procedures are applicable to any of the other individuals captured in the snapshots. However, in our future work, we will further investigate the *interrelations* among individuals and their impacts on the level of location privacy.

There are several observable characteristics of accumulated information. First, i can be linked to different trips from snapshot to snapshot. The differences are in the number, as well as the origins and destinations of the trips. We name the assortment of trips related to i in a snapshot a *trip constellation*. Second, accumulated information has two dimensions, i.e., the one extends into the diversity of trip constellations, and the other extends along the timeline. Third, given the fact that many individuals use vehicles to fulfill demands on activities on a daily basis [9], accumulated information is likely to contain an individual's *trip patterns*, i.e., regularly occurring trips with the same origins and destinations. Therefore, by *same trip* we mean two or more trips have the same origin and destination, e.g., the same garage, parking lot, or street parking space etc.

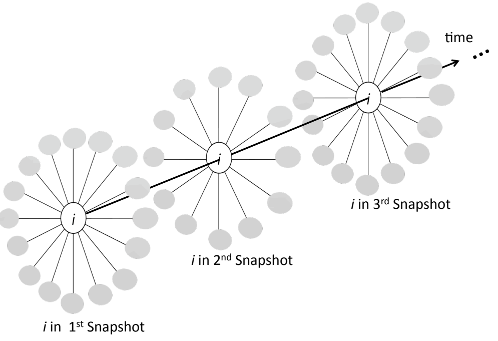

Fig. 3. Multiple snapshots of i in timely-ordered sequence

To model the accumulated information in multiple snapshots, we represent the hub-and-spoke structures in a more compact way. Let S be the set of all snapshots and let T be the set of all trips considered for an individual i, then snapshot S_t reflects the relation of i to a set of trips at the time period t. We define S_t to be

$$S_t{:=}\{(T_k,p_k)|\ T_k{\in}T,p_k{\in}]0,1],\textstyle\sum_k p_k{=}1,k{=}1,\ldots,n_t\}$$

where (T_k, p_k) is a tuple in which T_k denotes a specific trip (i.e., the k^{th} trip) and p_k is the corresponding probability of that trip. Only trips with probabilities bigger than 0 are assigned to i. As trip constellations can vary in snapshots, we denotes the number of possible trips at t by a variable n_t. For the t^{th} snapshot, each T_k represents a spoke and each p_k represents the corresponding probability on that spoke. For simplicity, the last spoke denoting the probability of an individual "staying at home" is also represented as one of the trips. As the metric uses entropy to quantify the uncertainty in the information (cf. Section II), the calculation of entropy of i at time t can be simplified as

$$H_t = -\sum_{k} p_k log(p_k) \qquad (1)$$

where p_k is the probability of the k^{th} trip in S_t.

Consider a simple example in Table I. We have five consecutive snapshots of an individual i, $t = 1, \dots, 5$. In the 1^{st} snapshot, i is probable to make one of the trips $\{T_1, T_2, T_3, T_4\}$ with corresponding probabilities given in the table. In the 2^{nd} snapshot, i is observed to make a new trip T_5. In the 4^{th} and 5^{th} snapshot, T_3 disappears from the observation. For clarity, non-existing trips (or tuples) are shown as blanks in the table. The probabilities show the adversary's information on the linkability of the vehicle trips to a particular individual over time. However, only one trip at each time (i.e., each row in the table) has actually happened.

TABLE I
A SIMPLE EXAMPLE WITH SIX CONSECUTIVE SNAPSHOTS OF i

t	T_1	T_2	T_3	T_4	T_5
$t=1$	0.2	0.2	0.3	0.3	
$t=2$	0.2	0.2	0.3	0.2	0.1
$t=3$	0.2	0.1	0.3	0.2	0.2
$t=4$	0.2	0.3		0.2	0.3
$t=5$	0.2	0.2		0.3	0.3
$t=6$	0.2	0.2	0.2	0.2	0.2

Now imagine that the 6^{th} snapshot is captured. Without considering snapshots accumulated in the past, the information contained in S_6 represents the highest uncertainty because all trips are equally probable. However, if we also take into account the five already existing snapshots, our intuition tells us that the historical data might provide us with some useful information.

Based on the observed characteristics, we are aware that to include accumulated information in the metric, we need approaches to *process* the information contained in the snapshots, *propagate* such information along the timeline to the following snapshots, and *utilize* the information in the measurement calculation.

IV. METRIC BASED ON MULTIPLE SNAPSHOTS

In this section, we propose two approaches to measure location privacy with accumulated information. Specifically, the existing trip-based location privacy metric is extended from a single snapshot to multiple timely-ordered snapshots. The extension to multiple snapshots takes into account the impact of accumulated information on location privacy.

