Measuring Privacy in Vehicular Networks

Isabel Wagner
De Montfort University
Leicester, United Kingdom
Email: isabel.wagner@dmu.ac.uk

***Abstract*—Vehicular communication plays a key role in near-future automotive transport, promising features like increased traffic safety, flexible insurance premiums, or wireless software updates. However, vehicular communication can expose driver locations and thus poses important privacy risks. Many schemes have been proposed to protect privacy in vehicular communication, e.g. the use of short-term pseudonyms instead of long-term identifiers. The effectiveness of these schemes is usually shown using *privacy metrics*. However, to the best of our knowledge, (1) different privacy metrics have never been compared to each other, and (2) it is unknown how strong the metrics are. In this paper, we argue that privacy metrics should be monotonic, i.e. that they indicate decreasing privacy for increasing adversary strength, and we evaluate the monotonicity of 37 privacy metrics on real and synthetic traffic with state-of-the-art adversary models. Our results indicate that (1) the strongest metrics are the expected estimation error and normalized entropy, and (2) most privacy metrics are weak at least in some situations. We therefore recommend to use *metrics suites*, i.e. combinations of privacy metrics, when evaluating new privacy-enhancing technologies.**

I. Introduction

Vehicular wireless communication technologies like dedicated short-range communications (DSRC) enable vehicles to communicate with other vehicles and infrastructure nodes to enable features like intersection collision avoidance, electronic road pricing, or cooperative adaptive cruise control. To realize these features, vehicles transmit sensitive data – often without encryption – for example their location, speed, and heading. This information can be used by anybody within wireless transmission range to track vehicles and their drivers on a large scale [5]. These privacy issues are well recognized, and many schemes have been proposed to protect privacy. For example, vehicles are often assumed to have pools of pseudonyms in addition to a long-term identifier, and different schemes have been proposed to switch between pseudonyms in a privacy-preserving way without compromising safety and accountability [9]. To evaluate how effectively these schemes protect privacy, researchers use various privacy metrics.

Because privacy is difficult to quantify, privacy metrics focus on quantities that are related to privacy, for example the number of vehicles that an adversary cannot distinguish, or the probability that an adversary can track a vehicle successfully. Many such metrics have been proposed, and researchers usually select one or two metrics to evaluate a new scheme.

However, there is a lack of research into the metrics themselves. In particular, we are not aware of research that compares different privacy metrics, or that analyzes how strong different privacy metrics are. This is an important open problem because the accurate measurement of privacy is essential to evaluate new privacy protections.

Contributions. In this paper, we make three contributions to research into privacy metrics: (1) We define monotonicity as a criterion for the strength of privacy metrics, i.e. we argue that a privacy metric should indicate low privacy for strong adversaries, and high privacy for weak adversaries. (2) We describe a methodology how monotonicity can be evaluated systematically and (3) we apply this methodology to study 37 privacy metrics proposed in the literature, including metrics that have not been used in vehicular privacy before.

II. Related Work

Privacy metrics for vehicular communications. Several privacy metrics have already been used in the vehicular communications field, most notably the anonymity set size, entropy, the adversary's success rate, and the maximum tracking time [11]. Many more privacy metrics have been proposed in other fields [13], but it is unclear whether they are suitable for vehicular privacy. In this paper, we analyze the strength of 37 privacy metrics from vehicular privacy and other fields.

Evaluation of metrics. To the best of our knowledge, there is no literature that compares or evaluates privacy metrics for vehicular privacy. We have previously studied privacy metrics for genomic privacy and proposed a methodology to systematically evaluate the strength of privacy metrics [12].

III. Methodology

To evaluate how good privacy metrics are, i.e. the *strength* of privacy metrics, we adapt the methodology we first introduced for genomic privacy [12]. Our method provides a controlled environment in which to experiment with privacy metrics by abstracting from many of the factors that affect privacy in the real world. For example, we assume that precise and timely position updates are available for all cars – a best-case scenario from the adversary's viewpoint – instead of considering network-level packet delays or losses.

Four components are needed to apply the method: (1) to model user driving behavior, we use real-world traffic traces, (2) to model the adversary, we implement a state-of-the-art tracking algorithm, (3) to show how strong privacy metrics are when applied to a specific combination of adversary and user actions, we define monotonicity as a strength indicator, and (4) we use 37 privacy metrics from the privacy literature. We implemented all four components in Python using the packages numpy, scipy, scikit-learn, and mpi4py [4].

TABLE I
TRAFFIC CHARACTERISTICS FOR TIME/DAY COMBINATIONS

City	Day	Time	Duration	Vehicles	Data points	Veh/km^2
Rome	Mon	1pm	1800s	182	68k	0.16
Rome	Tue	5pm	1800s	131	45k	0.18
Rome	Wed	10am	1800s	139	48k	0.47
Rome	Fri	8am	1800s	54	8k	0.04
Madrid	Mon	11am	1000s	1597	1.3m	160
Madrid	Tue	8am	1000s	2168	1.6m	217
Madrid	Wed	11am	1000s	2215	2.0m	222
Madrid	Fri	8am	1000s	2495	1.9m	250

A. Traffic Traces

The user actions in a vehicular network consist of traffic traces that determine the locations and movement patterns of vehicles, and thus the characteristics of the network traffic the adversary can observe. For a meaningful evaluation of privacy metrics, these traffic traces need to fulfill four requirements: they should (1) be realistic, (2) represent varied environments, (3) offer fine-grained position updates in the order of 1Hz, and (4) consist of timestamps and positions (velocities can be approximated from subsequent positions). We use two sets of traffic traces, inner city traffic and highway traffic, to fulfill these requirements.

