Evaluation of Privacy Preserving Algorithms Using Traffic Knowledge Based Adversary Models

Zhanbo Sun, Bin Zan*, Jeff (Xuegang) Ban, Marco Gruteser* and Peng Hao
Rensselaer Polytechnic Institute
110 8th St. Troy, NY 12180-3590
{sunz2, banx, haop}@rpi.edu
*WINLAB, Rutgers University
*671 Route 1 South, North Brunswick, NJ 08902-3390
*{zanb, gruteser}@winlab.rutgers.edu

Abstract—**By providing location traces of individual vehicles, mobile traffic sensors have quickly emerged as an important data source for traffic applications. In dealing with the privacy issues associated with this, researchers have been proposing different privacy protection algorithms. In this paper, we propose traffic-knowledge-based adversary models to attack privacy algorithms. By doing so, we can compare and evaluate different privacy algorithms in terms of both privacy protection and the convenience for traffic modeling. Results show that by having a relatively good privacy performance, the released datasets of both the 3.3 level of confusion entropy and the 0.1 individual likelihood can still be applied for a fine level of traffic applications.**

I. Introduction

MOBILE traffic sensors – those move with the traffic flow and have the potential to track the movement of individual vehicles by using location traces – have quickly emerged as an important data source and been widely used. Meanwhile, there are always privacy concerns associated with this approach [4], [7]. Some proposed privacy preserving methods try to address this issue via naïve anonymization techniques which simply remove vehicle identifiers [12]. However, by linking the driving pattern and sensitivity locations (e.g., home end, office building, etc.) with the driver, location traces can be easily re-identified. Other methods try to protect privacy by perturbing or reducing the accuracy for either spatial or temporal information [8], [11], [2]. In these cases, however, transportation modelers usually hesitate to use such datasets for traffic applications, especially for those requiring high accuracy spatial and temporal information (e.g., for arterial performance measurements). Similarly, reducing the sampling frequency [10] is also not a very promising approach, since the sampled location traces (say in a 1-minute sampling interval) can barely be used for a fine level of traffic applications.

Hoh et al. (2007) [4] introduced a novel time-to confusion metric to evaluate group-wide privacy in a set of location traces. They then propose an uncertainty-aware path cloaking algorithm which yields the time-to-confusion criterion, in which the uncertainty of tracking is measured by entropy. Hoh et al. (2008) [5] proposed the idea of Virtual Trip Line (VTL), and showed that by using VTLs to regulate location and speed reports, privacy violations can be reduced. On top of that, Zan et al. (2011) [6] proposed a VTL zone-based path cloaking algorithm. The algorithm predefines VTL zones over the intersections of interests and only those vehicle trajectories within VTL zones (which also need to satisfy the entropy criterion) can be released. Using the released datasets as input, the success rate of traffic applications (in this case, queue length estimation) is acceptable. These existing algorithms use purely statistical traffic knowledge; information such as path likelihood and travel time distributions is estimated using historical data. Interestingly enough, the link between two neighboring VTL zones is actually similar to the concept of mix-zone as proposed by Bereford and Stajano (2004) [1]. In fact, the VTL-zone based algorithm defines the areas where mobile data should be collected, which can be considered as the opposite way of the mix-zone algorithm that defines areas where data should be suppressed. We believe the VTL-zone concept, by focusing on where data should be collected, minimizes the data collection effort and can better satisfy the data needs for traffic applications. For more comparisons between VTL zone based path cloaking algorithm and mix zone approaches, one can refer to [6].

In this paper, we evaluate how resilient these algorithms are to adversaries with more sophisticated traffic knowledge. In particular, we consider knowledge of travel time in the mix zones and signal delay patterns. We then measure how effective these are in linking released vehicle trajectories from two VTL zones. Moreover, since entropy is not a very intuitive metric in the transportation community, we also present results in terms of tracking success probability. The algorithms are evaluated using microscopic traffic simulation. The results reveal that by having a relatively good privacy performance, the released datasets of both the 3.3 level of confusion entropy and the 0.1 individual likelihood can still be applied for a fine level of traffic applications.

The paper is comprised of 6 sections. In Section 2, we introduce several privacy algorithms. We then propose our traffic-knowledge-based adversary models in Section 3.

Manuscript received May 15, 2011.

Section 4 describes the performance measures we use to evaluate different privacy algorithms. Experiments and numerical results are shown in Section 5, followed by the conclusions in Section 6.

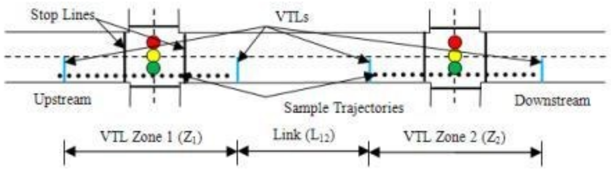

Fig. 1. VTL zones (Case 1)

II. THE PRIVACY PRESERVING ALGORITHMS

In this section, we introduce several privacy algorithms including the baseline case, random sampling, individual likelihood based and entropy based algorithms. Detailed discussions of some of those algorithms can be found in [6].