A. Frequency based approach

One way to "learn from the past" is to check whether the same trip has already been observed. Normally vehicle trips have some patterns. For example, we might drive from home to work on a daily basis. Hence the information on the frequency of a particular trip in the past gives hints on how probable the same trip will be repeated in the future. For this we define an auxiliary variable f_k^t which counts how often trip T_k has been linked to i over all snapshots up to time t, i.e., $f_k^t = |\{S_i | S_i \in S, i = 1, ..., t, \exists (T_k, p_k) \in S_i\}|$. For example, in Table I, at time $t = 6$, T_1 has occurred 6 times so $f_1^6 = 6$, whereas $f_3^6 = 4$ holds. Then the frequency-adjusted snapshot $\hat{S}_t^f$ for snapshot $S_t = \{(T_k, p_k)|...\}$ can be calculated as

$$\hat{S}_t^f = \{(T_k, \alpha p_k f_k^t), k = 1, \dots, n_t\} \qquad (2)$$

where $\alpha = 1/\sum_k p_k f_k^t$ is a normalization constant calculated by requiring that all probabilities in $\hat{S}_t^f$ sum to 1. Consequently, the frequency-adjusted S_6 is

$$\hat{S}_6^f \approx \{(T_1,0.22),(T_2,0.22),(T_3,0.15),(T_4,0.22),(T_5,0.19)\}$$

Comparing $\hat{S}_6$ with S_6, the probability distribution changes from equal to unequal. The corresponding entropy calculated by (1) is also decreased from 2.32 for S_6 to 2.31 for $\hat{S}_6$, i.e., the accumulated information slightly reduces the uncertainty of the current information.

However, using only the frequency of a particular trip does not consider the actual probability of that trip in each snapshot. Therefore, we lose information if we use only frequencies to adjust a snapshot. For example, in Table I, though T_1 and T_4 have the same value of f_k^t, T_4 has a higher average probability than T_1. To include actual values of the probabilities in the snapshots, we rewrite (2) as

$$\hat{S}_t^w = \{(T_k, \alpha p_k w_k^t), k = 1, \dots, n_t\} \qquad (3)$$

in which we replace f_k^t by the average probability of the same trip, i.e., $w_k^t = \sum_i p_k^i / f_k^t$ for $i = 1, \dots, t$. The normalization constant α is changed to $\alpha = 1/\sum_k p_k w_k^t$, accordingly. The probability of a non-existing trip (e.g., T_5 at $t = 1$) is treated as 0, so the equation can be kept in a generic form. Using (3), $\hat{S}_6^w$ turns out to be

$$\hat{S}_6^w \approx \{(T_1,0.18),(T_2,0.18),(T_3,0.24),(T_4,0.21),(T_5,0.19)\}$$

with an entropy value of 2.31. The result again shows that accumulated information, in terms of average probabilities, can change the current probability distribution and thus modify the level of uncertainty. Furthermore, the result reflects the value of probabilities of the trips in the past. For example, T_3 has the highest probability because it has been associated with high probabilities in the past (i.e., 0.3 at $t = 1, 2, 3$). On the other hand, even though T_1 and T_2 appear at all snapshots, the relatively low probabilities in the past cause these two trips to have the lowest value in the probability distribution of $\hat{S}_6^w$ (i.e., both are 0.18). A more extensive evaluation of this approach will be given in Section V.

B. Bayesian approach

Our second approach to process, propagate, and utilize the accumulated information is to use the Bayesian method to infer information from the historical data.

In principle, Bayesian method uses evidence to update a set of hypotheses expressed numerically in probabilities. The core of Bayesian method is the Bayes' theorem. Let h_k be the k^{th} *hypothesis* of a complete set of *hypotheses* H^1[1], the Bayes' theorem can be written as a function of h_k as

$$P(h_k|E) = \frac{P(E|h_k)P(h_k)}{\sum_{k} P(E|h_k)P(h_k)} \tag{4}$$

in which E is the evidence. $P(h_k|E)$ is the *posterior probability* of h_k because it is the conditional probability of h_k given the evidence E. $P(E|h_k)$ is the conditional probability of observing the evidence E if the hypothesis h_k is true. $P(h_k)$ is the *prior probability* of h_k because it is the probability of h_k before it is updated by E. The denominator in (4) is the sum of probabilities of observing the evidence E under all possible hypotheses.

The above description accounts for updating the hypotheses once. When applying Bayes' theorem to situations in which hypotheses are continuously updated by new evidence, the following steps are usually involved:

- Initially define an *exhaustive* and *mutually exclusive* set of hypotheses H^0.
- Before receiving new evidence E, generate a set of priori hypotheses H^-. H^- is the same as H^0 before the first update.
- After receiving the evidence E, calculate the set of posterior hypotheses H^+ using (4). H^+ will be used as the prior hypotheses H^- for the next update.