For *inner city traffic*, we use taxi traces recorded in Rome available from CRAWDAD [1], [3]. The dataset consists of one month of measurements for 320 taxis, with one data point every 7 seconds. Positions where the GPS precision was below 20m were filtered from the dataset. To offset these missing data points, we reduced the granularity to 15 seconds.

For *highway traffic*, we use synthetic traffic from highways near Madrid [7], [8]. The traffic characteristics were modeled after empirical real-world traffic conditions on three highways. The dataset provides traffic on three 10km stretches of highway: one day of traffic on the four-lane M30, and eight 30-minute intervals on each M40 and A6 (both three lanes). Measurements are available every 500ms.

Previous work showed that the characteristics of vehicular network graphs differ depending on times of day and days of the week [8]. To study whether privacy metrics are similarly affected, we selected different combinations of times of day and days of the week from each dataset (see Table I).

B. Adversary Models

The adversary in vehicular communications is often assumed to be a passive observer who aims to track vehicles. To evaluate the strength of privacy metrics, the adversary model needs to (1) represent a realistic and strong adversary, and (2) be adjustable to model adversaries of different strengths.

Tracking algorithm. To fulfill the requirement for a realistic and strong adversary, we implemented a state-of-the-art vehicle tracking algorithm, the joint probabilistic data association filter (JPDA) (which is sometimes referred to as a multiple hypothesis tracker (MHT) with zero-scan, i.e. a MHT that maintains only one hypothesis [14]). Originally described for radar tracking applications [2], [10], JPDA has since been applied successfully to vehicle tracking [6], [14]. The JPDA algorithm maintains a list of tracks, each representing one vehicle. Periodically, a set of new observations arrives, and the tracker attempts to find the best continuations for all tracks, based on the positions and velocities of existing tracks and observations, but not on vehicle identities. JPDA is based on Kalman filtering, extended by the capability to resolve situations where the association between existing tracks and new observations is not unique.

Our implementation of JPDA closely follows the description in [2], with inspiration for the definition of the state vector and covariance matrices taken from [6], [14]. Here, we briefly describe our choice of parameters. The state vector x for each track consists of 2D positions and velocities, and the state transition matrix A models the vehicle dynamics according to which the state evolves from one time step to the next (1). The tracker is subject to two kinds of noise: *process noise* represents uncertainty in state estimation due to random motion entering the system between observations, and *measurement noise* represents uncertainty in measurement. Both kinds of noise are assumed to be normally distributed white noise with covariances Q (process noise) and R (measurement noise) (2).

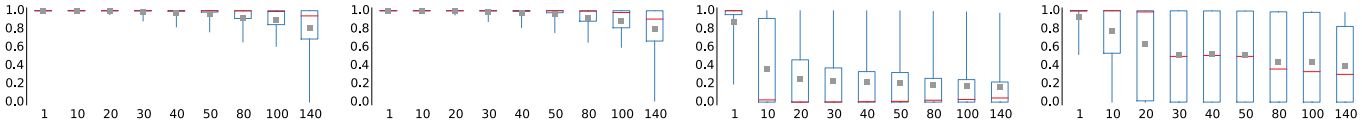
$$x = \begin{pmatrix} p_x \\ v_x \\ p_y \\ v_y \end{pmatrix}, A = \begin{pmatrix} 1 & t & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & t \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (1)$$

$$Q = q \begin{pmatrix} t^3/3 & t^2/2 & 0 & 0 \\ t^2/2 & t & 0 & 0 \\ 0 & 0 & t^3/3 & t^2/2 \\ 0 & 0 & t^2/2 & t \end{pmatrix}, R = rI_4 \quad (2)$$

To speed up the computations, the tracker uses *gating* to eliminate unlikely associations, e.g. observations that are outside of the realistic travel distance for a vehicle. A frequently used gating condition is the ellipsoidal gate that demands that the norm of the residual vector is not larger than the gating threshold G [10]. Our tracker uses $G = 35$.

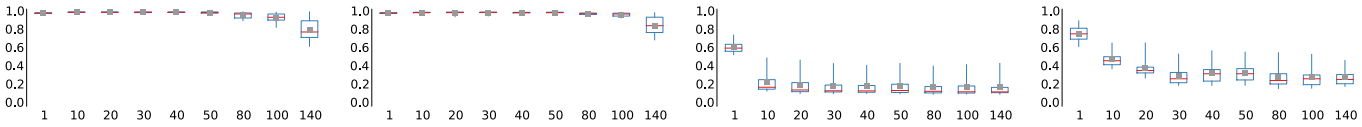
Ordered strength levels. To fulfill the requirement for adjustable adversary strengths, we adjusted the parameters for the JPDA tracker. Because tracker performance strongly depends on the values for the covariance matrices R and Q [14], and because variation of these values is easy to implement, we chose nine parameter levels for each r and q (Eq. (2) shows how they are related to R and Q), with $r = [1, 10, 20, 30, 40, 50, 80, 100, 140]$ and $q = 0.1r$.

