On the Strength of Privacy Metrics for Vehicular Communication

Yuchen Zhao and Isabel Wagner, *Member, IEEE*

Abstract—Vehicular communication plays a key role in near-future automotive transport, promising features such as increased traffic safety and wireless software updates. However, vehicular communication can expose drivers' locations and thus poses privacy risks. Many schemes have been proposed to protect privacy in vehicular communication, and their effectiveness is usually evaluated with *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 evaluate and compare the strength of 41 privacy metrics in terms of four novel criteria: Privacy metrics should be monotonic, i.e., indicate decreasing privacy for increasing adversary strength; their values should be spread evenly over a large value range to support within-scenario comparability; and they should share a large portion of their value range between traffic conditions to support between-scenario comparability. We evaluate all four criteria on real and synthetic traffic with state-of-the-art adversary models and create a ranking of privacy metrics. Our results indicate that no single metric dominates across all criteria and traffic conditions. We therefore recommend to use *metrics suites*, i.e., combinations of privacy metrics, when evaluating new privacy-enhancing technologies.

Index Terms—Privacy metrics, vehicular communications, vehicular networks, privacy, monotonicity, privacy-enhancing technologies

1 Introduction

Vehicular communication technologies allow vehicles to communicate with other vehicles and infrastructure nodes to enable features such as intersection collision avoidance and 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, which raises privacy concerns [1]. These privacy issues are well recognized, and many approaches have been proposed to protect privacy. For example, vehicles are often assumed to have a pool of pseudonyms in addition to a long-term identifier, and different schemes have been proposed to change pseudonyms in a privacy-preserving way without compromising safety and accountability [2]. Privacy metrics quantify how effectively these schemes protect privacy.

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 privacy metrics or analyzes how strong privacy metrics are. Strong privacy metrics are important to ensure an accurate and consistent measurement of privacy, which is essential to evaluate new privacy protection schemes.

Contributions. In this paper, we make two contributions to research on privacy in vehicular networks.

First, we contribute to the methodological foundations of privacy measurement by proposing a method to evaluate the strength of privacy metrics using four novel criteria:

- *Monotonicity* requires that metrics show decreasing privacy with increasing adversary strength. This prevents misjudging the effectiveness of new privacy-enhancing technologies (PET).
- *Extent* requires that metric values are spread over a large value range, and *evenness* requires that metric values are distributed uniformly.
- Together, extent and evenness support fine-grained privacy analysis *within a scenario*, e.g., between vehicles, over time, and between parts of a city, as well as visualization of privacy levels.
- *Shared value range* requires that metric values share a common value range when applied in different traffic conditions. This allows for comparisons *between scenarios*.

Second, we evaluate the strength of 41 privacy metrics for vehicular networks, rank the metrics according to their scores in the four criteria, and make specific recommendations for the use of privacy metrics in vehicular networks. In particular, our key findings and recommendations are:

- No single metric excels in all four criteria, and the strength of many metrics varies between traffic conditions. We therefore recommend to always use metrics suites that combine the strengths of different metrics.

- *The authors are with the Cyber Security Centre, De Montfort University, Leicester LE1 9BH, United Kingdom.*
  E-mail: {yuchen.zhao, isabel.wagner}@dmu.ac.uk.

Manuscript received 7 Dec. 2017; revised 14 Mar. 2018; accepted 17 Apr. 2018. Date of publication 3 May 2018; date of current version 7 Jan. 2019. (Corresponding author: Isabel Wagner.)
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TMC.2018.2830359

- There are significant weaknesses in some metrics that have been used to evaluate pseudonym-changing schemes in the past, for example the *mean tracking duration*, *time/distance to confusion*, and *maximum tracking time*. We therefore recommend to use these metrics with caution, if at all.

Our contributions advance the state of the art in privacy measurement and are of particular use to researchers who design privacy protections for vehicular networks and use privacy metrics to evaluate their systems.

2 Related Work

In this paper, we draw on related work on privacy and privacy metrics in vehicular communications, privacy metrics in other fields, and research on evaluating the strength of privacy metrics.

2.1 Privacy Metrics for Vehicular Networks

In the past 15 years, many different privacy metrics have been proposed to evaluate the effectiveness of new PETs [3]. In the vehicular networking context, privacy metrics have been used, for example, to evaluate new pseudonym-changing strategies. These strategies determine how and how often vehicles change their public broadcast identifiers to reduce the likelihood that an adversary can track them. Proposed strategies include silent periods [4], pseudonym swapping [5], and mix zones [6], and in each case the privacy provided by each strategy was evaluated with a privacy metric: maximum tracking time [4], entropy [5], and the adversary's success rate [6], respectively.

Despite the large variety of privacy metrics, there is no consensus in the community as to which privacy metrics should be used [7]. For example, Wasef and Shen [8] use anonymity set size to quantify location privacy, whereas Eckhoff et al. [5] use entropy to offset the weaknesses of anonymity set size. Shokri et al. [9] argue that individual metrics are insufficient to quantify location privacy and combine confidence intervals, entropy, and incorrectness. Many other privacy metrics have been used, including cumulative entropy [6] and the mean time to confusion [10]. Although some of these papers argue for or against certain privacy metrics, they do not evaluate the existing privacy metrics in a uniform scenario and against a formal set of criteria.

In this paper, we evaluate all 14 metrics that, to the best of our knowledge, have already been used in vehicular communications (see Table 2 in the next section). In addition, we evaluate 21 metrics from the wider privacy literature. These metrics have not been used in vehicular communications before, but they can be calculated because their computations use data that is available in vehicular communications scenarios. We also evaluate variations of 6 metrics, either to offset their weaknesses, for example normalized versions of the *hiding property* and *user-specified innocence*, or to explore alternate definitions, for example a version of the *maximum tracking time* that is based on the adversary's success rate instead of the anonymity set size.

Our work in this paper contributes to finding a set of consensus metrics by presenting a comprehensive evaluation and ranking of a large number of privacy metrics.

2.2 Criteria for Privacy Metrics

Many authors have proposed criteria that good privacy metrics should fulfill. For example, they should be understandable and indicate the adversary's chances of success [11]; they should show both the level of privacy and the potential for privacy violations [12]; they should integrate accuracy, uncertainty, and correctness as three components of the adversary's success [9]; and they should quantify the amount of resources an adversary needs to succeed [13].

These criteria can serve as a checklist of what a privacy metric should fulfill. However, they are not suitable to evaluate how well a privacy metric addresses each criterion, especially when comparing privacy metrics to each other. To address this issue, in previous work we have proposed the criterion of monotonicity to evaluate the strength of privacy metrics [14], [15].

In this paper, we propose three novel criteria in addition to monotonicity to evaluate the strength of privacy metrics for vehicular networks.

2.3 Evaluation of Privacy Metrics

When evaluating new PETs, it is important to select strong privacy metrics because weak privacy metrics may overestimate privacy and result in real-world privacy violations. However, despite the large number of privacy metrics and the existence of criteria for privacy metrics, we are not aware of systematic efforts to evaluate the strength of privacy metrics for vehicular communications. The most closely related work in this respect is Murdoch's evaluation of metrics for anonymous communication [16].

In our own previous work, we presented a method for the evaluation of privacy metrics in genomic privacy [15] and a preliminary adaptation of this method to vehicular privacy [14]. Our initial method was based on the idea that privacy metrics should be monotonic, and that we can systematically evaluate their monotonicity using appropriately defined models for user and adversary behavior. In this paper, we define these user and adversary models for vehicular communications, expand the set of studied metrics to include metrics that are relevant for vehicular communications, and introduce three new criteria for metric strength: extent, evenness, and shared value range.

We thus close the gap in knowledge about the strength of privacy metrics for vehicular networks by systematically evaluating the strength of 41 privacy metrics based on four formal criteria.

2.4 Privacy Visualization

The visualization of privacy can help privacy engineers design new privacy-enhancing technologies. For example, Reeder et al. [17] visualize privacy policies in an Expandable Grid and show that this interface can help privacy experts make decisions. In vehicular networks, privacy metrics are naturally associated with the locations of vehicles and can be visualized as a map overlay. However, the use of such visualizations of location privacy has not been investigated. Compared with existing work, we evaluate the conditions privacy metrics need to satisfy to produce good visualizations and briefly explore possible uses for such visualizations.

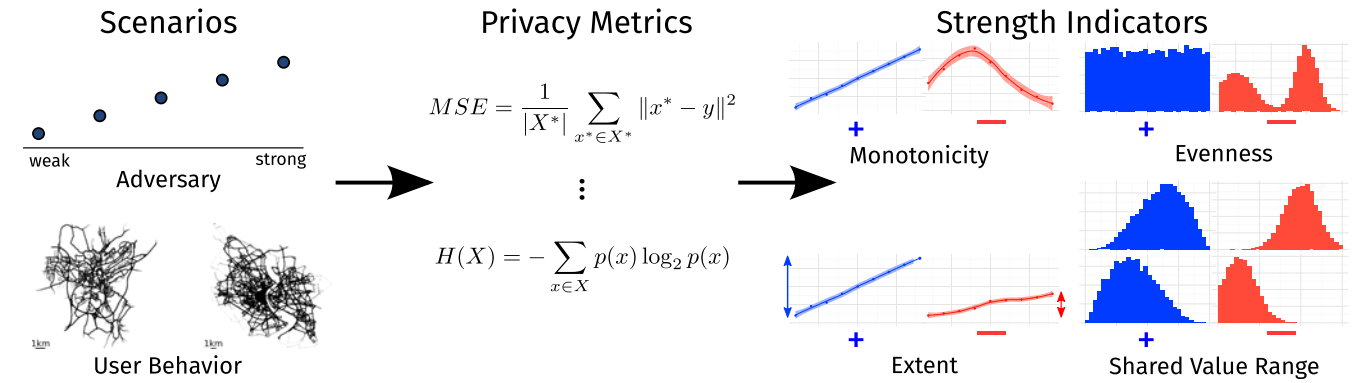


Fig. 1. Methodology to evaluate the strength of privacy metrics. (1) User behavior and adversary behavior are combined into scenarios. (2) Privacy metrics are applied to the scenarios. (3) The strength of privacy metrics in each scenario is evaluated with four strength indicators: monotonicity, extent of spread, evenness of spread, and shared value range.