In Bayesian method, the initial hypotheses can be subjective, i.e., we can assign probabilities to the hypotheses according to some preliminary knowledge. If there are enough evidence, the hypotheses will eventually be updated towards the objective truth.

The characteristics of the modeled accumulated information make it appropriate to apply Bayesian method. Specifically, S_t contains a set of possible trips and the corresponding probabilities. Each of the trips can be regarded as a hypothesis of an individual making that trip. S_t includes all the possible trips and only one of them can be true. Therefore, the hypotheses are complete and mutually exclusive. The corresponding probabilities are the evidence of those trips from observations. At each time step, S_t contains a new set of evidence, which can be used to update the hypotheses.

However, there is still an issue to be solved before we can apply Bayesian method. It is very likely that S_t contains a dynamic constellation of trips, e.g., $\{T_1, T_2, T_3, T_4\}$ in S_1 and $\{T_1, T_2, T_3, T_4, T_5\}$ in S_2 (see Table I). The implication of such dynamics is that the set of hypotheses H will be different from snapshot to snapshot. As Bayesian method works on a fixed set of hypotheses, i.e., it does not consider adding or removing one or more hypotheses during the evidence updating process, we need a "smart" solution to apply Bayesian method to solve our problem.

The solution is Algorithm 1 shown below. In general, for a given snapshot at time t the algorithm calculates the modified probability distribution for this snapshot using the Bayesian method. Specifically, for each existing snapshot $S_j, j = 1, \ldots, t$, the algorithm calculates the prior hypotheses H_j^- and uses the probability in S_j to calculate the posterior hypotheses H_j^+. The algorithm stores each H_j^+ in a belief table B. Entries in B can be regarded as *Belief*, i.e., posterior hypotheses updated by evidence which express the level of confidence of the algorithm on their "correctness". The algorithm also keeps tracks of *the latest posterior hypotheses with the same trip constellation*. For example, S_6 has the same trip constellation as S_3 in Table I, so H_3^+ will be the latest posterior hypotheses with the same trip constellation to S_6. Informally, we use $H_j^+ \equiv_{lph} S_i, i > j$ to denote that H_j^+ is the latest posterior hypotheses of S_j in B with the same trip constellation as S_i.

Algorithm 1 Calculate $\hat{S}_t$ using Bayesian method

```
Input: snapshots until time $t$, $S_1, \ldots, S_t$
Output: snapshot at time $t$ with modified probability
    distribution, $\hat{S}_t$
for $i = 1$ to $t$ do
    if found $H_j^+ \equiv_{lph} S_i$ then
        use $H_j^+$ as $H_i^-$
    else
        assign equal probabilities to $H_i^-$
    end if
    update $H_i^-$ with the probabilities in $S_i$, the result
        is $H_i^+$
    add $H_i^+$ to $B$
end for
replace the probability distribution in $S_t$ with $H_t^+$
    to obtain $\hat{S}_t$, return $\hat{S}_t$
```

To calculate $\hat{S}_t$, the algorithm takes all existing snapshots up to time t. Before processing a new snapshot S_i, the algorithm first consults B for the latest posterior hypotheses with the same trip constellation as S_i. If found, the posterior hypotheses H_j^+ will be used as the prior hypotheses H_i^- for the current snapshot S_i. If not found, the algorithm assigns H_i^- with equally distributed probabilities. The rationale is that we assign probabilities without any prejudices to the initial

[1] Notice that the notation H is conventionally used for both entropy and hypotheses. We keep the convention and assume that the meaning should be clear from the context.

hypotheses, believing that the evidence will eventually update the hypotheses towards the objective truth. Then H_i^- is updated by S_i to generate H_i^+. H_i^+ is added to B. Notice that for efficiency, B only needs to keep the latest H^+ with unique trip constellation. Finally, H_t^+ replaces the probability distribution in S_t to have $\hat{S}_t$. $\hat{S}_t$ reflects the current beliefs expressed in probabilities, which have been continuously updated by new evidence, on each of the trips in the trip constellation in S_t. In line 7 of the algorithm, when using the probabilities in S_i to update the prior hypotheses, the notions in (4) can be substituted and rewritten as

$$p_k^{H_i^+} = \frac{p_k^{S_i} p_k^B}{\sum\limits_k p_k^{S_i} p_k^B} \qquad (5)$$

in which $p_k^{H_i^+}$ and $p_k^{S_i}$ are the probabilities of the k^{th} trip in H_i^+ and S_i, respectively. p_k^B is defined as

$$p_k^B = \begin{cases} p_k^{H_j^+} & \text{if H}_j^+ \equiv_{lph} S_i \text{ found} \\ \frac{1}{n_i} & \text{if } H_j^+ \equiv_{lph} S_i \text{ not found} \end{cases} \qquad (6)$$

in which $p_k^{H_j^+}$ is the probability of the k^{th} trip of the latest posterior hypotheses in B with the same trip constellation as S_i, and n_i is the number of trips in S_i.