To evaluate whether privacy metrics indicate high privacy for weak adversaries and low privacy for strong adversaries, the adversary strength levels need to be ordered. Figs. 1 and 2 show the probability the adversary assigns to the ground truth (i.e. the probability that a vehicle's track is continued with the correct observation) and the percentage of correctly identified vehicles at each time step, respectively, for four of the eight data slices (the other figures show similar results). Each figure shows one box for each of the nine adversary strength levels, summarizing all vehicles and all time steps in the data slice.



(a) Madrid, A6, Monday 11am (b) Madrid, M40, Wednesday 11am (c) Rome, Monday 1pm (d) Rome, Friday 8am

Fig. 1. Probability the adversary assigns to the ground truth, by adversary strength level, ordered strongest to weakest from left to right



(a) Madrid, A6, Monday 11am (b) Madrid, M40, Wednesday 11am (c) Rome, Monday 1pm (d) Rome, Friday 8am

Fig. 2. Percentage of correctly tracked vehicles, by adversary strength level, ordered strongest to weakest from left to right

The boxes indicates the upper and lower quartiles of this data, the median is shown with a red line, and the mean with a grey square. From the ends of the boxes, lines extend to the 5% and 95% quantiles. It can be seen that 1 is consistently the strongest adversary level and 140 the weakest. The adversary's strength between the strongest and weakest levels is decreasing in almost every case, with the exception of Friday morning in Rome (Fig. 2d). The reason for this is likely that the vehicle density in this case (54 unique vehicles) is much lower than in all other cases (130-180 unique vehicles). Overall, Figs. 1 and 2 confirm that our parameter levels for R and Q do indeed result in ordered levels of adversary strength.

Note that the adversary's strength is lower overall for Rome compared to Madrid. This is because highway traffic (Madrid) is much more regular and thus easier to track than inner-city traffic (Rome), and also because position updates are available in a much higher frequency for Madrid (500ms) compared to Rome (15s).

C. Criteria for Metric Strength

To determine how strong a privacy metric is, we use the criterion of *monotonicity*: with increasing adversary strength, metrics should indicate decreasing privacy values. We have previously proposed a formalization of the monotonicity criterion and an algorithm to compute monotonicity values [12]. In brief, the algorithm uses two statistical tests (Welch's t-test and the Wilcoxon rank-sum statistic) for each pair of successive adversary strength levels to determine whether the difference between mean metric values is statistically significant and points in the expected direction (positive for higher-better metrics, negative for lower-better metrics). Each outcome of each statistical test is then assigned points: $+1$ for a difference in the expected direction, -1 for a difference in the wrong direction, and -2 for a change in direction (a peak means that strong and weak adversaries cannot be distinguished and is thus not desirable). The total monotonicity score, i.e. the addition of these point values, can be visualized in heat maps (see Section IV-G).

In addition to evaluating monotonicity, we discuss how the values of each metric can be interpreted, and whether the metrics are suitable to compare privacy between scenarios with different traffic conditions.

D. Privacy Metrics

We study 37 privacy metrics that have been proposed in the literature. Many of these have not been applied to vehicular network privacy before. Our aim in including these metrics is to identify strong privacy metrics that are not yet common in the vehicular networking community, and thus to increase the selection of metrics available.

IV. Results

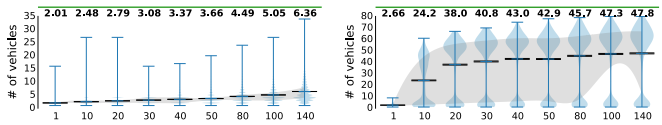
To compute the results on metric strength, we applied the JPDA tracker to 2000 time steps of the four Madrid data slices and 120 time steps of the four Rome data slices, and then evaluated each of the privacy metrics in all scenarios and time steps. In Sections IV-A to IV-F, we present detailed results for the metrics, including a brief description and a discussion how metric values can be interpreted and compared between scenarios. For full descriptions, equations, and references for each metric, we refer to [13]. Our presentation is grouped into categories of metrics according to which kind of output they measure (as proposed in [13]). Due to space limitations, we omit detailed results for some metrics and only compare two scenarios: Monday 11am in Madrid and Monday 1pm in Rome. We present the full set of results on monotonicity for all metrics and all scenarios in Section IV-G.

For each privacy metric and adversary strength level, we plot one vertical violin plot. The vertical bars show the range of the data, with horizontal lines indicating the minimum, mean, and maximum. Two thin horizontal lines show the confidence intervals for the mean. In addition, a kernel density plot on each side of the bar indicates the probability density. To illustrate the statistical distribution of the bulk of metric values, we added the first and third quartile to the plot. We fitted cubic splines to the quartiles and shade the area between them to emphasize how the quartiles change depending on adversary strength. In addition, we print the value of the mean in boldface at the top of each violin. The green bars indicate whether higher or lower values indicate higher privacy.