3 Methodology

Our goal is to evaluate the strength of privacy metrics for vehicular networks. To do this, we adapt the method we first introduced for genomic privacy [15] to vehicular network privacy and introduce three new criteria that measure the strength of privacy metrics.

Assumptions. Our method provides a controlled environment 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. This ensures that the evaluation of metric strength is not influenced by network communication artifacts.

In addition, we abstract from the application of privacy-enhancing technologies because a strong PET has a similar observable effect on privacy metrics as a weak adversary (and, conversely, a weak PET "looks" the same as a strong adversary). In other words, the adversary's success and the user's privacy are two sides of the same coin [9], [18], and we focus on modeling the adversary's side.

Overview. To apply our methodology (see Fig. 1), we first define *scenarios* consisting of users and an adversary, where the adversary aims to infer user behavior. Second, we calculate the values of a range of *privacy metrics* in each scenario and finally we measure the strength of each privacy metric using four *strength indicators*: monotonicity, extent of spread, evenness of spread, and shared value range. We used open-source Python packages including NumPy [19], SciPy [19], scikit-learn [20], scikit-gof, and mpi4py [21] to implement our methodology.

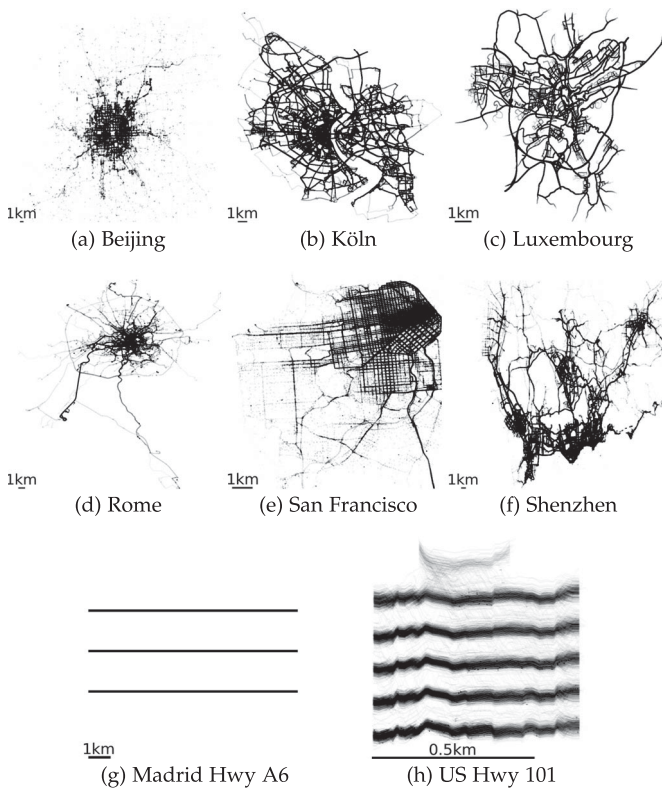


Fig. 2. Maps of the traffic traces used in our evaluation. Grayscale indicates the density of traffic (black=dense). The traffic data of Beijing, Rome, San Francisco, Shenzhen, and US Highway 101 are from real-world vehicles. The data of Madrid, Köln, and Luxembourg are synthetic. Note that the y axes for Madrid and US 101 are not to scale.

3.1 User Behavior: Real-World Traffic Traces

We model user behavior using spatio-temporal traffic traces. These traces of physical movement determine the characteristics of the network traffic the adversary can observe. We use eight sets of traffic traces, representing combinations of real and synthetic traffic as well as inner city traffic and highway traffic, to model realistic traffic in varied environments. Fig. 2 plots the coordinates of all vehicles at all time steps for each of the eight traffic traces.

For *inner city traffic*, we use taxi traces recorded in Rome [22], Beijing [23], and Shenzhen [23] as well as synthetic traffic based on measurements of real traffic in Köln [24] and Luxembourg [25]. The Köln traffic traces were generated by the microscopic mobility simulator SUMO, based on detailed travel and activity patterns collected by the German Federal Statistical Office. The Luxembourg traces were also generated by SUMO and are based on synthetic traffic demand that combines the real population demographics, road network, and traffic volume.

For *highway traffic*, we use real traffic from highway 101 near Los Angeles [26] and synthetic traffic from highway A6 near Madrid. The synthetic traffic is based on real-world traffic counts and has been generated by a microscopic vehicular mobility simulator [27]. The resulting traffic traces represent unidirectional, free flowing highway traffic.

TABLE 1
Traffic Characteristics for Time/Day Combinations

City	Day	Time	Length	Granularity	Cars	Cars/km^2	Type	Road layout	Reference
Rome	Mon	1 pm	2700s	15s	182	0.16	taxi	city	[22]
Rome	Tue	5 pm	2700s	15s	131	0.18	taxi	city	[22]
Rome	Wed	10 am	2700s	15s	139	0.49	taxi	city	[22]
Rome	Fri	8 am	2700s	15s	54	0.04	taxi	city	[22]
Madrid	Mon	11 am	1000s	0.5s	1597	26620	synthetic	highway	[27]
Madrid	Tue	8 am	1000s	0.5s	2215	36921	synthetic	highway	[27]
Köln	weekday	11 am	600s	1s	17980	23.3	synthetic	city	[24]
Luxembourg	weekday	11 am	900s	1s	6167	39.9	synthetic	city	[25]
Shenzhen	Mon	2 pm	1000s	1s	10359	4.61	taxi	city	[23]
US 101	Wed	7:50 am	250s	1s	1993	135036	car	highway	[26]
US 101	Wed	8:05 am	220s	1s	1533	69431	car	highway	[26]
US 101	Wed	8:20 am	120s	1s	1298	56656	car	highway	[26]
Beijing	Mon	12 pm	900s	15s	7972	1.27	taxi	city	[23]
San Francisco	Tue	1 am	3600s	5s	406	2.85	taxi	city	[23]
San Francisco	Mon	8 am	3600s	5s	322	2.31	taxi	city	[23]

Because the characteristics of vehicular network graphs can depend on the time of day and day of the week [27], we selected different combinations of time slots and days from the full traffic traces where possible. Table 1 summarizes the characteristics of each dataset.

We note that scenarios with low traffic density, such as Rome and Beijing, can be used to approximate the situation during roll-out of a new vehicular networking technology, when the percentage of vehicles equipped with the new technology is still low.

3.2 Adversary Behavior

The adversary in vehicular communications is often assumed to be a passive observer who aims to track vehicles [5]. 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.

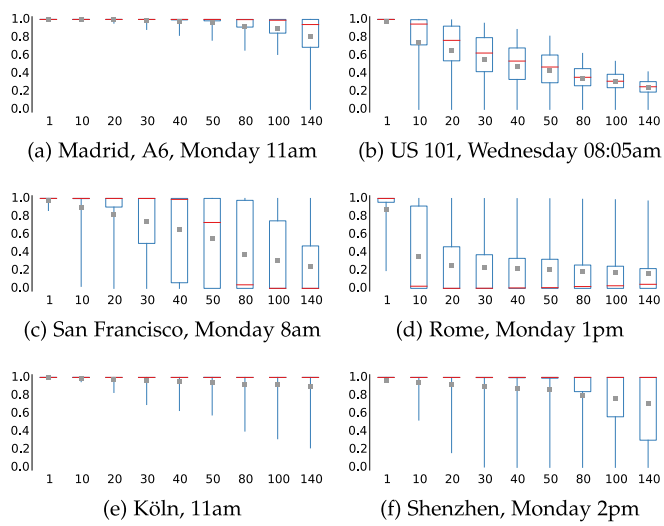


Fig. 3. Probability that a vehicle's track is continued with the correct observation, by adversary strength level. The subfigures show that an increase in process noise weakens the adversary regardless of the traffic condition.

Tracking Algorithm. To fulfill the requirement for a realistic and strong adversary, we implemented a state-of-the-art tracking algorithm, the joint probabilistic data association filter (JPDA) (also called multiple hypothesis tracker (MHT) with zero-scan [2]). Originally described for radar tracking [28], JPDA has already been applied to vehicle tracking [2], [29]. The JPDA algorithm maintains a list of tracks, each representing one vehicle. Whenever new observations arrive, the tracker computes the best continuations for all tracks, based only on positions and velocities of existing tracks and observations. JPDA uses Kalman filtering and can resolve non-unique associations between existing tracks and new observations. Our implementation of JPDA follows [28], with inspiration for the definition of the state vector and covariance matrices taken from [2], [29]. The tracker is subject to two kinds of noise: *process noise* that represents random motion in the system between observations, and *measurement noise* that represents uncertainty in measurement. JPDA assumes that both kinds of noise are normally distributed white noise with covariances Q (process noise) and R (measurement noise).