A. *VTL Zone and Baseline Case*

Being comprised of two Virtual Trip Lines (VTLs, [5]), one VTL zone covers one direction of an intersection. The two VTLs are designated to avoid major deceleration and acceleration processes due to traffic signals. GPS trace data is only collected within VTL zones: starting with the sample right before a vehicle enters into a VTL zone and ending with the sample right before a vehicle leaves a VTL zone. See Fig. 1. In this paper, the metric of privacy is based on the chance that vehicle trajectories may be linked at two VTL zones.

In the baseline case, all the vehicle trajectories within VTL zones are released. Because of the discontinuity of vehicle trajectories at the link between two VTL zones, it is not a trivial task for the adversary to keep tracking the vehicle trajectories between two VTL zones.

B. *Random Sampling*

On top of the baseline case, in order to enhance the level of privacy, a naïve random sampling approach is proposed. In which a proportion of the sample trajectories is randomly selected and released at each VTL zones, so that it is even harder for the adversary to continuously tracking vehicle trajectories between two VTL zones.

C. *Individual Likelihood*

Inspired by the individual likelihood of being tracked [6], we use equation (1) as a privacy metric, which illustrates the likelihood of taking a given time period for a particular vehicle to go from one VTL zone to another, normalized over the summation of the likelihoods of all the possible vehicles which cannot be distinguished from the object vehicle.

$$p^d_{v\to c} = \frac{\rho_{v\to c} * p_T(t^d_{v\to c})}{\sum_{v'\in V\backslash c}\sum_{d'\in D_{v'}} \rho_{v'\to c} * p_T(t^{d'}_{v'\to c})} \qquad (1)$$

Where $D$is the whole set of vehicle IDs, V is the whole set of VTL zones, d is the target vehicle we are trying to track, which have the latest disclosed trajectory in VTL zone v, c is the current VTL zone we are looking at, $\rho_{v\to c}$ is the likelihood of any vehicle go through the path from v to c, which is calculated using historical data, $t^d_{v\to c}$ is the time period it takes for vehicle d to go from VTL zone v to c, and $p_T(\cdot)$ is a discrete probability density function of the travel time distribution (from v to c), which yields a three parameter Log-normal distribution, parameterized by Least Square Estimation (LSE) using historical data.

Notice that the trajectory of vehicle d at VTL zone c cannot be released when the individual likelihood $p^d_{v\to c}$ is larger than a predefined level, e.g., 0.1.

D. *Entropy*

Some researchers have used entropy to measure the tracking uncertainty [4], [6] and equation (2) is proposed to calculate the system-wide tracking uncertainty.

$$H = -\sum_{v\in V\backslash c}\sum_{d\in D_v}(p^d_{v\to c}\log p^d_{v\to c}) \qquad (2)$$

Where H is the entropy metric and other related terms have already been defined in equation (1). By comparing the privacy metric H with a predefined confusion level α, we can then determine if one sample can be released or not. Intuitively, entropy is related to the factor that among how many vehicles the target vehicle can be indistinguishable. It is also an aggregated measure of the probability of not tracking vehicles between two VTL zones. For example, when the entropy is equal to 0.95, this is about to say that one target vehicle is indistinguishable between two vehicles. We also try to combine both individual likelihood and entropy as the privacy metrics, and the overall performance has no major deviation from the one of individual likelihood dataset.

III. TRAFFIC-KNOWLEDGE-BASED ADVERSARY MODEL

In this section, we propose traffic-knowledge-based adversary models, which can be used to attack the released datasets by trying to link vehicle trajectories from two different VTL zones together. Two general cases are considered.

A. *Case 1*

Consider two neighboring VTL zones (Z_1 and Z_2) which cover two consecutive intersections, and with one link (L_{12}) in between these two VTL zones, as shown in Fig. 1. On L_{12}, since vehicles are usually proceeding at a speed close to the Free Flow Speed and the acceleration/deceleration processes are usually unnoticeable, the travel time on L_{12} is very stable. In this paper, we propose two methods to estimate the travel time on L_{12}, one is simply the Free Flow Travel Time, which can be obtained by the length of L_{12} divided by the design speed; the other is an adjusted travel time, which can be obtained by the length of L_{12} divided by the estimated average speed on this link. The estimated average speed is calculated by taking the average of the speed of the last sample in Z_1 and the first sample in Z_2. In essence, the adversary does not know which travel time is closer to the ground truth; however, it is conservative to assume that the adversary will always choose the travel time which has the best performance, referred as T_{12}.

Now target on a set of released vehicles (Ω_1) which go through Z_1. Then for vehicle $n \epsilon \Omega_1$, based on its trajectory in Z_1, it is easy to tell when it leaves Z_1, referred as T_1^n. The time that n enters Z_2 can then be approximated as $T_1^n + T_{12}^n$, referred as T_2^n. See equation (3).

$$T_2^n \approx T_1^n + T_{12}^n \quad (3)$$

If we slightly relax the travel time estimation and give a threshold T_t, then the vehicle enters VTL zone 2 within time period $[T_2^n - T_t, T_2^n + T_t]$ are very likely to be the same vehicle as vehicle n. In other words, consider a set of released vehicles (Ω_2) which go through Z_2, then for vehicle $m \epsilon \Omega_2$, with T_2^m as the time that vehicle m enters into Z_2, if $T_2^n - T_t \leq T_2^m \leq T_2^n + T_t$, we add m into a suspect list (S), meaning that m is likely to be the same vehicle as n. If S is not empty, we can then choose the vehicle $\hat{k}$ which satisfies equation (4) as an inference.

$$\hat{k} = \text{argmin}_{k \in S} |T_2^k - T_2^n| \quad (4)$$

That means $\hat{k}$ is the vehicle whose entrance time to Z_2 is the closest to the estimated entrance time of vehicle n (T_2^n). If the inference is correct ($\hat{k} = n$), the vehicle trajectories at two neighboring VTL zones can be linked, the privacy is thus violated.