A. Uncertainty Metrics

The *anonymity set size* indicates how many vehicles the adversary cannot distinguish. We approximate the anonymity set size by counting the vehicles to which the tracker assigns a non-zero probability. Fig. 3 shows that the anonymity set

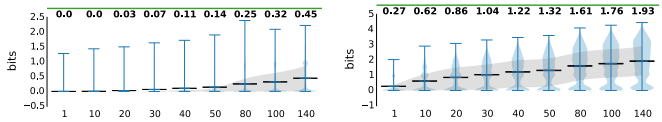
size in Madrid (highway traffic) is much lower than in Rome (inner-city taxi traffic), despite Rome having a much lower traffic density. This may caused by waiting areas where inner-city taxis wait for customers, and indicates that the anonymity set size is sensitive to differences in traffic characteristics.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 3. Anonymity set size

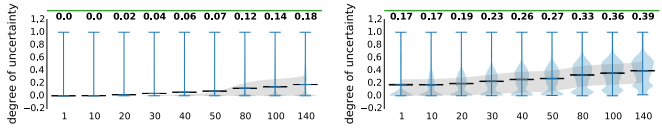
Entropy indicates the adversary's uncertainty as to which member of the anonymity set is the true target. Fig. 4 shows the value of entropy in bits, indicating that the uncertainty in Madrid is lower than in Rome. However, the absolute value of entropy is not easily comparable across scenarios because it is influenced by the size of the anonymity set.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 4. Entropy

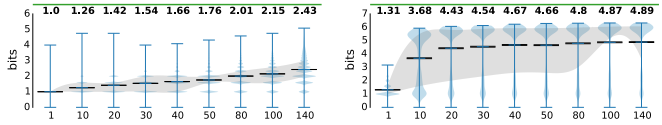
For this reason, *normalized entropy* uses the maximum amount of entropy in a scenario to normalize entropy values to a $[0, 1]$ range which indicates the adversary's degree of uncertainty. Fig. 5 shows that inner-city traffic has a higher degree of uncertainty than highway traffic. Note that the difference is much smaller than the differences in entropy or anonymity set size, which indicates that normalized entropy is better suited for comparisons between scenarios.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 5. Normalized entropy

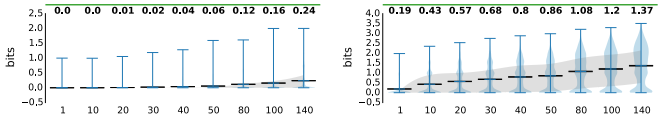
Some variations of entropy are based on Rényi entropy, a parameterized generalization of the entropy introduced above (where entropy has the parameter $\alpha = 1$). *Max-entropy* ($\alpha = 0$) indicates the maximum uncertainty the adversary can have when all members of the anonymity set are equally likely. Max-entropy represents an upper limit on privacy, and Fig. 6 shows that its behavior is similar to the anonymity set size.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 6. Max-entropy

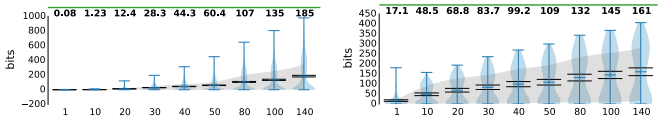
In contrast, *min-entropy* ($\alpha = \infty$, Fig. 7) focuses on the target for which the adversary has the largest probability and thus indicates a lower limit on privacy. *Collision entropy* ($\alpha = 2$, not shown) indicates the uncertainty that two independent samples from the adversary's estimate are the same. Both min-entropy and collision entropy behave similarly to entropy, but indicate lower uncertainty values.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 7. Min-entropy

Cumulative entropy assumes that subsequent independent applications of privacy mechanisms (e.g., mix zones) increase the privacy of vehicles. Cumulative entropy thus adds up the entropies from each application. In this study, we assume a privacy mechanism is applied at every time step, and Fig. 8 shows the result at the last time step. The metric values within each scenario behave as expected (increasing values with decreasing adversary strength), but are hard to compare between scenarios if the scenarios apply privacy mechanisms a different number of times.

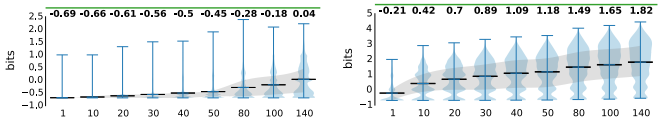


(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 8. Cumulative entropy

Because entropy is strongly influenced by low-probability outliers, *quantiles on entropy* computes entropy based on only those parts of the adversary's estimated probability distrubtion that are above a certain quantile. Our results, based on the 5% quantile, are very similar to the results for entropy.