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 [2], we chose nine parameter levels for each r and q, with $r = [1, 10, 20, 30, 40, 50, 80, 100, 140]$ and $q = 0.1r$. To evaluate monotonicity, i.e., whether privacy metrics indicate high privacy for weak adversaries and low privacy for strong adversaries, these adversary strength levels need to be ordered. We illustrate this ordering in Fig. 3, which shows box plots of the probability that the adversary can continue a vehicle's track correctly in six different traffic conditions. In each plot, the boxes indicate the upper and lower quartiles and the median (red line), summarizing all vehicles and time steps for one adversary strength level. The plots also show the mean values (grey squares), and lines extend to the 5 and 95 percent quantiles.

Fig. 3 confirms that the nine parameter levels for r and q result in *ordered* levels of adversary strength, with 1

TABLE 2
Location Privacy Metrics Used in Our Evaluation

Category	Metric	per time	per vehicle	High/Low	Monotonicity	Extent	Evenness	Sh. Value Range
Uncertainty	Anonymity set size (ASS)*	✓	✓	H	+	o	o	–
	Collision entropy	✓	✓	H	++	o	o	+
	Conditional entropy	✓	✓	H	++	o	+	+
	Conditional privacy	✓	✓	H	++	–	o	o
	Cross entropy	✓	✓	H	+	–	–	+
	Cumulative entropy*		✓	H	+	o	o	o
	Entropy*	✓	✓	H	++	o	o	+
	Inherent privacy	✓	✓	H	++	–	o	o
	Max-entropy	✓	✓	H	+	o	+	o
	Min-entropy	✓	✓	H	++	o	o	+
	Normalized entropy	✓	✓	H	++	o	o	++
	Quantiles on entropy	✓	✓	H	++	o	o	+
	User-centric location privacy, l=0.1*	✓	✓	H	o	–	–	o
	User-centric location privacy, l=2*	✓	✓	H	o	–	–	+
Inform. gain/loss	**Amount of leaked inform.***	✓		L	o	o	+	–
	Conditional privacy loss	✓	✓	L	+	o	o	++
	Increase in adversary belief	✓	✓	L	o	–	+	++
	Information surprisal	✓	✓	L	–	–	–	+
	Loss of anonymity	✓		L	o	–	–	+
	Mutual information	✓	✓	L	+	o	o	++
	Pearson correlation	✓	✓	L	++	o	+	+
	Relative entropy	✓	✓	H	+	–	–	+
Error	**Expected distance error***		✓	H	++	–	++	–
	Expected distortion*	✓	✓	H	+	–	–	–
	Expected estimation error*	✓	✓	L	+	–	–	–
	Incorrectness*	✓	✓	H	++	+	o	++
	Mean squared error	✓	✓	H	o	–	–	+
	Perc. incorrectly classified*	✓		H	o	o	+	o
Sim	Normalized variance	✓	✓	H	+	o	o	+
Adv.'s success prob.	Adversary's success rate*	✓		L	o	o	+	+
	Hiding property, s=0.5	✓		H	+	o	o	–
	Norm. hiding property, s=0.5	✓		H	+	o	o	o
	Privacy breach level	✓	✓	L	++	o	o	++
	User-specified innocence, s=0.5	✓		H	+	o	+	–
	Norm. user-specified innocence, s=0.5	✓		H	+	o	+	o
Time	Distance to confusion, h=0.1		✓	L	o	o	o	o
	Distance to confusion, h=3		✓	L	–	o	o	o
	Dist. to first confusion, h=0.1		✓	L	o	o	–	o
	Max. tracking time (ASS=1)*		✓	L	o	–	–	–
	Max. tracking time (tracking success)		✓	L	–	o	o	o
	Mean tracking duration*		✓	L	o	o	+	o
	Time to confusion, h=0.1*		✓	L	o	o	+	o
	Time to confusion, h=3*		✓	L	o	o	+	o
	Time to first conf., h=0.1		✓	L	o	o	o	o

Metrics in bold are explained in Section 3.3. Starred metrics have previously been used in vehicular communications. H/L: high (H) or low (L) values indicate high privacy. Ratings for the four criteria are based on their average normalized scores: < 0.3*: –,* $\in [0.3, 0.7[$*: o,* $\in [0.7, 0.9[$*: +,* $\in [0.9, 1]$*: ++.*

consistently the strongest adversary level and 140 the weakest (adversary strengths for the other traffic conditions are ordered as well, see Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TMC.2018.2830359).

3.3 Privacy Metrics

We study 41 privacy metrics that have been proposed in the literature, both in the vehicular networking literature and the wider literature on privacy measurement in other application domains. Table 2 summarizes the metrics we have analyzed as well as our results for metric strength (we introduce the criteria in Section 3.4). The table also indicates whether the metrics evaluate privacy in each time step, for each vehicle, or both. For example, *entropy* generates values both per-time and per-vehicle, the *adversary's success rate* aggregates over all vehicles, i.e., it generates one value per-time, and the *mean tracking duration* aggregates over all time steps and thus generates values per-vehicle.

TABLE 3
Notation for All Privacy Metrics

$H(\cdot)$	Entropy
$I(\cdot;\cdot)$	Mutual information
$v \in V$	Set of vehicles the adversary aims to track
T	Total observation time
$X_{v,t}$	Adversary's estimated probabilities for vehicle v at time t
$Y_{v,t}$	Adversary's observed data (may be obfuscated)
$X^*_{v,t}$	True assignment of observations to vehicles
$d(x, x^*)$	Distance between the estimated and true location
h	Threshold for entropy
l	Rate of privacy decay
s	Threshold for adversary's probability

Due to the large number of metrics, we will not introduce all metrics in detail, but instead focus on the strongest metrics, according to our analysis in Sections 4 and 5. For equations and references for the remaining metrics, we refer to our previous work [3]. We present the metrics grouped by the output they measure, according to the taxonomy we introduced in [3]. Table 3 shows the notation used to formally describe the metrics. For clarity, we omit the indices for time steps and vehicles except where a metric is based on two or more time steps or aggregates over vehicles.

3.3.1 Uncertainty Metrics

Many metrics rely on the concept of the anonymity set, i.e., the set of vehicles V that the adversary cannot distinguish. In our evaluation, the anonymity set consists of all vehicles v to which the tracker assigns a non-zero probability. Most uncertainty metrics use variants of the *entropy* of the anonymity set [30] to quantify privacy, indicating how uncertain the adversary is about their estimate $p(x)$.

Rényi Entropy is a parameterized description of entropy. By adjusting the parameter α, several popular variants of entropy can be represented in terms of Rényi entropy, for example *Shannon entropy* ($\alpha = 1$) and *collision entropy* ($\alpha = 2$). *Min-entropy* ($\alpha = \infty$) focuses on the target for which the adversary has the highest probability and thus indicates a lower limit on privacy. *Max-entropy* ($\alpha = 0$) indicates the maximum uncertainty the adversary can have when all members of the anonymity set are equally likely and thus represents an upper limit on privacy.

$$priv_{RE} \equiv H_\alpha(X) = \frac{1}{1-\alpha} log_2 \sum_{x \in X} p(x)^\alpha.$$

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 distribution that are above a certain quantile (we used the 5 percent quantile in our evaluation).

Normalized Entropy uses *max-entropy* to normalize its values to $[0, 1]$, indicating the adversary's degree of uncertainty. The bounded value range is likely to make normalized entropy suitable for between-scenario comparisons.

$$priv_{NE} \equiv \frac{H(X)}{H_0(X)}.$$

Conditional Entropy describes how much information (in bits) is needed to describe the true mapping X^* between observations and existing tracks, conditioned on the adversary's estimate X.

$$priv_{COE} \equiv H(X^*|X) = -\sum_{x \in X, x^* \in X^*} p(x, x^*) \log_2 p(x^*|x)$$

Inherent Privacy and *conditional privacy* are based on *entropy* and *conditional entropy*, respectively. Both metrics indicate how many yes/no questions the adversary would have to answer correctly to describe the ground truth.

$$priv_{IP} \equiv 2^{H(X)}; priv_{CP} \equiv 2^{H(X^*|X)}.$$

3.3.2 Information Gain/Loss Metrics

Information gain/loss metrics measure how much information the adversary gains (or how much privacy the user loses) through the adversary's observation.

Amount of Leaked Information indicates how many vehicles v the adversary can track correctly, i.e., all cases in which the observation with the highest probability corresponds to the correct vehicle. Its values strongly depend on the total number of vehicles in a scenario.

$$priv_{ALI} \equiv |V|, \forall v \in V : \max p(x_v) = x_v^*.$$

Mutual Information indicates the amount of information shared between the distribution of the adversary's estimate X and the true mapping X^*.

$$priv_{MI} \equiv I(X^*; X) = H(X^*) - H(X^*|X).$$

Conditional Privacy Loss is based on *mutual information* and measures the fraction of privacy lost through the adversary's estimate.

$$priv_{CPL} \equiv 1 - 2^{-I(X^*;X)}.$$

Pearson Correlation measures the degree of linear dependence between the adversary's estimate and the ground truth, with a lower coefficient indicating higher privacy.

$$priv_{PCC} \equiv \frac{cov(X^*, X)}{\sigma_{X^*} \sigma_X}.$$

3.3.3 Error Metrics

Error metrics measure how far the adversary's Estimate is from the ground truth, either in terms of probabilities or in terms of geographical distance.