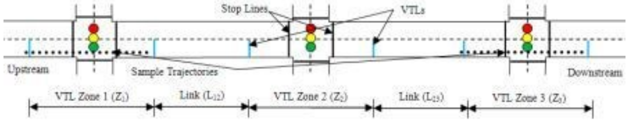

Fig. 2. VTL zones (Case 2)

B. Case 2

Consider two VTL zones which are not neighbored with each other (e.g., Z_1, Z_2 and Z_3, which cover a corridor with three intersections, and the adversary model is trying to link the vehicle trajectories from Z_1 to Z_3, as shown in Fig. 2). Following the same logic as Case 1, the travel times on the links (L_{12}, L_{23}) between two consecutive VTL zones are very stable, which can be estimated as T_{12} and T_{23}, using the same approaches as Case 1.

Now we attempt to link the trajectories from a set of released vehicles (Ω_1) which go through Z_1 to a set of released vehicles (Ω_3) which go through Z_3. For vehicle $n \epsilon \Omega_1$, the time that n enters Z_3 can then be approximated as equation (5).

$$T_3^n \approx T_1^n + T_{12}^n + T_{2,D}^n + T_{23}^n \quad (5)$$

Where T_3^n is the estimated time that n enters Z_3, T_1^n is the time that n leaves Z_1; T_{12}^n and T_{23}^n are the estimated travel time of n on the links between Z_1 and Z_2, and between Z_2 and Z_3, respectively; $T_{2,D}^n$ is the travel time of vehicle n within Z_2, which can be estimated via the delay pattern of Z_2 [14]. See Fig. 3. Notice that the delay pattern of Z_2 is reconstructed using the trajectories of a set of released vehicles (Ω_2) which go through Z_2, and the signal timing information. Thus for vehicle $m \epsilon \Omega_3$, with T_3^m as the time that vehicle m enters into Z_3, if $T_3^n - T_t \leq T_3^m \leq T_3^n + T_t$, we add m into a suspect list (S). If S is not empty, we can then choose the vehicle $\hat{k}$ which satisfies equation (6) as an inference.

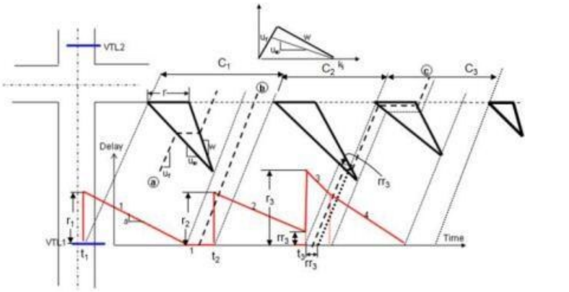

Fig. 3. Reconstruction delay pattern. A typical signalized intersection is shown here. Given the trajectories of a set of released vehicles $\Omega_2$that go through Z_2 (dashed lines) and the signal timing information, we can use the bold solid triangles (in the upper part) to represent how queue forms and dissipates. By analyzing the geometry of the triangles, we can then construct the theoretical delay curve (piecewise linear curves at the bottom), which indicates the travel time an imaginary vehicle will experience when arriving at the intersection at a given time.

$$\hat{k} = \text{argmin}_{k \in S} |T_3^k - T_3^n| \quad (6)$$

If the inference is correct ($\hat{k} = n$), the vehicle trajectories at Z_1 and Z_3 can be linked, the privacy is thus violated.

IV. Performance Measures

In this section, we describe the measures we use to evaluate the performance of the privacy algorithms, with respect to both privacy protection and the convenience for traffic modeling.

A. Privacy Protection

In terms of privacy protection, the performance of the privacy models can be evaluated by applying the adversary models. In this paper, we use two measures to indicate this, namely, the percentage of tracked trajectories (P_1) and the percentage of correct inference (P_2). Obtained by using the number of correct inference divided by the total number of trajectories (do not necessarily need to be released) going through both the VTL zones (e.g., both Z_1 and Z_2 in Case 1; Z_1, Z_2 and Z_3in Case 2), P_1 indicates the probability that the trajectories of one vehicle can be successfully linked at the two VTL zones. Different from P_1, P_2 is obtained by using the number of correct inference divided by the total number of inference, which indicates how accurate the inferences are.

From the perspective of individuals, P_1 may seem to be more important, since it reveals the potential risk that one vehicle trajectory can be tracked. However, P_2 is indeed equally important, because even though the trajectories can be linked, the linkage is highly possible to be incorrect, thus the adversary has no way to prove that the linked trajectories correspond to the same driver.

Notice that in our adversary models, one released vehicle trajectory in the current VTL zone (e.g., Z_1, as of Case 1) can at most correspond to one inference (which is selected from the suspect list) based on equation (4) or (6). It is also worthy to mention that there are some situations in which the suspect list (S) is empty, therefore no inference can be made in these situations.