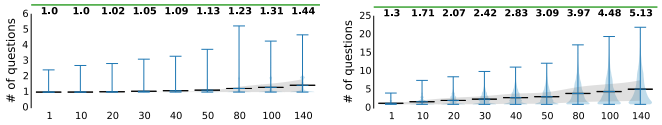
Conditional entropy (Fig. 9) describes how much information (in bits) is needed to describe the ground truth, i.e. the true mapping between observations and existing tracks, given that the adversary's estimated probabilities are known. The results for conditional entropy are similar to the results for entropy.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 9. Conditional entropy

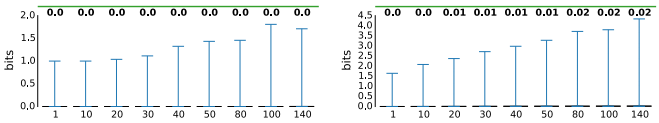
Inherent privacy (Fig. 10) and *conditional privacy* (not shown) are based on entropy and conditional entropy, respectively, and behave similarly to their base metrics. Instead of expressing the adversary's uncertainty in bits, these metrics indicate how many yes/no questions the adversary would have to ask to describe the ground truth.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 10. Inherent privacy

User-centric location privacy is based on entropy, but assumes that after the application of a privacy mechanism, a user's privacy decays over time with a user-specified rate l. We assume a fresh application of a privacy mechanism whenever a vehicles anonymity set size is greater than 1. Fig. 11 shows that the results for $l = 0.9$ are almost always zero, and thus does not allow to distinguish strong and weak adversaries in most cases (the results for other parameter settings are similar, see Fig. 33).

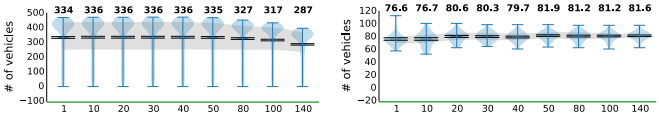


(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 11. User-centric location privacy with $l = 0.9$

B. Information Gain/Loss Metrics

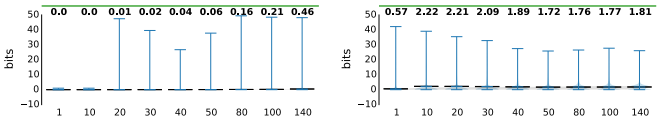
The *amount of leaked information* (Fig. 12) indicates how many vehicles the adversary was able to track correctly. The values depend strongly on the total number of vehicles. In Madrid, weaker adversaries track fewer vehicles than stronger adversaries (as expected), however, the metric gives opposite results in Rome, where weaker adversaries track more vehicles correctly.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 12. Amount of leaked information

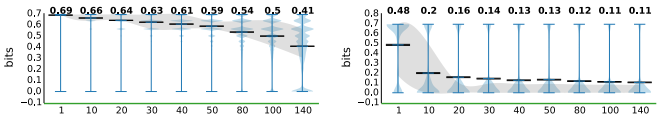
Relative entropy measures the distance between between the adversary's estimated probabilities and the ground truth and expresses the amount of information the adversary loses by approximating the ground truth with estimated probabilities. Fig. 13 shows that the metric behaves as expected for Madrid, indicating higher values for weaker adversaries. However, the behavior is reversed in Rome.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 13. Relative entropy

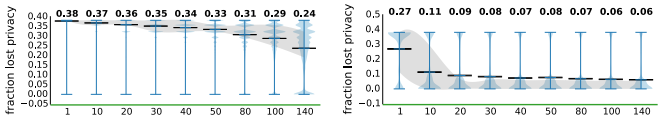
Mutual information indicates the amount of information that is shared between the distribution of the adversary's estimate and the ground truth. Fig. 14 shows that the amount decreases with decreasing adversary strength, but not all adversary strengths in the Rome scenario can be distinguished.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 14. Mutual information

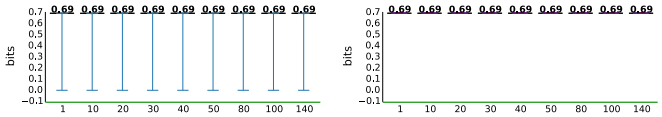
Conditional privacy loss is based on mutual information and measures the fraction of privacy lost through the adversary's estimate. As expected, the privacy loss decreases with decreasing adversary strength. However, the privacy loss in Rome seems very small compared to the privacy indicated by other metrics such as the adversary's success rate (Fig. 21) or the amount of information leaked (Fig. 12).



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 15. Conditional privacy loss

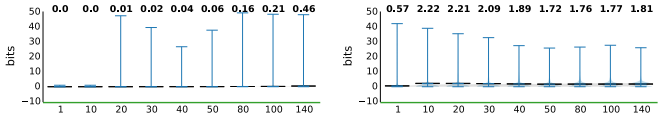
Loss of anonymity describes how much information can be learned from the adversary's estimate (in terms of mutual information) for the least private distribution of true outcomes. As Fig. 16 shows, this always corresponds to the maximum amount of mutual information across all adversary levels and thus does not allow to distinguish between adversaries of different strengths.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 16. Loss of anonymity

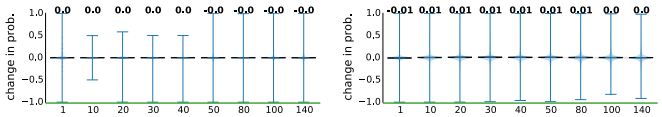
Information surprisal describes how much information is contained in the adversary's estimate of the true outcome, indicating how surprised the adversary would be upon learning the true outcome. Fig. 17 shows that the metric behaves as expected for Rome, but shows increasing instead of decreasing values for Madrid.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 17. Information surprisal

Increase in adversary's belief measures the difference between the adversary's estimates in the current and previous time step. As Fig. 18 shows, the differences are very small and can be both positive and negative. Therefore, this metric does not seem suitable to indicate privacy against a tracking adversary in vehicular communications.