We demonstrate how the algorithm works by calculating the same example from Table I. The results at each time step are shown in Fig. 4. We also include H^- at each time step to show how they are assigned and how they are updated by S to generate H^+. For example, at $t = 2$, since the trip constellation of S_2 appears for the first time, H^- is assigned a equal probability distribution. Look further down, at $t = 6$, the latest snapshot with the same trips constellation can be found at $t = 3$. So the posterior probabilities H^+ at $t = 3$ is copies to the prior probabilities H^- at $t = 6$. $\hat{S}_6$ has the same value as H^+ at $t = 6$

$$\hat{S}_6 \approx \{(T_1,0.19),(T_2,0.1),(T_3,0.42),(T_4,0.19),(T_5,0.09)\}$$

with entropy of 2.08. Comparing with the results from the frequency based approach in Section IV-A, we witness a more dramatic change in the probability distribution and the decrease in entropy. We will further compare and evaluate these approaches in the next section.

t	S_t(Evidence)					B (Belief)					
	T_1	T_2	T_3	T_4	T_5		T_1	T_2	T_3	T_4	T_5
t=1	0.2	0.2	0.3	0.3		H^-	0.25	0.25	0.25	0.25	
						H^+	0.2	0.2	0.3	0.3	
t=2	0.2	0.2	0.3	0.2	0.1	H^-	0.2	0.2	0.2	0.2	0.2
						H^+	0.2	0.2	0.3	0.2	0.1
t=3	0.2	0.1	0.3	0.2	0.2	H^-	0.2	0.2	0.3	0.2	0.1
						H^+	0.19	0.1	0.42	0.19	0.1
t=4	0.2	0.3		0.2	0.3	H^-	0.25	0.25		0.25	0.25
						H^+	0.2	0.3		0.2	0.3
t=5	0.2	0.2		0.3	0.3	H^-	0.2	0.3		0.2	0.3
						H^+	0.16	0.24		0.24	0.36
t=6	0.2	0.2	0.2	0.2	0.2	H^-	0.19	0.1	0.42	0.19	0.1
						H^+	0.19	0.1	0.42	0.19	0.09

Fig. 4. Example of Algorithm 1

V. Evaluation

A. Evaluation criteria

Our goal is to evaluate whether the privacy metric can *really* reflect the underlying value of user location privacy in V2X communication systems. For this purpose, we define two use-case-based evaluation criteria. The use cases specify scenarios likely to happen in V2X systems. The criteria are the expected impacts of the scenarios on user location privacy. We simulate the use cases. The simulation results will then be compared with the criteria. The results give us clues as how good the metric can be used to measure the location privacy in V2X systems. We define the evaluation criteria as

1) if an individual has irregular trips with quite different origins and destinations at each time, accumulated information should provide less or even no additional information;
2) if an individual has regular trip patterns, accumulated information should provide additional information. With this additional information, it should be possible to detect an individual's trip patterns.

In our metric, the uncertainty of information is quantified in entropy. A decrease in entropy indicates that additional information leads to a decrease in uncertainty, i.e., a decrease in user location privacy.

B. Evaluation setup

We identify three parameters to have main influences on the outcome of the metric. Among them are the trip constellations in each snapshot, their corresponding probability distributions, and the number of snapshots. First, the trip constellation specifies the number of trips and their appearances observed in a snapshot. Second, the probability distribution of the corresponding trips specifies the information captured by a snapshot. Finally, the number of snapshots specifies the duration of the measurement. Implicitly, it specifies the amount of accumulated information available to the metric.

By specifying these parameters, we can create use cases to check whether the metric meets the evaluation criteria. The use cases are the mock-ups of scenarios in the real world. We have created a set of use cases to evaluate the metric. However, due to the page limit, we include only three selected use cases in this paper.

The first two use cases represent two opposite extremes. In the first use case, each of the snapshots has different trip constellations. A series of such snapshots contain irregular trips. We imagine that such scenario will happen, if either an individual makes different trips each time or the observation of an adversary is of very bad quality such that there are high confusions or uncertainties associated with the obtained information. For each snapshot, the simulation first generates a random trip index in the range of 1 to 100, then it generates the corresponding probabilities. To avoid any subjectiveness in the probability assignment, the probabilities are randomly generated from the uniform distribution. The process is repeated for 60 snapshots.