Expected Distance Error measures the Expected euclidean distance $d(x, x^*)$ between the true location and the estimated location over multiple time steps t.

$$priv_{EDE} \equiv \frac{1}{|V|T} \sum_{t \in T} \sum_{v \in V} \sum_{x \in X} p(x_{v,t}) d(x, x^*).$$

Incorrectness indicates the adversary's probability of error. It replaces the Euclidean distance with an indicator function $\hat{d}$ that yields 0 if the adversary was able to track a vehicle correctly, and 1 if the tracking was not successful.

$$priv_{INC} \equiv \sum_{x \in X} p(x) \hat{d}(x, x^*).$$

3.3.4 Adversary's Success Metrics

Adversary's success metrics quantify how likely it is that the adversary succeeds.

Privacy Breach Level indicates the posterior probability the adversary assigns to the true vehicle, given the observations y from the current time step.

$$priv_{PBL} \equiv p(x = x^*|y).$$

3.3.5 Time Metrics

Time metrics are based on the time during which the adversary can (or cannot) successfully track a vehicle.

Time to Confusion indicates the cumulative time during which entropy is below a threshold h, i.e., the time during which the adversary is not confused.

$$priv_{TC} \equiv \textit{ Time during which } H(X) < h.$$

3.4 Criteria for Metric Strength

We use four criteria to evaluate the strength of these privacy metrics: monotonicity, the spread of the value range in terms of extent and evenness, and the portion of the value range that is shared across scenarios.

3.4.1 Monotonicity

The most important requirement for privacy metrics is monotonicity, i.e., metrics should indicate decreasing privacy values with increasing adversary strength. Non-monotonic metrics may indicate the same privacy level for weak and strong adversaries, or for strong and weak PETs. The use of non-monotonic metrics can thus lead to misjudging the strength of privacy protections, and subsequently to real-world privacy violations.

We have previously proposed an algorithm to compute monotonicity scores [15] (adapted to vehicular networks in Fig. 4). In brief, the algorithm uses two statistical tests 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 statistically significant difference in the expected direction, -1 for a statistically significant difference in the wrong direction, -2 for a change in direction (such a peak means that strong and weak adversaries cannot be distinguished and is thus not desirable), and -0.2 for a change that is either zero or not statistically significant (slight penalty for metrics that have similar values for successive adversaries). The total monotonicity score is the

Input metric values for each traffic condition c and each adversary strength a_i
Output monotonicity scores m_c for one privacy metric

$tests \leftarrow$ {Welch's t-test, Wilcoxon rank-sum statistic}
foreach traffic condition c **do**
  $m_c \leftarrow 0$
  foreach $test \in tests$ **do**
    $prevResult \leftarrow 0$
    foreach pair of succ. adv. strengths (a_i, a_{i+1}) **do**
      apply $test$ to (a_i, a_{i+1})
      $p \leftarrow$ statistical significance of test
      $result \leftarrow$ value of test statistic
      if $p < 0.05$ **then**
        if $result > 0$ (< 0 for LB metrics) **then**
          $m_c \leftarrow m_c + 1$
        else if $result < 0$ (> 0 for LB metrics) **then**
          $m_c \leftarrow m_c - 1$
        else
          $m_c \leftarrow m_c - 0.2$
        end if
      else
        $m_c \leftarrow m_c - 0.2$
      end if
      if sign($result$) $\neq$ sign($prevResult$) **then**
        $m_c \leftarrow m_c - 2$
      end if
      $prevResult \leftarrow result$
    end for
  end for
end for
normalize all m_c to $[0, 1]$
return monotonicity scores m_c

Fig. 4. Algorithm to calculate monotonicity scores. LB (lower-better) refers to metrics where lower values indicate higher privacy.

addition of these point values. We normalize monotonicity scores to $[0, 1]$ based on the monotonicity values for all metrics in our study.

3.4.2 Extent and Evenness of Spread

The spread of a metric's value range indicates how suitable a metric is to distinguish privacy levels *within* scenarios. A large spread makes it easier to identify statistically significant differences between privacy levels. This helps to judge whether a PET works equally well in different parts of a city, for example in areas of high or low traffic density, allows to compare vehicles to each other, and allows to evaluate privacy levels over time. To support these within-scenario comparisons, the metric's value range should spread *evenly* over a *large* value range.

To measure the extent of the spread, we first calculate the standard deviation σ of the normalized metric values for all adversary strengths individually. The extent score then corresponds to the average standard deviation over all adversary strengths, normalized to $[0, 1]$ based on the extent values for all metrics in our study.

To measure the evenness of the spread, we analyze the uniformity of metric values, i.e., how close the distribution of values is to a uniform distribution. We use the Cramér-von Mises criterion, which measures the goodness of fit between a theoretical distribution and an empirical distribution, to analyze the fit between the uniform distribution $U(0, 1)$ and the normalized metric values for all adversary strengths combined. Because the Cramér-von Mises criterion is influenced by the number of samples, we normalize the criterion by the number of metric values.

3.4.3 Shared Value Range

How much of a metric's value range is shared across traffic conditions indicates how suitable a metric is to compare privacy levels *between* different scenarios, for example with different traffic characteristics or different road layouts. This helps to judge whether the performance of a new PET is independent of specific traffic patterns, that is, whether PETs work equally well regardless of the time of day, day of the week, or city in which they are deployed. To support these between-scenario comparisons, metrics should not be influenced by the number of vehicles or the size of the area. For example, we expect that metrics that use some form of normalization will have a large shared value range.

To formalize this criterion, we measure how much of a metric's value range is shared between traffic conditions. We first calculate the global value range for each metric across all traffic conditions and then compute the percentage of the global value range used in each traffic condition.

3.4.4 Discussion of Criteria

Of the four criteria for metric strength we have defined in this section, monotonicity is the most important requirement that all metrics should satisfy. The other three criteria focus on more specific requirements: extent and evenness are important to compare privacy levels *within* a scenario, and shared value range is important to compare privacy levels *between* scenarios. Their usefulness thus depends on what kinds of comparisons the metrics are being used for.

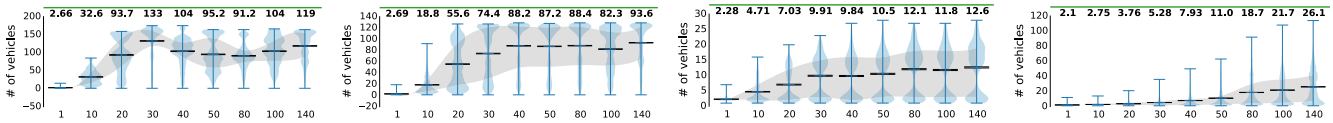
4 Results

We have applied our methodology to all nine levels of adversary strength in all fifteen traffic conditions, and evaluated 41 privacy metrics with respect to our four criteria for metric strength.

For each criterion, we first present detailed results to illustrate the criterion. Due to the volume of result data (~ 800 GB and more than 2000 individual plots), we present only a small subset of our results in detail. We then present the full set of results in aggregated heat maps and show how the strength of metrics can depend on the traffic condition. Finally, we rank metrics by their strength for each criterion and derive specific recommendations for metric selection in Section 5.

4.1 Monotonicity

To illustrate our results for the monotonicity requirement, Fig. 5 shows one metric, the *anonymity set size*, in four traffic conditions. Each subfigure shows the distribution of metric values for the nine adversary strength levels using violin plots, and additionally indicates confidence intervals (horizontal lines), the area between quartiles (shaded), mean values (bold numbers), and whether higher or lower numbers indicate higher privacy (green line). The full set of violin plots for all metrics and traffic conditions is included in the supplementary material, available online.



(a) US 101 8:20, monotonicity=0.51 (b) US 101 7:50, monotonicity=0.63 (c) Rome 8am, monotonicity=0.71 (d) SFO 8am, monotonicity=1.00

Fig. 5. Anonymity set size in four traffic conditions, ordered from lowest to highest monotonicity.

The anonymity set size for US highway 101 at 8:20 am (Fig. 5a) has the lowest monotonicity rating, caused by changes in the wrong direction between adversary strengths 30 and 80, and by the negative peak at adversary strength 80. US highway 101 at 7:50 am (Fig. 5b) has the next highest monotonicity rating, caused by the negative peak at adversary strength 100. In Rome (Fig. 5c), the anonymity set size is monotonic, but several strength levels have no statistically significant difference (e.g., 30/40 and 80/100). In the figure, this lack of a statistically significant difference can be seen in the overlapping confidence intervals between neighboring violins. San Francisco (Fig. 5d) has the highest monotonicity rating for the adversary's success rate because the metric is monotonic and all strength levels are clearly distinguishable.