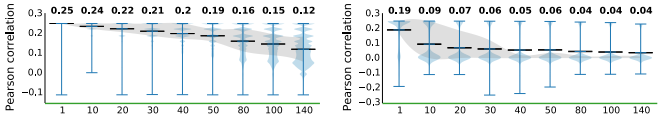


(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 18. Increase in adversary's belief

Pearson's correlation coefficient, a standard tool in statistics, measures the degree of linear dependence between the adversary's estimate and the ground truth, with a lower coefficient indicating higher privacy. Fig. 19 shows that the

correlation is decreasing with decreasing adversary strength, but with low correlation values throughout.

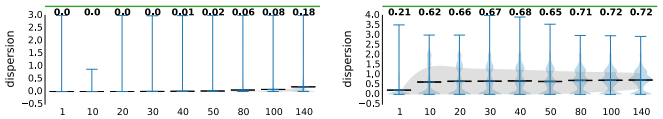


(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 19. Pearson's correlation coefficient

C. Data Similarity Metrics

Normalized variance measures the dispersion between the adversary's estimate and the ground truth. Fig. 20 shows good results for Madrid (increasing dispersion with decreasing adversary strength), but inconsistent dispersion values for Rome.

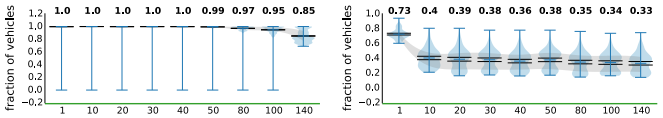


(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 20. Normalized variance

D. Adversary's Success Probability Metrics

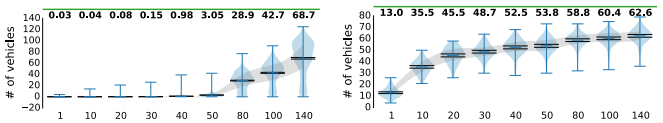
The *adversary's success rate* indicates the percentage of vehicles that were tracked correctly. The success rate (Fig. 21) shows decreasing privacy for increasing adversary strength in both scenarios. In Madrid, the metric indicates 100% success, i.e. zero privacy, for the five strongest adversary levels. This is in contrast to other metrics that do indicate some amount of privacy in these cases, e.g. cumulative entropy or the anonymity set size.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 21. Adversarys success rate

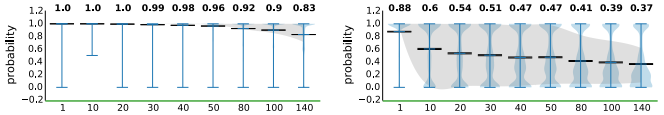
User-specified innocence indicates the number of vehicles for which the adversary's probability for the true outcome is below a user-specified threshold t, indicating how many vehicles are unlikely to be tracked. Fig. 22 shows user-specified innocence for a threshold of $t = 0.7$ (results for other thresholds are in Section IV-G). As expected, the values increase with adversary strength in both scenarios. However, the metric values are hard to interpret without knowing the total number of vehicles in a scenario.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 22. User specified innocence with $t = 0.7$

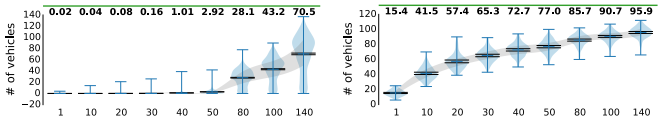
The *privacy breach level* indicates the adversary's probability for the true outcome. Fig. 23 shows that, as expected, the privacy breach level decreases with the adversary strength in both scenarios.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 23. Privacy breach level

The *hiding property* indicates the number of vehicles for which the highest probability in the adversary's estimate is below a threshold t. This can be interpreted as the number of vehicles for which the adversary is uncertain of how to continue their tracks. Fig. 24 shows that the values for a threshold of $t = 0.7$ increase with decreasing adversary strength (as expected). However, the values are hard to interpret without knowing the total number of vehicles in a scenario.

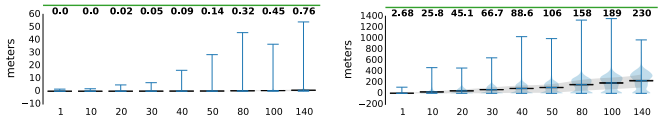


(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 24. Hiding property with $t = 0.7$

E. Error Metrics

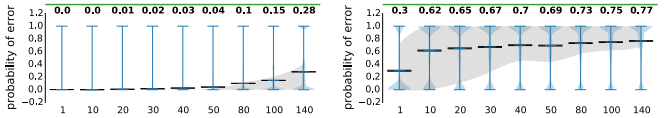
The *expected estimation error* measures the expected Euclidean distance (in meters) between the adversary's estimated location and the vehicle's true location. As expected, the distance error increases with decreasing adversary strength in both scenarios. The error is much smaller in Madrid which is likely caused by more regular traffic patterns on a highway and more frequent location updates.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 25. Expected estimation error

Incorrectness is a variant of the expected estimation error that replaces the Euclidean distance with a distance that indicates 0 for the correct guess and 1 otherwise. That is, incorrectness indicates the adversary's probability of error and thus mirrors the results for the adversary's success rate. As Fig. 26 shows, weaker adversaries have higher incorrectness in both scenarios.