B. Convenience for Traffic Modeling

A tradeoff usually exists between privacy protection and traffic applications, meaning by having a high level of privacy, transportation researchers may have to, to some extent, sacrifice the ease of traffic modeling. Thus it is important to make sure that after applying the privacy algorithms, the released datasets can still be used for traffic modeling, especially (and to the greatest interest of the author) arterial performance estimation (e.g., queue length estimation, delay pattern estimation, etc.). Two measures are used in our work: one is the released number of trajectories in each VTL zone (compared with the total number of trajectories in the baseline case); the other is the percentage of the number of cycles (out of the total number of cycles in the simulation) for which queue length estimation can be successfully performed, which is defined as the success rate in [13].

V. Experiment and Numerical results

In this section, we evaluate the privacy algorithms using the performance measures we mentioned in section IV. Detailed results of the numerical experiment are presented here.

A. Simulation Settings

The traffic simulation is run in Paramics for about an hour. Vehicle trajectory data is then extracted from a sub-network of the SR41 corridor located at the city of Fresno, CA [9]. The selected network covers over 90 signalized intersection and 15 ramp metering controller. A total number of 102 VTL zones are deployed. See Fig. 4.

B. Evaluation in Terms of Privacy Preserving

By using the traffic-knowledge-based adversary model to attack the released datasets, we can evaluate the performance of the privacy algorithms in terms of privacy preserving.

1) Case 1

For Case 1, we list the results of 10 pairs of VTL zones, in which the traffic volumes are relatively large. Table 1 shows the performance of baseline case: column 2 and column 3 are the upstream and downstream VTL zones, expressed by the link sequence that one VTL zone covers; column 4 is the number of released trajectories which go through both Z_1 and Z_2 (as of the baseline case, trajectories of all vehicles are released); column 5 and column 6 are the number of released trajectories within Z_1 and Z_2, respectively; column 7 is the number of inferences can be made, notice that one released trajectory in Z_1 can at most correspond to one inference (it is possible that S is empty, no inference can be made in this case); column 8 is the number of correct inferences; column 9 corresponds to P_1 in section IV, which can be obtained by using column 8 divided by the actual number of trajectories which go through both Z_1 and Z_2; column 10 corresponds to P_2 in section IV, which can be obtained by using column 8 divided by column 7.

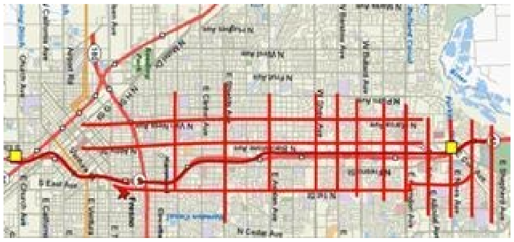

Fig. 4. Simulation network

Table 1 essentially illustrates that the idea of only releasing vehicle trajectories within VTL zones is, to some extent, useful to preserve privacy; however, this may not be sufficient since there are still a large proportion (about 35% in average) of vehicles trajectories can be successful tracked.

The results for random sampling (50% trajectories are released) case are shown in Table 2. Compared with the baseline case, random sampling models are able to protect privacy by filtering out some sample trajectories at each VTL zone, so that fewer trajectories can be tracked and more inaccurate the inference becomes. As it is shown in Fig. 5, when sampling rate decreases, more trajectories are filtered out, and as a result, the first privacy measure (P_1) drops. However, the second privacy measure (P_2) may not necessarily decrease as the sampling rate increases, meaning the random sampling method cannot guarantee the accuracy of the inferences will decrease as more samples are filtered out, see Fig. 6.

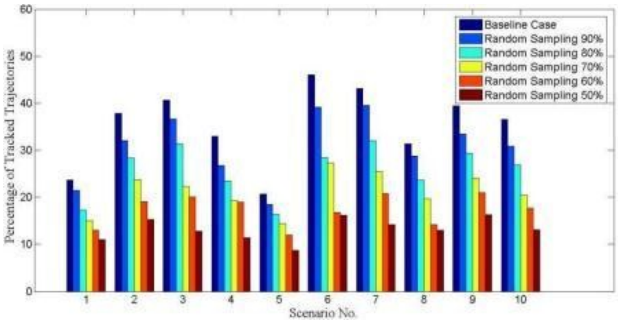

Fig. 5. Baseline case vs. random sampling (P_1)

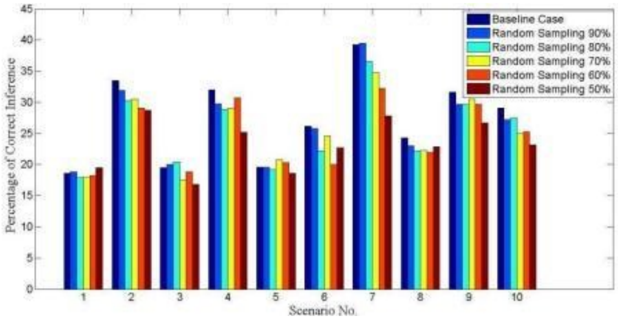

Fig. 6. Baseline case vs. random sampling (P_2)

Table 1. Privacy performance of the baseline dataset (Case 1)