In Fig. 6, we use a heat map to visualize monotonicity scores in a compact way. Each square represents one set of results presented in detailed violin plots above, computed according to our algorithm in Fig. 4. For example, the last square in the third row summarizes Fig. 5a (*anonymity set size* for US highway 101, 8:20 am). The heat map thus summarizes the results for 15 traffic conditions and 44 metrics, i.e., 660 individual results.

The heat map shows that several metrics have high monotonicity regardless of the traffic condition, for example *entropy* and the *privacy breach level*. Very few metrics are non-monotonic throughout and therefore not recommended, for example *information surprisal*. The monotonicity of most other metrics varies depending on the traffic condition. For example, the *adversary's success rate* is very strong in three traffic conditions but only of medium strength in the other conditions. If these metrics are selected to evaluate a new PET, it is necessary to validate their monotonicity for the specific scenario.

Heat maps visualize a large number of results by traffic condition, but they do not show the overall ranking of privacy metrics. To do this, we aggregate the heat map into box plots, such that each row in the heat map is represented by one box. We then sort the boxes by their mean values and plot the 15 best metrics (we show the full plot in Appendix C, available in the online supplemental material).

Fig. 7 shows that the metric with the highest monotonicity score is *entropy*, followed by seven other metrics that are derived from entropy. We note that several of the metrics that have been proposed to evaluate PETs for vehicular networks, such as the *maximum tracking time*, the *time to confusion*, and the *mean tracking duration*, are not among the strongest metrics (in fact, their average monotonicity scores are below 0.5).

The normalized monotonicity scores of the top metrics are higher than 0.5 in all cases, indicating that the metrics are mostly monotonic and therefore suitable to evaluate and compare new PETs.

A score below 0.5 is generally undesirable because it indicates the presence of non-monotonic behavior, for example cases where a metric indicates higher (instead of lower) privacy for a stronger adversary. These metrics are not suitable to evaluate the performance of PETs because they may misjudge not only how well a new PET protects privacy, but also how two PETs compare to each other.

Fig. 6. Heat map for monotonicity. The colors indicate the monotonicity score (from yellow = low to blue = high).

4.2 Extent and Evenness of Spread

To illustrate our requirement for spread, we plot the privacy values as colors on city maps, such that light colors indicate high privacy and dark colors indicate low privacy (Fig. 8). The light/dark color sequence corresponds to the global value range for each metric.

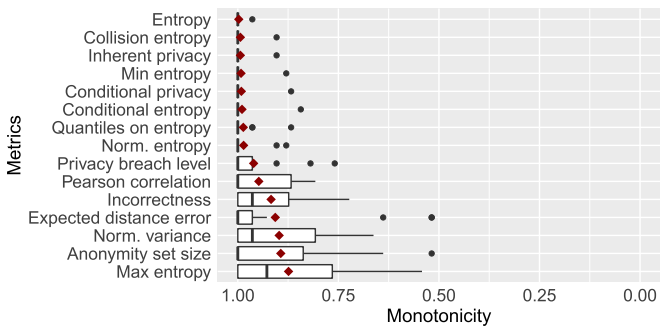
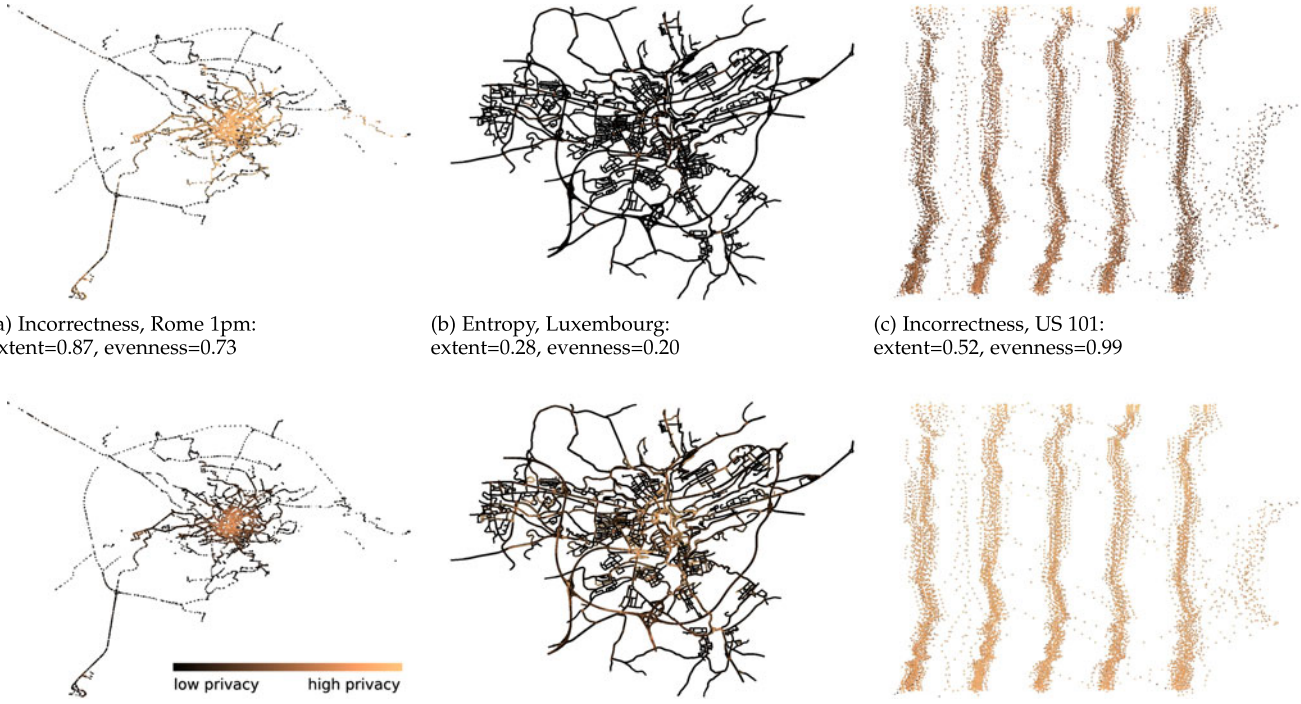


Fig. 7. Distribution of the monotonicity for the 15 best metrics across all traffic conditions. The top-8 metrics all belong to the uncertainty category.



(a) Incorrectness, Rome 1pm: extent=0.87, evenness=0.73

(b) Entropy, Luxembourg: extent=0.28, evenness=0.20

(c) Incorrectness, US 101: extent=0.52, evenness=0.99

(d) Entropy, Rome 1pm: extent=0.58, evenness=0.84

(e) Mutual Information, Luxembourg: extent=0.65, evenness=0.55

(f) Mutual Information, US 101: extent=0.24, evenness=0.58

Fig. 8. Spread: extent versus evenness. These plots show an overplot of five adversary strengths (1, 20, 40, 80, 140) with 20 percent transparency. All metrics in this plot have high monotonicity, but show varying degrees of extent and evenness.

In the Rome (1 pm) traffic condition, *incorrectness* (Fig. 8a) has a large extent (0.87), which can be seen in the clear representation of both the darkest and lightest colors, indicating that there is non-negligible probability mass on a large range of privacy values. The evenness is lower (0.73), which can be seen through a lower proportion of medium browns compared to roughly equal proportions of the darkest and lightest colors. In contrast, *entropy* (Fig. 8d) has a lower extent (0.58), which is visible in the absence of very light colors, but a higher evenness (0.87), indicated by the clear visibility of light, medium, and dark colors.

In Luxembourg, *entropy* (Fig. 8b) has both low extent and low evenness, which can be seen in the complete absence of light colors and a large overrepresentation of dark colors. *Mutual information* (Fig. 8e) has a higher extent, indicated by the presence of lighter colors in the center, but the evenness is still low, again indicated by the overrepresentation of dark colors. For this traffic condition, *mutual information* is thus more suitable to visualize differences in privacy between the dense city center and the less-dense outskirts.

For traffic on the US highway 101, *incorrectness* (Fig. 8c) shows a much lower extent compared to Rome, but a very high evenness. This highlights an interesting property of the *incorrectness* metric: we find that its distribution is bimodal for all inner-city traffic conditions, but not for the highway traffic conditions. As a result, *incorrectness* is one of the few metrics where we find a statistically significant difference between city traffic and highway traffic (i.e., no overlap in 95 percent confidence intervals for extent). Other metrics with a similar behavior include *conditional privacy loss*, *privacy breach level*, and *mutual information* (Fig. 8f).

In Fig. 9, we show the extent and evenness scores on heat maps. Even though the extent scores are lower on average than the evenness scores, the two heat maps show a similar pattern of high and low scores, indicating that extent and evenness may be correlated. We discuss correlations between our four criteria in Section 4.5.

The heat maps show that some metrics have a low extent in all scenarios, e.g., *expected distance error* and *user-centric location privacy*, and some metrics have low evenness throughout, e.g., *cross entropy* and *user-centric location privacy*. Even though the monotonicity of these metrics may be high, their low spread in terms of extent and/or evenness makes them less suitable to measure differences in privacy within a scenario or to visualize privacy levels on a map.

Some metrics score highly in extent but low on evenness, for example *incorrectness* and *conditional privacy loss*. The values of these metrics generally have a bimodal distribution. *Incorrectness*, for example, has most of its probability mass on the values 0 and 1, and very little probability mass in between. Although these metrics can clearly separate vehicles that enjoy high resp. low privacy, they are less suitable for visualization and fine-grained analysis than metrics that score highly on evenness, such as *max-entropy*.