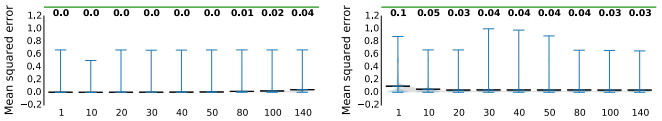


(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 26. Incorrectness

The *mean squared error* is often used in statistical parameter estimations and indicates the error between the adversary's estimated distribution and the true outcome. Fig. 27 shows that the error values are very small and almost indistinguishable in

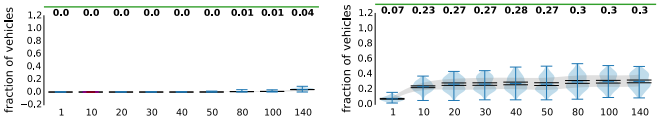
both scenarios. In addition, contrary to expectation, the metric indicates smaller errors for weaker adversaries in Rome.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 27. Mean squared error

Percentage incorrectly classified (Fig. 28) indicates the percentage of vehicles that were not tracked correctly. This is a normalized and inverted version of the amount of information leaked (see Fig. 12). Contrary to the earlier metric, however, normalization makes this metric more accurate in that weaker adversaries now consistently show a higher percentage of incorrect classifications.

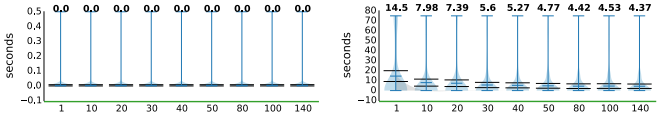


(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 28. Percentage incorrectly classified

F. Time Metrics

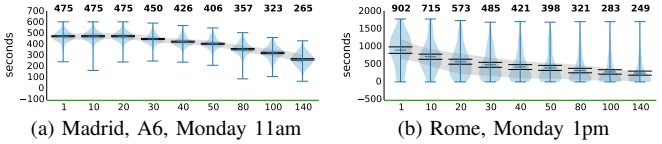
Maximum tracking time assumes that a vehicle is certain to be tracked if its anonymity set contains only the vehicle itself. The metric thus indicates the cumulative time (in seconds) during which a vehicle's anonymity set size is 1. Fig. 29 shows that this time is very small in both scenarios, indicating that the anonymity set size is rarely 1 (we have already seen that this does not impede the adversary's ability to track vehicles, see Fig. 21). Against an adversary who can track successfully even with larger anonymity sets, this metric does not seem to provide a useful indication of privacy.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 29. Maximum tracking time

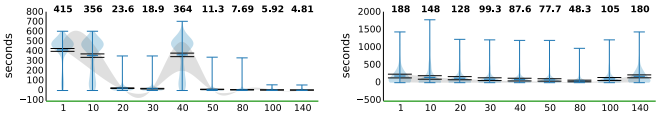
Time to confusion assumes that the adversary is confused when entropy is below a threshold h. The metric gives the cumulative time (in seconds) during which this is the case. Fig. 30 uses a confusion threshold of $h = 0.5$ (results for other thresholds are in Fig. 33) and shows that, as expected, weaker adversaries have a shorter time during which they are confused.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 30. Time to confusion with $h = 0.5$

Time to first confusion gives the first time when entropy drops below the threshold h. Fig. 31 shows that this metric gives much less consistent results than the cumulative time to confusion, most likely caused by brief drops of entropy early in the scenario.

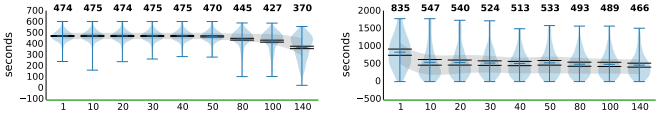


(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 31. Time to first confusion with $h = 0.5$

Distance to confusion indicates the cumulative distance traveled while the adversary was confused, and *distance to first confusion* indicates the distance until the first time of confusion. The results for these metrics mirror the results for the time to confusion metrics.

Mean tracking duration indicates how long, on average, each vehicle could be tracked correctly. Fig. 32 shows that the values for the two scenarios are of similar magnitude, even though the Rome scenario is almost twice as long as the Madrid scenario. In addition, the metric does not take into account that vehicles in the highway scenarios pass through the highway in a linear fashion, thus spending a limited amount of time, whereas vehicles in the inner-city scenario remain for a longer period. This metric thus seems suitable for within-scenario comparisons, but of limited value in comparing different scenarios.