No.	Z_1	Z_2	No. of Released Trajectories (Both Z_1 and Z_2)	No. of Released Trajectories(Z_1)	No. of Released Trajectories (Z_2)	No. of Inferences	No. of Correct Inferences	Percentage of Tracked Trajectories	Percentage of Correct Inferences
1	597_598_599	601_602	1136	1496	1225	1446	269	23.7%	18.6%
2	554_555	556_557	517	586	1144	585	196	37.9%	33.5%
3	556_557	559_560	452	1144	622	944	184	40.7%	19.5%
4	559_560	562_563	579	622	757	597	191	33.0%	32.0%
5	612_613_614	616_617	886	969	1112	934	183	20.7%	19.6%
6	658_659	660_661_662	495	999	642	871	228	46.1%	26.2%
7	780_781_782	784_785	424	475	923	466	183	43.2%	39.3%
8	690_691_692	692_693_694	625	915	672	807	196	31.4%	24.3%
9	692_693_694	694_695_696	514	672	599	643	203	39.5%	31.6%
10	529_530	531_532_533	673	892	765	846	246	36.6%	29.1%

Table 2. Privacy performance of the 50% random sampling dataset (Case 1)

No.	Z_1	Z_2	No. of Released Trajectories (Both Z_1 and Z_2)	No. of Released Trajectories(Z_1)	No. of Released Trajectories (Z_2)	No. of Inferences	No. of Correct Inferences	Percentage of Tracked Trajectories	Percentage of Correct Inferences
1	597_598_599	601_602	278	729	604	640	125	11.0%	19.5%
2	554_555	556_557	136	309	564	275	79	15.3%	28.7%
3	556_557	559_560	104	564	306	345	58	12.8%	16.8%
4	559_560	562_563	136	306	380	262	66	11.4%	25.2%
5	612_613_614	616_617	231	497	559	413	77	8.7%	18.6%
6	658_659	660_661_662	135	510	314	353	80	16.2%	22.7%
7	780_781_782	784_785	97	237	443	216	60	14.2%	27.8%
8	690_691_692	692_693_694	169	459	362	353	81	13.0%	22.9%
9	692_693_694	694_695_696	140	362	299	315	84	16.3%	26.7%
10	529_530	531_532_533	172	427	389	380	88	13.1%	23.2%

Table 3. Privacy performance of the 0.1 individual likelihood dataset (Case 1)

No.	Z_1	Z_2	No. of Released Trajectories (Both Z_1 and Z_2)	No. of Released Trajectories(Z_1)	No. of Released Trajectories (Z_2)	No. of Inferences	No. of Correct Inferences	Percentage of Tracked Trajectories	Percentage of Correct Inferences
1	597_598_599	601_602	341	1481	440	890	95	8.4%	10.7%
2	554_555	556_557	177	519	870	412	75	14.5%	18.2%
3	556_557	559_560	88	870	343	407	21	4.6%	5.2%
4	559_560	562_563	36	343	466	262	11	1.9%	4.2%
5	612_613_614	616_617	227	560	673	462	76	8.6%	16.5%
6	658_659	660_661_662	51	468	413	266	11	2.2%	4.1%
7	780_781_782	784_785	91	334	709	296	58	13.7%	19.6%
8	690_691_692	692_693_694	159	669	307	387	60	9.6%	15.5%
9	692_693_694	694_695_696	90	307	401	261	60	11.7%	23.0%
10	529_530	531_532_533	292	750	480	613	109	16.2%	17.8%

Table 4. Privacy performance of the 3.3 level of confusion entropy dataset (Case 1)

No.	Z_1	Z_2	No. of Released Trajectories (Both Z_1 and Z_2)	No. of Released Trajectories(Z_1)	No. of Released Trajectories (Z_2)	No. of Inferences	No. of Correct Inferences	Percentage of Tracked Trajectories	Percentage of Correct Inferences
1	597_598_599	601_602	622	1492	709	1214	185	16.3%	15.2%
2	554_555	556_557	150	570	778	396	84	16.2%	21.2%
3	556_557	559_560	174	778	461	414	66	14.6%	15.9%
4	559_560	562_563	41	461	286	172	30	5.2%	17.4%
5	612_613_614	616_617	125	496	574	307	46	5.2%	15.0%
6	658_659	660_661_662	30	385	364	118	13	2.6%	11.0%
7	780_781_782	784_785	28	345	550	93	18	4.2%	19.4%
8	690_691_692	692_693_694	100	505	276	172	43	6.9%	25.0%
9	692_693_694	694_695_696	80	276	363	209	51	9.9%	24.4%
10	529_530	531_532_533	226	583	478	488	107	15.9%	21.9%

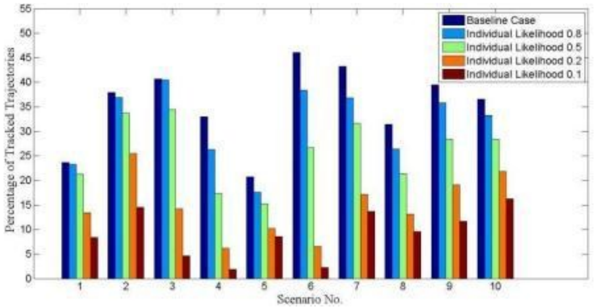

Fig. 7. Baseline case vs. individual likelihood (P_1)