Metrics with both high extent and high evenness, such as *max-entropy* and *privacy breach level*, are most desirable, because they can show within-scenario differences clearly and can highlight how privacy levels change between areas of low and high privacy.

Generally, metrics with a high monotonicity score do not necessarily score highly in spread. For example, the metric with the highest monotonicity score, *entropy*, only has medium extent and evenness scores. Comparing the ranking of metrics according to monotonicity (Fig. 7) and extent (Fig. 10a), we find that only four metrics occur in both top 15 lists (*incorrectness*, *privacy breach level*,

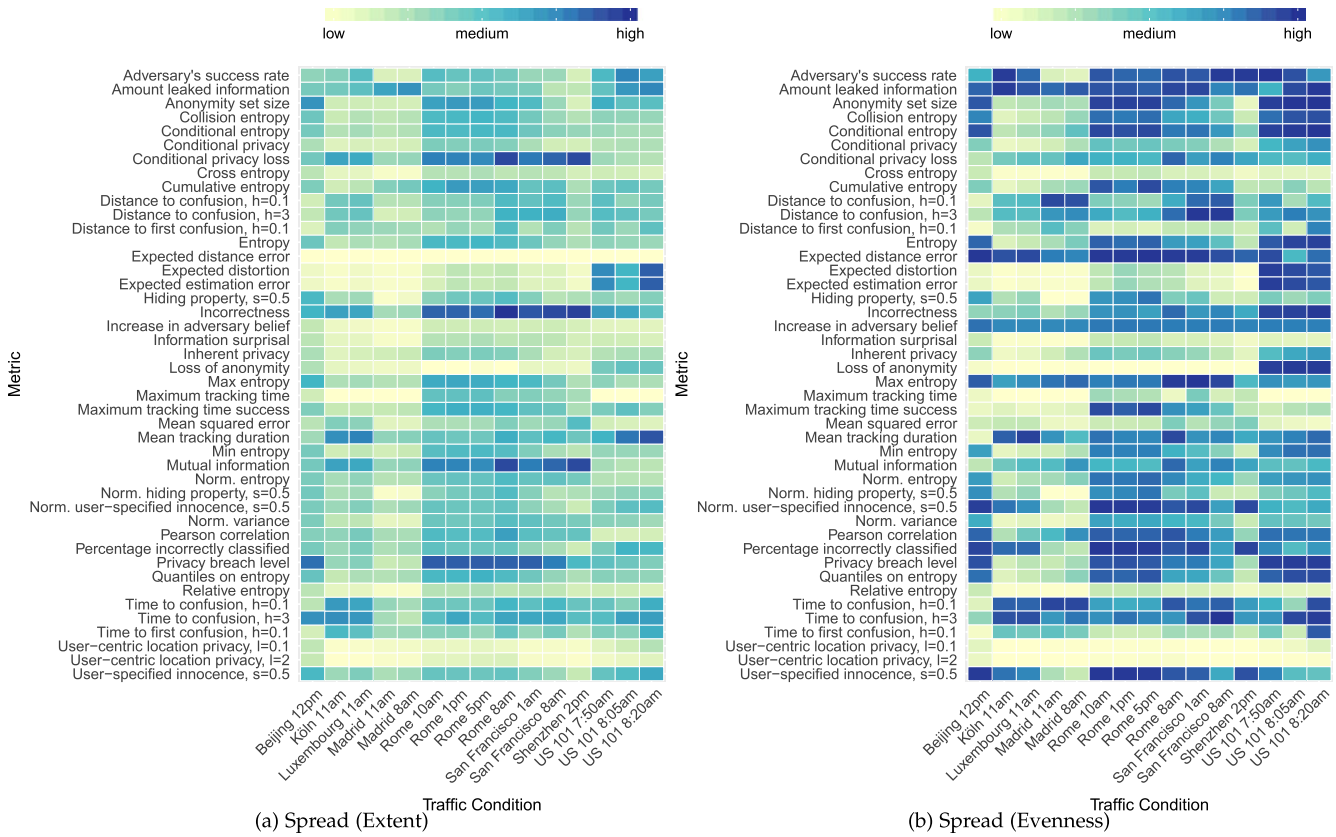


Fig. 9. Heat maps for the spread (extent and evenness) of privacy metrics. The colors indicate the value of each criterion (from yellow = low to blue = high).

anonymity set size, and *max-entropy*). This indicates that in some cases it may be necessary to trade-off monotonicity against extent. The choice of metrics in such a situation depends on the purpose of the evaluation. One possible choice is to find a compromise metric with relatively high scores in both criteria, e.g., *privacy breach level* or *max-entropy*, to both evaluate the effectiveness of PETs and visualize privacy levels on a map. Another choice is to combine the best metrics from each criterion, i.e., *entropy* and *incorrectness*, in a metrics suite.

Compared with the distribution of monotonicity in Fig. 7, the boxes in Fig. 10 have wider ranges. This means that even the metrics with the highest extent and evenness scores may be weak in some traffic conditions. Their suitability for within-scenario comparisons can therefore be condition-specific and should be validated before the metrics are used.

4.3 Shared Value Range

Metrics that use the same value range regardless of the traffic condition are more suitable to compare privacy levels between scenarios. To illustrate this requirement for a shared value range, Fig. 11 shows the *anonymity set size* and *distance to confusion* in two traffic conditions each (top row) and the *normalized entropy* in four traffic conditions (bottom row). We can see that the value range for *anonymity set size* depends heavily on the traffic conditions, ranging up to 80 in Rome (not shown), 120 in Shenzhen (Fig. 11a), and 800 in Beijing (Fig. 11b). For Beijing, the shared value range is 1.00 because the metric values cover the entire global value range. Rome (0.09) and Shenzhen (0.14) indicate much lower values for the shared value range.

A similar observation holds for the *distance to confusion*, which ranges up to 14000 on Madrid's A6 highway (Fig. 11c), but only up to 800 on the US 101 highway (Fig. 11d). This

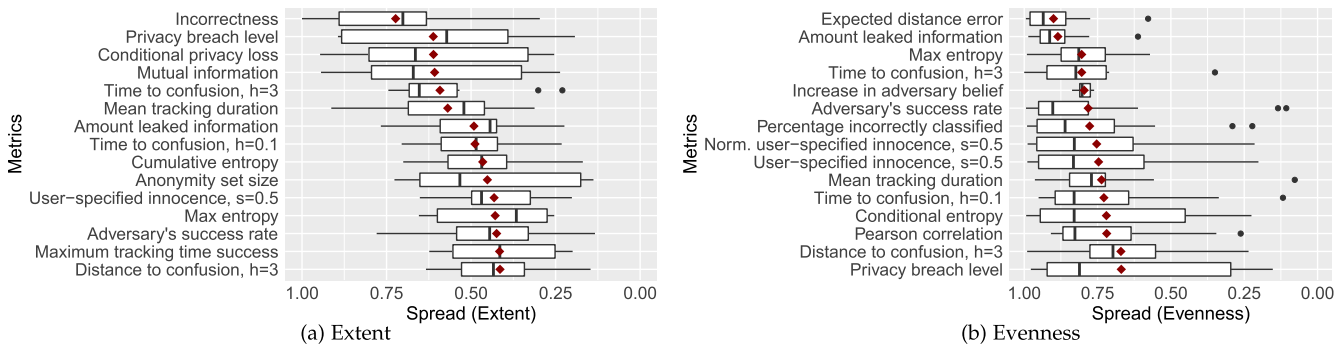
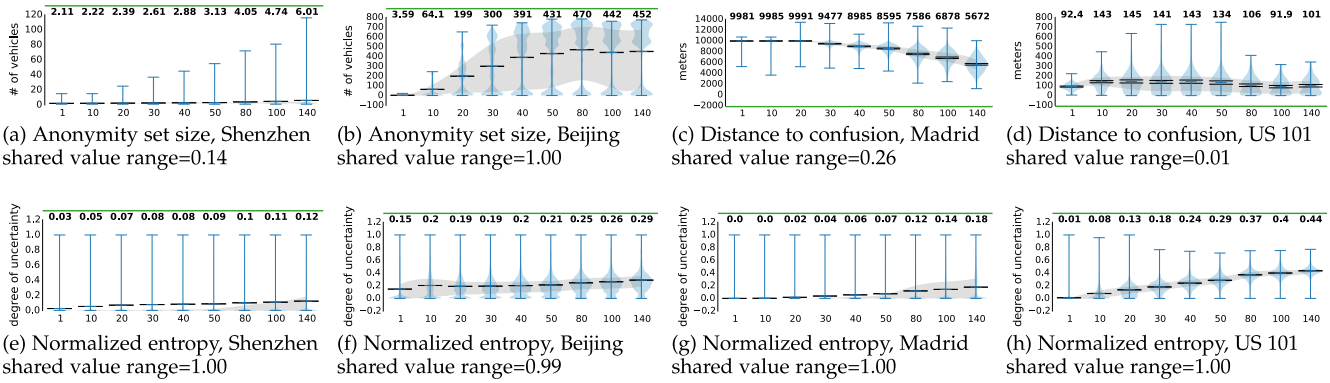


Fig. 10. Distribution of the metric's spread for the best 15 metrics across all traffic conditions.