In the second use case, all snapshots have the same trip constellation. However, only one trip in the constellation actually happens. Hence the snapshots contain a regular trip hidden among other observed trips. This scenario happens if an adversary has correctly observed the regular trip such as driving from home to work, but somehow cannot distinguish it from other trips observed at the same time. To simulate such scenario, we generate 60 snapshots with the same trip constellation, each with 100 trips. We set the trip T_1 in the constellation as the one actually happened and assign a fixed probability, called the p-value, to it. The remaining 99 trips are assigned with probabilities from the uniform distribution. We set the p-value to be the average, i.e., $p = 0.01$, and normalize the probabilities of the remaining 99 trips to be $\sum_{i=2}^{i=100} p_i = 0.99$. The choice and impact of the p-value will be further discussed in Section V-C.

The third use case locates on the spectrum between the two extreme cases described before, and contains several re-occurring trips. Thus it is a mock-up of a more realistic and common scenario. In this use case, we simulate the trip patterns specified in Table II. Imagine there is a series of snapshots capturing an individual's vehicle trips for several weeks. All snapshots cover a time period somewhen in the morning, so all the trips are from home to somewhere. We simulate this by four trip constellations. The first trip constellation for snapshots (Mon. – Wed.) contains trips $(T_1, T_4, \ldots, T_{100})$. We set T_1 as the trip actually happened and assign a p-value of 0.012. The corresponding probabilities of $(T_4, \ldots, T_{100})$ are assigned with probabilities from the uniform distribution, and normalized to be $\sum_{i=4}^{i=100} p_i = 0.988$. The second trip constellation for snapshots (Thur. – Fri.) contains trips $(T_2, T_4, \ldots, T_{100})$. We set T_2 as actually happened and also assign a p-value of 0.012, and the normalized probabilities to $(T_4, \ldots, T_{100})$. The third trip constellation for snapshots (Sat.) contains trips $(T_3, T_4, \ldots, T_{100})$. We assign a p-value of 0.012 to T_3 and the normalized probabilities to $(T_4, \ldots, T_{100})$. The last trip constellation for snapshots (Sun.) has trips $(T_4, \ldots, T_{100})$. To simulate random destinations on Sundays, we assign all the trips with probabilities from the uniform distribution. The simulation setup is also summarized in Table II. We repeat the process and generate 56 snapshots to simulate 8 weeks of snapshots with re-occurring trips.

TABLE II
3^{nd} USE CASE SETUP

Scenario		Simulation	
Week days	Trip (from Home to)	Trip constellation	Probability assignment
Mon. – Wed.	Office A	$(T_1, T_4, \ldots, T_{100})$	$p_1 = 0.012$ $\sum_{i=4}^{i=100} p_i = 1-p_1$
Thur. – Fri.	Office B	$(T_2, T_4, \ldots, T_{100})$	$p_2 = 0.012$ $\sum_{i=4}^{i=100} p_i = 1-p_2$
Sat.	Shopping mall C	$(T_3, T_4, \ldots, T_{100})$	$p_3 = 0.012$ $\sum_{i=4}^{i=100} p_i = 1-p_3$
Sun.	A random destination	$(T_4, \ldots, T_{100})$	$\sum_{i=4}^{i=100} p_i = 1$

In the simulations, the snapshot data of each use case is fed to the metric. The outcome of the metric is analyzed along the evaluation criteria. For our analysis, we choose the following entropy values: 1) H_{max}, the theoretical maximum entropy based on each single snapshot; 2) H, the entropy based only on single snapshot; 3) H_f, the entropy based on the snapshots modified by frequencies of occurrence; 4) H_w, the entropy based on the snapshots modified by average probabilities; 5) H_B, the entropy based on the snapshots modified by Bayesian method.

To analyze the impact of accumulated information on the actual level of uncertainty, we further define H_d as a measurement of the *decrease in uncertainty*

$$H_d = \frac{H_B - H}{H} 100\% \tag{7}$$

which bases the calculation on the difference of the entropy using Bayesian method and the entropy based on single snapshot without any additional information.

C. Simulation

Fig. 5 shows the simulation result from the first use case, in which each snapshot contains a randomly generated trip constellation. We can see from the figure that the entropies of H, H_f, H_w, and H_B are so close that they overlap each other most of the time. This means neither frequency based approach nor Bayesian approach are able to benefit from the accumulated information. Besides, these entropies are very close to the upper-bound H_{max}, due to the fact that the probabilities in each snapshot are from uniform distributions. For illustrative reason, the lower part of the figure includes a bar chart showing the number of trips in each of the snapshots. Notice that the actually trip constellations are not shown in the bar chart.