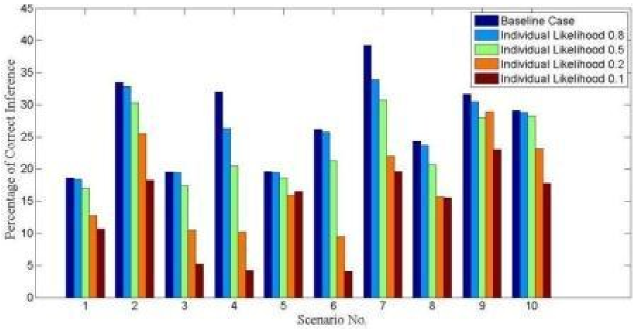

Fig. 8. Baseline case vs. individual likelihood (P_2)

Table 3 indicates the results of 0.1 individual likelihood dataset. By compared with Table 2, one can tell that while releasing much more samples than the random sampling dataset (e.g., Scenario No. 1, 2) or almost the same number of samples at each VTL zone (e.g. Scenario No. 5, 8), the individual likelihood dataset has overall higher level of privacy.

We also show the results for different individual likelihoods (e.g., 0.2, 0.5, 0.8), as shown in Fig. 7 and Fig. 8. In contrast to the random sampling method, a stricter threshold here leads to decreases in both P_1 and P_2, resulting a higher level of privacy.

The results for the 3.3 level of confusion entropy dataset are shown in Table 4. By comparing this with Table 3, one can find that in terms of privacy protection, these two datasets are comparable with each other, at least under the attack models considered here. For P_1, 0.1 individual likelihood dataset has better performance in 5 out of 10 examples; for P_2, 0.1 individual likelihood dataset has better performance in 8 out of 10 examples. As expected, a higher level of privacy can be obtained by using a higher level of confusion, as shown in Fig. 9 and Fig. 10.

Fig. 11 and Fig. 12 compare the performance of datasets generated by different privacy algorithms. Notice that the amounts of sample trajectories these three datasets release at each VTL zone are generally comparable. In terms of P_1, the performance of these three datasets are very close to each other, the 0.1 individual likelihood and the 3.3 level of confusion entropy dataset are slightly better than 50% random sampling dataset; in terms of P_2, 0.1 individual likelihood dataset has the best performance, followed by the 3.3 level of confusion entropy dataset and then the 50% random sampling dataset.

2) Case 2

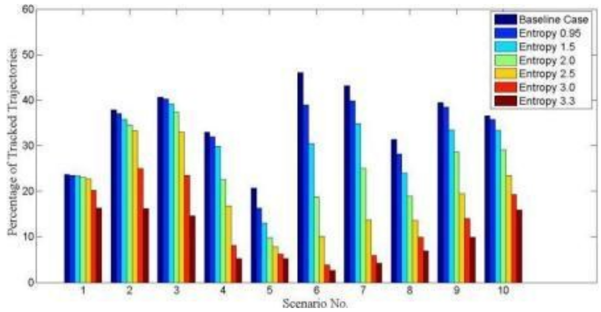

Fig. 9. Baseline case vs. entropy (P_1)

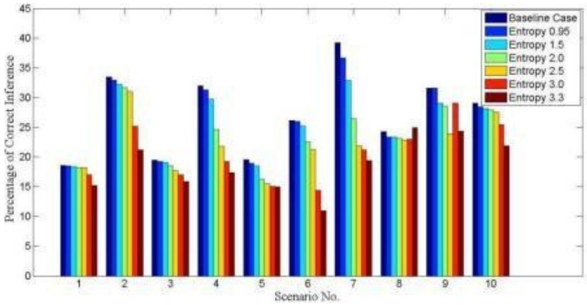

Fig. 10. Baseline case vs. entropy (P_2)

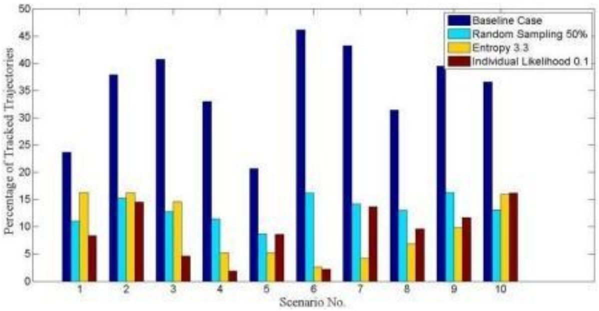

Fig. 11. Privacy performance of different datasets (P_1)

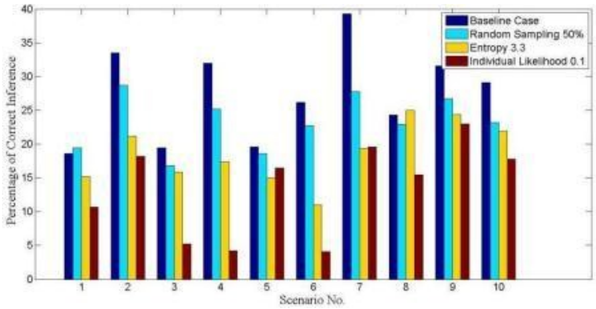

Fig. 12. Privacy performance of different datasets (P_2)

For case 2, here we list the results for 5 pairs of non-neighboring VTL zones. Table 5 shows that for the baseline case, the average percentage of tracked vehicles is about 12.6%, and among all the inferences, about 6.8% in average are correct. This implies that even though the adversary has access to the signal information and a relatively good traffic-knowledge-based travel time estimation model, it is still very difficult to successfully track vehicle trajectories from non-neighboring VTL zones.