(a) Anonymity set size, Shenzhen shared value range=0.14
(b) Anonymity set size, Beijing shared value range=1.00
(c) Distance to confusion, Madrid shared value range=0.26
(d) Distance to confusion, US 101 shared value range=0.01
(e) Normalized entropy, Shenzhen shared value range=1.00
(f) Normalized entropy, Beijing shared value range=0.99
(g) Normalized entropy, Madrid shared value range=1.00
(h) Normalized entropy, US 101 shared value range=1.00

Fig. 11. Shared value range: metrics in the top row show very different value ranges (y axis), while the bottom row (normalized entropy) has a shared value range of [0,1] across all scenarios.

indicates that neither the *anonymity set size* nor the *distance to confusion* are suitable for between-scenario comparisons. In contrast, the bottom row shows the value of *normalized entropy* in the same four traffic conditions as in the top row. In each case, the metric is valued between 0 and 1, indicating that *normalized entropy* supports comparisons between traffic conditions.

Fig. 12 summarizes our results for the shared value range in a heat map. Most metrics with a low shared value range across traffic conditions, e.g., *amount of leaked information*, *expected estimation error*, and *hiding property*, are calculated from absolute values such as the number of vehicles, distance, and time, which do not have a natural upper limit. These metrics vary significantly between different scenarios, resulting in a small shared value range. Metrics that have a large shared value range across traffic conditions, e.g., *conditional privacy loss*, *incorrectness*, and *normalized entropy*, use fractions or ratios to calculate their values. As a result, their value ranges have defined upper and lower limits and a larger portion of it is shared across traffic conditions.

Fig. 13 ranks the 15 metrics that score highest on shared value range. We note that some metrics with a high shared value range score very low on monotonicity, e.g., *increase in adversary belief* and *information surprisal*. Despite their shared value range, these metrics cannot be recommended to compare privacy between scenarios because they may misjudge the strength of the adversary or PET in the scenario.

Eight of the metrics in Fig. 13 also occur in the list of top metrics for monotonicity, e.g., *incorrectness* and *normalized entropy*. These metrics can be recommended for between-scenario comparisons.

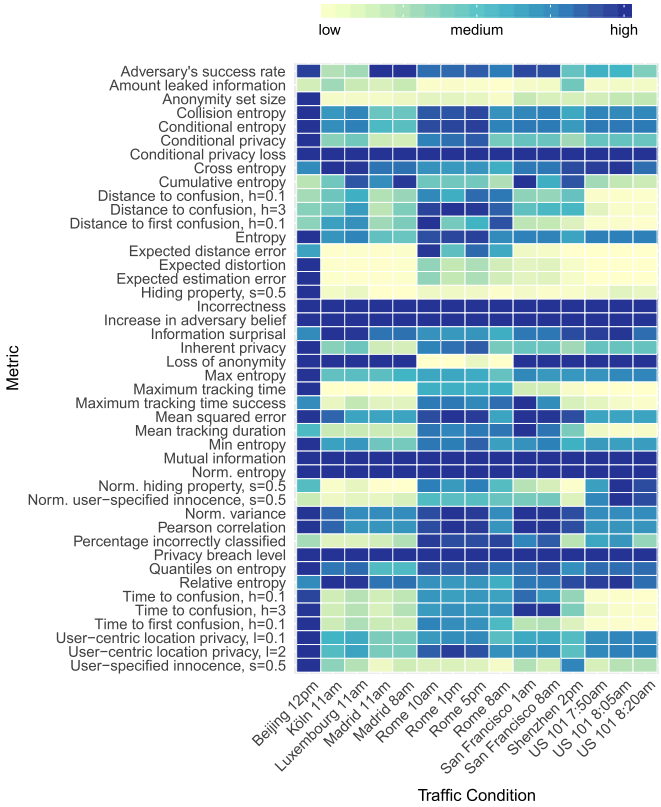


Fig. 12. Heat maps for the shared value range of privacy metrics. The colors indicate the size of the shared value range (from yellow = low to blue = high).

4.4 Influence of Parameter Settings

We studied nine metrics that are configurable with a parameter: *(normalized) hiding property*, *(normalized) user-specified innocence*, *time/distance to (first) confusion*, and *user-centric location privacy*. Our experiments show that the metric values depend on the parameter setting in each case, i.e., the privacy level indicated by metrics depends on the parameter value. In this section, we analyze whether the metric strength in terms of monotonicity, extent and evenness of spread, and shared value range depends on the parameter setting as well.

(Normalized) hiding property and *(normalized) user-specified innocence* use a threshold s for the adversary's probability. We find that the value of s does not influence the strength

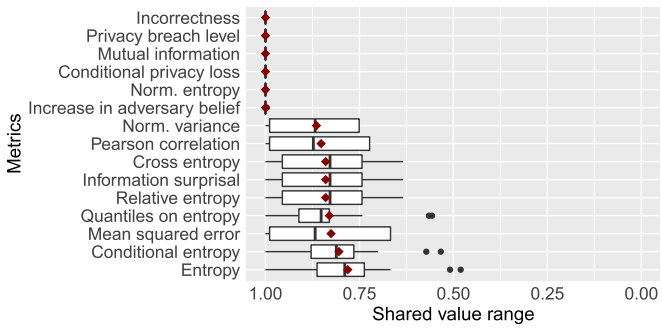
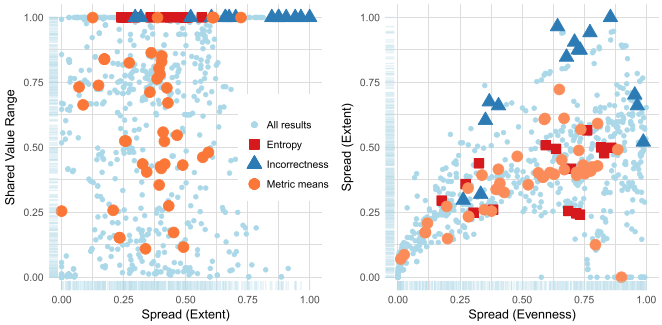


Fig. 13. Distribution of the shared value range for the best 15 metrics across all traffic conditions.



(a) Extent vs. Shared value range (b) Evenness vs. Extent

Fig. 14. Correlation between criteria. Pearson correlation coefficient $r = 0.60$ for evenness/extent, $r < 0.16$ for all others.

of the metrics (the full results are shown in the heat maps in Appendix D, available in the online supplemental material). We therefore discuss only a single parameter level, $s = 0.5$, in the paper.

Time/distance to (first) confusion use a threshold h on the adversary's uncertainty (i.e., entropy). We find that the strengths of *time/distance to first confusion* are not influenced by the value of h, and therefore we discuss only $h = 0.1$ for these metrics. In contrast, the strengths of *time/distance to confusion* vary depending on the parameter setting. Specifically, the spread (both extent and evenness) and shared value range improve with higher parameter values, while monotonicity improves with lower parameter values. We therefore discuss $h = 0.1$ as well as $h = 3$ in the paper.

User-centric location privacy uses the parameter l to express the rate of privacy decay over time. We find that the evenness of the spread is not influenced by l, whereas monotonicity and the extent of the spread improve with lower parameters, and the shared value range improves with higher values of l. We therefore discuss $l = 0.1$ and $l = 2$ in the paper.

4.5 Correlation between Criteria

To show that all four of our criteria are necessary to measure the strength of privacy metrics, we evaluate whether they are independent or correlated with each other.

Fig. 14 shows two of the pairwise correlations between the four criteria in scatter plots (we show the remaining pairwise correlations in Appendix B, available in the online supplemental material). The plots show one small blue circle for each combination of metric and traffic condition. Large orange circles indicate the average value for each metric. To show how individual metrics behave in different traffic conditions, we highlight *entropy* with red squares and *incorrectness* with blue triangles.

It is clear from Fig. 14a that extent and shared value range are not correlated ($r = 0.15$), and similar results hold for most of the other correlations. This indicates that all criteria are necessary to evaluate the strength of privacy metrics because they evaluate independent aspects of the behavior of privacy metrics.