Fig. 6 shows what the metric returns from the second use case, which simulates the scenario that a regular trip is blurred by other false observations in each snapshot. The result shows that the frequency based approaches

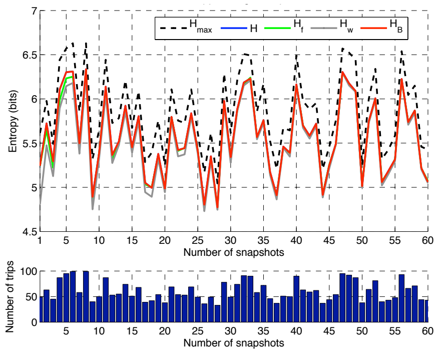

Fig. 5. Entropy of irregular trips

can barely utilize the accumulated information. As a result, H_f and H_w mostly overlap H, with the exception that H_w has slightly lower entropies at the first few snapshots. On the other hand, Bayesian method has significantly decreased the entropy level from 6.3 bits to as low as 0.79 bits at the 33^{th} snapshot. Obviously, at 0.79 bits, the uncertainty is very low, i.e., the privacy level is very low. The shape of the curve of H_B suggests that Bayesian method is able to process and benefit from the accumulated information.

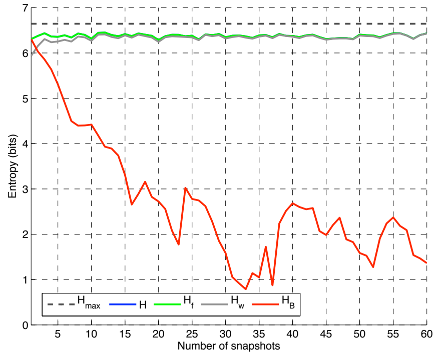

Fig. 6. Entropy of regular trips

Fig. 7 shows the simulation result from the third use case. The third use case simulates weekly re-occurring trips. H_f and H_w have similar outcomes as those in Fig. 6, i.e., frequency based approaches can not really benefit from accumulated information. Again, Bayesian method has significantly decreased the entropy value. Interestingly, this time the curve of H_B has a cascading and downward pattern. The reason is that we have simulated four types of re-occurring trips in this use case. The first three trips are regularly occurred trips and the fourth one (i.e., the Sunday trip) is chosen to be random. Therefore, while the overall curve of H_B demonstrates a downward trend, the entropies corresponding to the first three trips decrease much faster than the entropy of the Sunday trip. Notice that the entropy of the Sunday trip also exhibits a downward trend. The reason is that even though the probability distributions of the Sunday trip are from the uniform distribution, their values are slightly different among each others. As a result, the probabilities are modified by Bayesian approach towards a non-uniform distribution. In other words, given consecutive snapshots, our algorithm regards some of the trips are “more likely to have happened” than others. The result again demonstrates that Bayesian approach can take advantage of the accumulated information caused by regularly occurring trips.

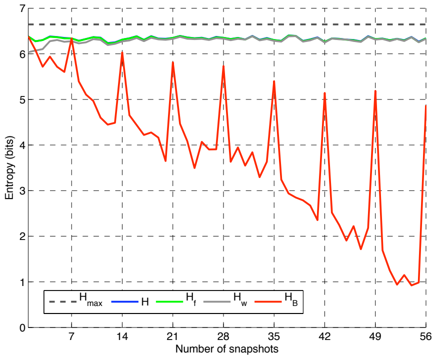

Fig. 7. Entropy of re-occurring trips

As the next step, we use H_d to analyze the decrease in uncertainty in each of the use cases. Since a new set of random values is generated each time a use case is simulated, we run each use case 100 times to take into account the effects of the variations of random variables. We calculate the mean value of the results from the three use cases and plot them in Fig. 8. For irregular trips, taking more snapshots into the metric does not decrease information uncertainty. In some cases, it even increases the level of uncertainty. This means based on the metric, accumulated information does not provide any additional information due to the randomness in the captured information. For regular trips, we can see that there is a constant decrease in uncertainty as more and more snapshots are added in the sequence. The decrease reaches -84.6% at the 60^{th} snapshot. The outcome of the metric shows that with regular trips, accumulated information can significantly reduce the uncertainty in the information related to user location privacy. For re-occurring trips, despite the spikes on each Sunday due to the randomness of the trips on that day, there is also a constant decrease in uncertainty as the time elapses. Because there are several regular trip patterns involved in this use case, the speed of the decrease in uncertainty is slower than the one in the previous use case. The result demonstrates again that the accumulated information can cause considerable decreases in the level of uncertainty, i.e., users’ location privacy. Notice that the curves in Fig. 8 correspond to those appeared in Fig. 5, 6, and 7, i.e., the observations we made before on single simulation result also hold in general cases.