Similar experiments have been done using random sampling datasets, individual likelihood datasets and entropy datasets. The results and implications generally match with

Table 5. Privacy performance of the baseline dataset (Case 2)

No.	Z_1	Z_3	No. of Released Trajectories (Both Z_1 and Z_3)	No. of Released Trajectories(Z_1)	No. of Released Trajectories (Z_3)	No. of Inferences	No. of Correct Inferences	Percentage of Tracked Trajectories	Percentage of Correct Inferences
1	554_555	559_560	197	586	622	495	39	19.8%	7.9%
2	556_557	562_563	417	1144	757	972	60	14.4%	6.2%
3	690_691_692	694_695_696	467	915	599	764	59	12.6%	7.7%
4	529_530	533_534_535	553	892	730	651	55	9.9%	8.4%
5	797_798	802_803_804	185	749	341	317	12	6.5%	3.8%

Table 6. Entropy vs. individual likelihood (convenience of traffic modeling)

No.	Intersection	No. of Released Trajectories		Success Rate for Queue Length Estimation			Winner
		3.3 level of confusion	0.1 individual likelihood	3.3 level of confusion	0.1 individual likelihood	Baseline	
1	597_598_599	1492/1496	1481/1496	79.6%	79.6%	81.8%	Entropy
2	601_602	709/1225	440/1225	64.4%	57.8%	75.6%	Entropy
3	554_555	570/586	519/586	96.4%	96.4%	96.4%	Entropy
4	556_557	778/1144	870/1144	64.8%	64.8%	70.1%	Individual likelihood
5	559_560	461/622	343/622	79.0%	84.2%	92.1%	Individual likelihood
6	562_563	286/757	466/757	23.3%	31.4%	34.9%	Individual likelihood
7	612_613_614	496/969	560/969	65.2%	69.6%	91.3%	Individual likelihood
8	616_617	574/1112	673/1112	76.7%	83.7%	88.4%	Individual likelihood
9	658_659	385/999	468/999	40.2%	45.7%	66.3%	Individual likelihood
10	660_661_662	364/642	413/642	89.3%	89.3%	92.9%	Individual likelihood
11	780_781_782	345/475	334/475	96.9%	96.9%	96.9%	Entropy
12	784_785	550/923	709/923	64.9%	71.9%	76.2%	Individual likelihood
13	690_691_692	505/915	669/915	69.2%	71.2%	71.2%	Individual likelihood
14	692_693_694	276/672	307/672	57.1%	64.3%	80.1%	Individual likelihood
15	694_695_696	363/599	401/599	36.9%	35.4%	38.5%	Entropy
16	529_530	583/892	750/892	58.2%	59.1%	59.1%	Individual likelihood
17	531_532_533	478/765	480/765	57.0%	55.9%	66.7%	Entropy

those of Case 1. It is worthy to mention that the level of privacy of baseline dataset is already quite high. Even though further implementation of other privacy algorithm can improve the performance, not too much improvement can be obtained. For example, for 0.1 individual likelihood dataset, the value of P_1 is 9.1% in average; and the value of $P_2$is 7.8%, which do not improve much from the baseline dataset.

C. *Evaluation in Terms of Modeling Convenience*

On top of an acceptable level of privacy, the modeling convenience of a released dataset should also be evaluated. Fig. 13 indicates the vehicle trajectories for the baseline dataset and Fig. 14 indicates the vehicle trajectories for the 0.1 individual likelihood dataset. As we can tell, in order to protect privacy, some of the trajectories have been filtered out in Fig. 14. And we want to test if the remaining trajectories are still enough for queue length estimation [13].

In particular, we compared 0.1 individual likelihood dataset and 3.3 level of confusion dataset, and the results are shown in Table 6. Notice that both datasets release almost all the trajectories at the first intersection. The reason for this is that this intersection is the very first intersection which is close to the entrance of the road network, thus both algorithms pretty much release all the trajectories at this intersection. For the cases where success rates are relatively low (e.g., about 30% for intersection 6 and about 37% for intersection 15), that is because some of the cycles for those intersections are uncongested, thus there is insufficient data (even for the baseline case) to support queue length estimation. There are also some cases in which one dataset releases more trajectories but ends up having a smaller success rate (e.g., intersection 5, intersection 15 and intersection 17). That is because the algorithm is releasing more samples in the cycles which already have many samples, but is not releasing enough samples in some cycles which have little samples. Compared with the baseline case,

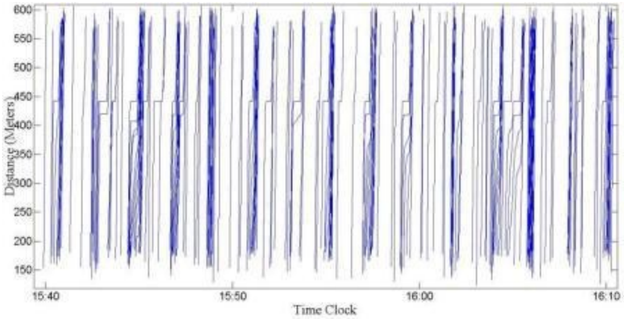

Fig. 13. Released vehicle trajectories (baseline case, 556_557)