(a) Madrid, A6, Monday 11am (b) Rome, Monday 1pm

Fig. 32. Mean tracking duration

G. Aggregated Results

Fig. 33 shows the full results for monotonicity for all metrics, all parameter levels, and all scenarios. Each square of the heatmap is determined by one set of results presented in detailed plots above. For example, the top-left square summarizes the results of Fig. 21a (adversary's success rate for Madrid, Monday 11am).

Influence of parameter settings. We evaluated how parameter settings influence the strength of the parameterized metrics. Fig. 33 shows that time/distance to confusion are stronger if the threshold is small (e.g., $h = 0.1$), whereas the strength of time/distance to first confusion and user-centric location privacy does not depend on the choice of threshold. Hiding property and user-specified innocence are stronger if the threshold is large (e.g., $t = 0.9$).

Influence of traffic conditions. Fig. 33 illustrates the influence of traffic conditions on the strength of privacy metrics. The heat map clearly shows that the strength of most metrics fluctuates between scenarios and traffic conditions (for example, time/distance to confusion, mean squared error, relative entropy). If these metrics are selected to evaluate a new privacy protection technology, it is neccessary to validate their strength for the specific scenario.

Metrics suites. To offset weaknesses in metrics, several metrics can be combined in a metrics suite that includes

strong metrics, metrics from each of Sections IV-A to IV-F, and metrics for between-scenario comparisons. An example metrics suite could thus consist of normalized entropy, mutual information, normalized variance, the adversary's success rate, the expected estimation error, and the time to confusion.

V. Conclusion

We have evaluated the strength of 37 privacy metrics for vehicular communications privacy in terms of their monotonicity and comparability between traffic conditions. We found that only few metrics are strong across all conditions, while most metrics are weak at least in some conditions. It is not immediately clear what causes these weaknesses, and therefore we have two recommendations how privacy metrics for the evaluation of new privacy protection technologies should be selected: (1) To evaluate the strength of the selected privacy metrics in each specific scenario before committing to their usage, and (2) to always use *metrics suites*, i.e. to combine several strong metrics from different categories.

Acknowledgment

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/P006752/1 and used the ARCHER UK National Supercomputing Service.

References

[1] R. Amici, M. Bonola, L. Bracciale, A. Rabuffi, P. Loreti, and G. Bianchi, "Performance Assessment of an Epidemic Protocol in VANET Using Real Traces," *Procedia Computer Science*, vol. 40, pp. 92–99, Jan. 2014.

[2] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*. Boston: Artech House Publishers, Jun. 1999.

[3] L. Bracciale, M. Bonola, P. Loreti, G. Bianchi, R. Amici, and A. Rabuffi, "CRAWDAD dataset roma/taxi (v.2014-07-17)," *CRAWDAD wireless network data archive*, Jul. 2014.

[4] L. Dalcín, R. Paz, M. Storti, and J. D'Elía, "MPI for Python: Performance improvements and MPI-2 extensions," *Journal of Parallel and Distributed Computing*, vol. 68, no. 5, pp. 655–662, May 2008.

[5] D. Eckhoff and C. Sommer, "Driving for Big Data? Privacy Concerns in Vehicular Networking," *IEEE Security & Privacy*, vol. 12, no. 1, pp. 77–79, 2014.

[6] K. Emara, W. Woerndl, and J. Schlichter, "On evaluation of location privacy preserving schemes for VANET safety applications," *Computer Communications*, vol. 63, pp. 11–23, Jun. 2015.

[7] M. Gramaglia, M. Fiore, and M. Calderon, "Measurement-Based Modeling of Interarrivals for the Simulation of Highway Vehicular Networks," *IEEE Communications Letters*, vol. 18, no. 12, pp. 2181–2184, Dec. 2014.

[8] M. Gramaglia, O. Trullols-Cruces, D. Naboulsi, M. Fiore, and M. Calderon, "Vehicular networks on two Madrid highways," in *2014 Eleventh Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, Jun. 2014, pp. 423–431.

[9] J. Petit, F. Schaub, M. Feiri, and F. Kargl, "Pseudonym Schemes in Vehicular Networks: A Survey," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 228–255, Firstquarter 2015.

[10] D. Reid, "An algorithm for tracking multiple targets," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 843–854, Dec. 1979.

[11] I. Wagner and D. Eckhoff, "Privacy assessment in vehicular networks using simulation," in *Winter Simulation Conference (WSC)*, Dec. 2014, pp. 3155–3166.

[12] I. Wagner, "Evaluating the Strength of Genomic Privacy Metrics," *ACM Trans. Priv. Secur.*, vol. 20, no. 1, pp. 2:1–2:34, Jan. 2017.

[13] I. Wagner and D. Eckhoff, "Technical Privacy Metrics: A Systematic Survey," *arXiv:1512.00327 [cs, math]*, Dec. 2015.

[14] B. Wiedersheim, Z. Ma, F. Kargl, and P. Papadimitratos, "Privacy in Inter-vehicular Networks: Why Simple Pseudonym Change is Not Enough," in *Proc. 7th International Conference on Wireless On-Demand Network Systems and Services (WONS)*, 2010, pp. 176–183.

Fig. 33. Strength of privacy metrics shown in a heat map. The colors indicate the strength of the metric (from blue=strong to yellow=weak).