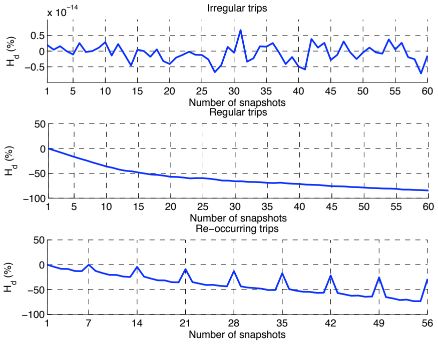

Fig. 8. Change of uncertainty

We know that the main reason behind the significant decrease in uncertainty is because of the algorithm based on Bayesian method. The algorithm processes, propagates, and utilizes the accumulated information by continuously updating the probabilities of each hypothesis each time it receives a set of new evidence contained in a snapshot. The updated hypotheses are kept in the belief table B. As a result, the probability distributions in the belief table converge toward the "real happened" trips. The changing of probability distributions leads to lower entropy values hence the decrease in uncertainty. However, so far we have not shown whether the algorithm is able to update probability distributions in a correct way. We test the correctness of Algorithm 1 by tracing the change of beliefs in the algorithm. In this sense, the second and the third use case are quite similar. Therefore, we only show the study on the second use case here. Same as before, we assign the first trip as the one actually happened. Furthermore, we assign different probabilities to study the effect of the p-values on the performance of the algorithm. The p-values are $\{0.009, 0.01, 0.011\}$, which correspond to 10% lower than the average, the average, and 10 % higher than the average of the probability of the 100 trips in the trip constellation. Again, we run the simulation 100 times to account for the variations in the random data set and take the means of the values of the first trip from the belief table.

Fig. 9 shows the result. At 10% below the average, the algorithm fails to detect the trip. However, as soon as the p-value is of the average value, there is a steady rise of the probability. If we assume that 0.5 is the threshold to select a trip as the one really happens, the first trip will be selected at the 59^{th} snapshot. Only slightly increase the p-value 10% higher, the probability of the first trip exhibits a sharp rise and passes the 0.5 threshold at the 32^{th} snapshot.

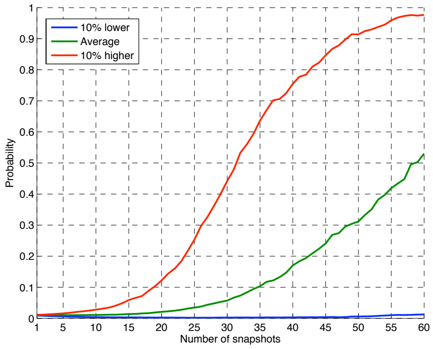

Fig. 9. Change of beliefs with different p-values

D. Discussion

The simulation results show that the privacy metric fulfills both evaluation criteria defined in Section V-A. However, as already shown, the algorithm only functions well on regular trip patterns. We can consider this issue from two angles. On one hand, one can develop heuristics to assist the functioning of the algorithm. For example, trimming and keeping only the trips with high probabilities can limit the number of trips in the constellations and facilitate the "learning from the past" process. On the other hand, a privacy-protection mechanism might exploit this feature by virtually creating irregular trip patterns to render accumulated information useless to a potential adversary.

VI. Related work

Anonymity and (un)linkability are two common approaches to express user privacy in communication systems. A definition on these two terms is given in [10] and unlinkability is further refined in [11].

The size of the anonymity set is a popular metric of anonymity. The authors of [12], [13] point out that the size of the anonymity set does not reflect different probabilities of the members in an anonymity set, and propose to use entropy as the metric for anonymity. Beresford et al. [14] use entropy to quantify the information obtained by an adversary on the user movements through mix zones. Applying the same principle, the authors in [15] and [16] use the entropy provided by the mix zones to evaluate the level of location privacy achieved by the vehicles in vehicular ad hoc neworks (VANET). Tracking, which learns a vehicle's movement by linking a series of messages from that vehicle, is another common approach to measure location privacy. Gruteser et al. [17] propose to use tracking algorithms to characterize the level of location privacy. Sampigethaya et. al [3] use maximum tracking time to evaluate the location privacy of vehicles in VANET. Hoh et. al [18] use the mean time to confusion to measure the privacy level of vehicles sending GPS traces in a traffic monitoring system. Fischer et. al [19] propose to measure unlinkability of sender-message relations based on the outer and inner structures of the set partitions of the observed messages. Most approaches to location privacy focus on location information. In [5], we propose that metrics for location privacy in V2X systems should take both individuals and their vehicle movements into considerations.

The impact of accumulated information on location privacy has not been explicitly addressed in most of these approaches so far. Mostly, it is assumed that an adversary's knowledge on a system already reflects its longtime observations at the time of attack. Empirical studies such as [20] use two weeks of recorded pseudonymous location tracks to infer home addresses and identities of the drivers with partial successes. Outside the communication domain, the authors of [21] find out that snapshot-based, time-invariant approaches cannot cope with the emergence of time series data mining, and propose to add the time dimension to the current research on privacy-preserving data mining.