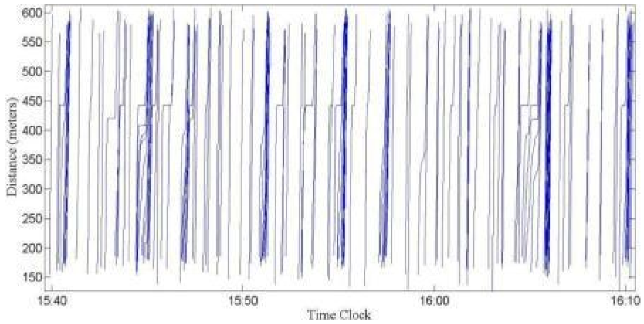

Fig. 14. Released vehicle trajectories (0.1 individual likelihood, 556_557)

the success rate of both entropy dataset and individual likelihood dataset do not decrease much, implying that by filtering out some sample trajectories (to protect privacy), the remaining samples can still be applied for traffic applications. Moreover, one can find that in 11 out 17 cases, the 0.1 individual likelihood dataset has better performance than the entropy dataset with 3.3 level of confusion, which suggests that the individual likelihood datasets are more preferable in terms of the convenience for traffic applications.

VI. CONCLUSIONS

In this paper, we developed traffic-knowledge-based adversary models to link vehicle trajectories between two different VTL zones. By applying this to different datasets, we can evaluate the performance of different privacy algorithms in terms of both privacy protection and the convenience for traffic modeling. It is found that the idea of only releasing trajectory data within VTL zones helps to protect privacy. And by comparing the performance of different privacy algorithms, we can conclude that privacy algorithms, especially those based on the metrics of individual likelihood and entropy, can enhance the level of privacy. Meanwhile, the released datasets of these algorithms can still be applied to traffic applications with satisfactory performances.

ACKNOWLEDGMENT

This research is supported by the National Science Foundation grant CMMI-1031452.

REFERENCES

[1] A. Beresford and F. Stajano, “Mix zones: user privacy in location-aware services,” in proceedings of the second IEEE Annual Conference on Pervasive Computing and Communication Workshops, 2004.

[2] B. Gedik and L. Liu, “Location privacy in mobile systems: A personalized anonymization model,” in distributed Computing Systems, 2005, ICDCS 2005, in proceedings. 25th IEEE International Conference on, 2005, pp. 620 –629.

[3] B. Hoh and M. Gruteser, “Protecting location privacy through path confusion,” in proceedings of IEEE/Create-Net SecureComm, Athens, Greece, September 2005.

[4] B. Hoh and M. Gruteser, “Preserving privacy in GPS traces via uncertainty-aware path cloaking,” in proceedings of ACM CCS 2007, 2007.

[5] B. Hoh, M. Gruteser, R. Herring, J. Ban, D. Work, J.-C. Herrera, A. M. Bayen, M. Annavaram, and Q. Jacobson, “Virtual trip lines for distributed privacy-preserving traffic monitoring,” in proceeding of the 6th international conference on Mobile systems, applications, and services, ser. MobiSys ’08. New York, NY, USA: ACM, 2008, pp. 5–28. [Online]. Available: http://doi.acm.org/10.1145/1378600.1378604

[6] B. Zan, P. Hao, M. Gruteser, X., Ban, “VTL zone-based path cloaking algorithm,” submitted to IEEE ITS Conference, 2011.

[7] F. Dotzer, “Privacy issues in vehicular ad hoc networks,” in proceeding of the 2nd ACM international workshop on Vehicular ad hoc networks. ACM Press, 2005.

[8] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, “Random data perturbation techniques and privacy preserving data mining,” in IEEE ICDM. IEEE Press, 2003.

[9] H. Liu and S. Jabari, “Evaluation of corridor traffic management and planning strategies using microsimulation: a case study,” Transportation Research Record, 2088, 2008, pp. 26-35.

[10] K. P. Tang, P. Keyani, J. Fogarty, and J. I. Hong, “Putting people in their place: an anonymous and privacy-sensitive approach to collecting sensed data in location-based applications,” in proceedings of CHI ’06, 2006, pp. 93–102.

[11] M. Gruteser, and D. Grunwald, “Anonymous usage of location-based services through spatial and temporal cloaking,” in proceedings of the First International Conference on Mobile Systems, Applications, and Services, 2003.

[12] S. Rass, S. Fuchs and M. Schaffer, “How to protect privacy in floating car data systems,” in proceeding of the fifth ACM international workshop on VehiculAr Inter-NETworking, 2008.

[13] X. Ban, P. Hao, Z. Sun, “Real time queue length estimation for signalized intersections using sampled travel times,” Transportation Research, Part C, in press, 2011.

[14] X. Ban, R. Herring, P. Hao, and A. Bayen, “ Delay pattern estimation for signalized intersections using sampled travel times,” Transportation Research Record 2130, 2009, pp. 109-119.