VII. Conclusion and future work

In this paper we present a trip-based location privacy metric for measuring location privacy of the users of V2X communication systems. To reflect the true underlying privacy values, the metric includes accumulated information and reflects the impact in the privacy measurement. We model the accumulated information and develop approaches to process, propagate, and utilize the accumulated information in the metric. We evaluate the viability and correctness of the metric by various case studies and extensive simulations. Our simulations show that under certain conditions, accumulated information can significantly decrease users' location privacy. We show that our metric is a valuable tool to evaluate and develop privacy-protection mechanisms for the users of V2X systems.

In future work, we will further evaluate our metric with more scenarios and realistic V2X applications. The evaluation will also include existing privacy-protection mechanisms proposed to V2X systems. The current metric only measures privacy of individual users. The possible interrelations among individuals and their impacts on the level of location privacy will be investigated to determine location privacy in a global view. The metric is extensible, which means when it is necessary, we can add other identified attacks on location privacy to the metric in the future.

Acknowledgment

This work was partially supported by the European Commission through the PRECIOSA project (IST-224201).

References

[1] Jean-Pierre Hubaux, Srdjan Čapkun, and Jun Luo, "The security and privacy of smart vehicles," *IEEE Security and Privacy*, vol. 4, pp. 49–55, 2004.

[2] F. Dötzer, "Privacy issues in vehicular ad hoc networks," in *Workshop on Privacy Enhancing Technologies*, 2005.

[3] K. Sampigethaya, M. Li, L. Huang, and R. Poovendran, "Amoeba: Robust location privacy scheme for vanet," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 8, pp. 1569 – 1589, Oct. 2007.

[4] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Enhancing security and privacy in traffic-monitoring systems," *IEEE Pervasive Computing*, vol. 5, no. 4, pp. 38–46, 2006.

[5] Z. Ma, F. Kargl, and M. Weber, "A location privacy metric for v2x communication systems," in *IEEE Sarnoff Symposium*, Princeton, NJ, USA, March 2009.

[6] P. Papadimitratos, L. Buttyan, T. Holczer, E. Schoch, J. Freudiger, M. Raya, Z. Ma, F. Kargl, A. Kung, and J.-P. Hubaux, "Secure vehicular communications: Design and architecture," *IEEE Communications Magazine*, vol. 46, no. 11, pp. 100–109, November 2008.

[7] Z. Ma, F. Kargl, and M. Weber, "Pseudonym-on-demand: a new pseudonym refill strategy for vehicular communications," in *WiVeC 2008*, Calgary, Canada, September 2008.

[8] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, pp. 379–423, 623–656, July, October 1948.

[9] M. E. Ben-akiva and J. L. Bowman, "Activity based travel demand model systems," in *Equilibrium and Advanced Transportation Modeling*. Kluwer Academic, 1998, pp. 27–46.

[10] A. Pfitzmann and M. Hansen, "Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management – a consolidated proposal for terminology," Tech. Rep., Feb. 2008, v0.31. [Online]. Available: http://dud.inf.tu-dresden.de/Anon_Terminology.shtml

[11] S. Steinbrecher and S. Köpsell, "Modelling unlinkability." in *Workshop on Privacy Enhancing Technologies*, 2003, pp. 32–47.

[12] A. Serjantov and G. Danezis, "Towards an information theoretic metric for anonymity," in *Workshop on Privacy Enhancing Technologies*, 2002.

[13] C. Diaz, S. Seys, J. Claessens, and B. Preneel, "Towards measuring anonymity," in *Workshop on Privacy Enhancing Technologies*, 2002.

[14] A. R. Beresford and F. Stajano, "Location privacy in pervasive computing," *IEEE Pervasive Computing*, vol. 2, no. 1, pp. 46–55, 2003.

[15] L. Buttyan, T. Holczer, and I. Vajda, "On the effectiveness of changing pseudonyms to provide location privacy in VANETs," in *ESAS 2007*, July 2007.

[16] J. Freudiger, M. Raya, M. Felegyhazi, P. Papadimitratos, and J.-P. Hubaux, "Mix-zones for location privacy in vehicular networks," in *WiN-ITS*, 2007.

[17] M. Gruteser and B. Hoh, "On the anonymity of periodic location samples." in *Security in Pervasive Computing 2005, Boppard, Germany*, vol. 3450, 2005, pp. 179–192.

[18] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Preserving privacy in GPS traces via density-aware path cloaking," in *ACM Conference on Computer and Communications Security (CCS)*, 2007.

[19] L. Fischer, S. Katzenbeisser, and C. Eckert, "Measuring unlinkability revisited," in *WPES '08: Proceedings of the 7th ACM workshop on Privacy in the electronic society*, Alexandria, Virginia, October 27 2008.

[20] J. Krumm, "Inference attacks on location tracks," in *Fifth International Conference on Pervasive Computing*, Toronto, Canada, May 2007, pp. 127–143.

[21] Y. Zhu, Y. Fu, and H. Fu, "On privacy in time series data mining," in *PAKDD*, 2008, pp. 479–493.