The only exception is the correlation between extent and evenness ($r = 0.60$, Fig. 14b), i.e., between two criteria that measure the spread of metric values. Although the correlation coefficient of 0.60 indicates a positive correlation, the correlation is not very high. As a result, there are several

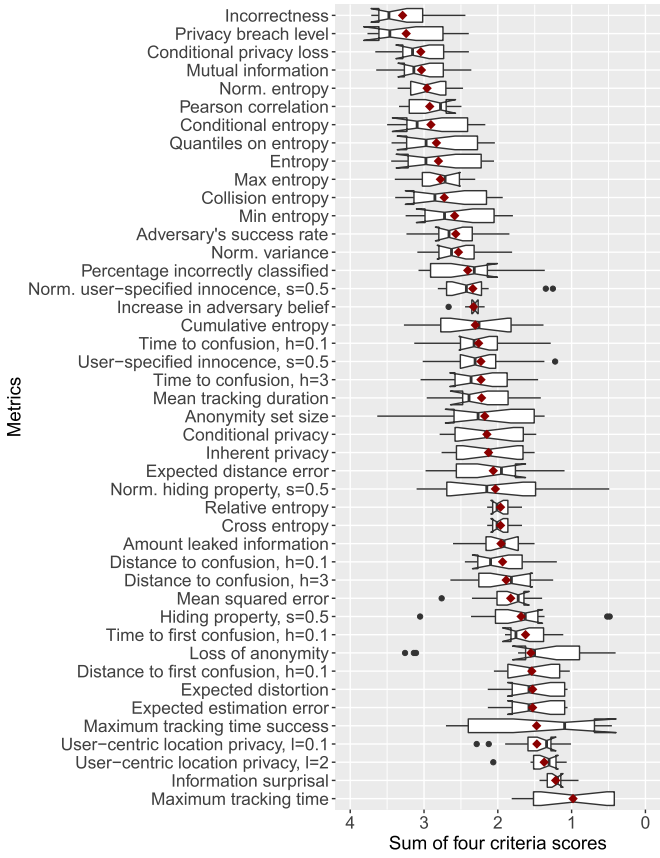


Fig. 15. Overall ranking of metric strength. The x axis shows the sum of the four criteria (Monotonicity + Spread (Extent) + Spread (Evenness) + Shared value range), each normalized to $[0, 1]$, with equal weights.

cases where extent and evenness diverge and thus it is beneficial to evaluate both aspects of a metric's spread. For example, while the mean values of most metrics (orange circles) in Fig. 14b follow the linear correlation, many values for *incorrectness* (blue triangles) show much higher extent than would be expected, while other metrics show much lower extent than expected, e.g., *expected distance error* and *increase in adversary belief* (the two bottom-right orange circles).

5 Discussion

We have discussed and ranked 41 privacy metrics according to four criteria: monotonicity, extent, evenness, and shared value range. To make the choice of privacy metrics easier, we now aggregate the four criteria into a single ranking. We calculate the overall score for each metric by adding up the normalized values for each criterion. Fig. 15 shows that all metrics fall short of the maximum score of 4: the metric with the best mean value across all scenarios is *incorrectness* (3.29), and the best strength score in a single traffic condition is 3.76 (*privacy breach level*, Rome 10 am).

To compare whether the differences in overall metric scores are statistically significant, Fig. 15 shows notched box plots. The notches depend on the inter-quartile range (IQR) and extend to $1.58 * IQR/\sqrt{n}$, indicating a roughly 95 percent confidence interval for the median. We note that the notches for several metrics are overlapping, indicating that there is no statistically significant difference between the medians. In particular, the confidence interval for the first metric

incorrectness overlaps with the confidence intervals of seven other metrics (*privacy breach level* down to *collision entropy*, with the exception of *normalized entropy*, *Pearson correlation*, and *max-entropy*). As a result, we cannot decide on a clear "winner" metric. In addition, the individual rankings for the four criteria show that there is no single metric that outperforms the others in all criteria and for all traffic conditions.

5.1 Weak Metrics

We note that some of the metrics that have been proposed specifically for use in vehicular networks score low on monotonicity and are in the bottom half of our overall metric ranking. These metrics include the *mean tracking duration*, *time/distance to confusion*, and *maximum tracking time*. Even though these metrics may make sense intuitively, they can make PETs appear stronger than they are (low monotonicity) or skew comparisons between scenarios (low shared value range). Therefore we recommend to replace these metrics with other metrics that are stronger in vehicular network scenarios.

5.2 Visualization to Support PET Design

Our visualization of privacy metrics on city maps (Fig. 8) showed that privacy often depends on the road layout and traffic density in a city. For example, privacy levels were often higher in the city center, and decreased towards the outskirts of a city. A possible consequence for the design of PETs is that it may make sense to apply one PET in city centers with dense traffic, and choose another PET for outskirts with less dense traffic, or to adjust parameter settings to provide adequate privacy in all areas.

Visualizing privacy levels on a map can support these design decisions because it highlights in which areas different PETs are most effective. Metrics that have high extent and evenness and generate per-time and per-vehicle values, e.g., *max-entropy* or *privacy breach level*, are suitable to create such visualizations.

5.3 Metrics Suites

Even the best metrics in our experiments do not perform well in all traffic conditions, as indicated in our box plots and heat maps. One solution to this problem is to validate all metrics before applying them to new traffic conditions. Depending on the traffic data, this may take a long time and may not always be feasible.

A better solution is therefore to combine several metrics into a metrics suite, i.e., to always work with multiple metrics. This approach can offset weaknesses in metrics, especially if the metrics in the suite are chosen carefully. We recommend to consider three aspects when choosing a metrics suite:

- Only use metrics with a high monotonicity score
- Include metrics from different categories, e.g., uncertainty, information gain/loss, and error (see Section 3.3 and Table 2)
- Include metrics that are particularly strong for within-scenario comparisons as well as metrics that are strong for between-scenario comparisons.

An example metrics suite could thus consist of *normalized entropy* (uncertainty, high shared value range), *conditional privacy loss* (information gain/loss, high extent), *incorrectness* (error, high extent and shared value range), *privacy breach level* (adversary's success probability, high extent and shared value range, good evenness), and *time to confusion* with $h = 0.1$ (time, high evenness). This metrics suite has an average monotonicity score of 0.86.

To allow for the construction of metrics suites that meet custom requirements, we publish our dataset with detailed results for all four criteria in the supplementary material, available online.

6 Conclusion

We have introduced four novel criteria to evaluate the strength of privacy metrics: monotonicity, extent, evenness, and shared value range. These criteria measure the consistency of privacy metrics and their suitability for within-scenario and between-scenario comparisons of privacy levels. In extensive experiments, we have applied these criteria to 41 privacy metrics in fifteen traffic conditions. Our results allowed us to reason about the strength of privacy metrics and generate an overall ranking of privacy metrics.

Our key findings are that (1) several existing metrics have low monotonicity scores, i.e., they can misjudge the strength of new privacy-enhancing technologies, (2) no single metric dominates across all criteria and traffic conditions, and (3) visualization can highlight where privacy depends on road layout and can thus support the design of PETs.

Based on these findings, we recommend to always use metrics suites when evaluating new PETs, i.e., to combine several privacy metrics that have high monotonicity scores, measure different outputs, and are strong for either within-scenario or between-scenario comparisons.

Acknowledgments

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] D. Eckhoff and C. Sommer, "Driving for big data? Privacy concerns in Vehicular Networking," *IEEE Secur. Privacy*, vol. 12, no. 1, pp. 77–79, Jan. 2014.

[2] B. Wiedersheim, Z. Ma, F. Kargl, and P. Papadimitratos, "Privacy in inter-vehicular networks: Why simple pseudonym change is not enough," in *Proc. 7th Int. Conf. Wireless On-Demand Netw. Syst. Serv.*, Feb. 2010, pp. 176–183.

[3] I. Wagner and D. Eckhoff, "Technical privacy metrics: A systematic survey," *ACM Comput. Surveys*, vol. 51, no. 3, Art. no. 57, Apr. 2018.

[4] K. Sampigethaya, M. Li, L. Huang, and R. Poovendran, "AMOEBA: Robust location privacy scheme for VANET," *IEEE J. Select. Areas Commun.*, vol. 25, no. 8, pp. 1569–1589, Oct. 2007.

[5] D. Eckhoff, C. Sommer, T. Gansen, R. German, and F. Dressler, "Strong and affordable location privacy in VANETs: Identity diffusion using time-slots and swapping," in *Proc. IEEE Veh. Netw. Conf.*, Dec. 2010, pp. 174–181.

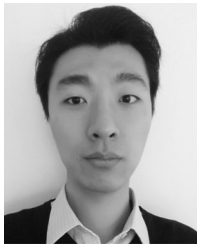
[6] J. Freudiger, M. Raya, M. Félegyházi, P. Papadimitratos, and J.-P. Hubaux, "Mix-zones for location privacy in vehicular networks," presented at the *1st Int. Workshop Wireless Netw. Intell. Transp. Syst.*, Vancouver, BC, Canada, Aug. 2007.

[7] I. Wagner and D. Eckhoff, "Privacy assessment in vehicular networks using simulation," in *Proc. Winter Simulation Conf.*, Dec. 2014, pp. 3155–3166.

[8] A. Wasef and X. Shen, "REP: Location privacy for VANETs using random encryption periods," *Mobile Netw. Appl.*, vol. 15, no. 1, pp. 172–185, Feb. 2010.

[9] R. Shokri, G. Theodorakopoulos, J. Y. Le Boudec, and J. P. Hubaux, "Quantifying location privacy," in *Proc. IEEE Symp. Secur. Privacy*, May 2011, pp. 247–262.

[10] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Preserving privacy in GPS traces via uncertainty-aware path cloaking," in *Proc. 14th ACM Conf. Comput. Commun. Secur.*, Oct. 2007, pp. 161–171.
[11] J. Alexander and J. Smith, "Engineering privacy in public: Confounding face recognition," in *Proc. 3rd Int. Workshop Privacy Enhancing Technol.*, vol. 2760, pp. 88–106, Mar. 2003.
[12] E. Bertino, D. Lin, and W. Jiang, "A survey of quantification of privacy preserving data mining algorithms," in *Privacy-Preserving Data Mining: Models and Algorithms*. New York, NY, USA: Springer, Jul. 2008, no. 34, ch. 8, pp. 183–205.
[13] P. Syverson, "Why I'm not an entropist," in *Proc. 17th Int. Workshop Secur. Protocols*, Apr. 2013, pp. 213–230.
[14] I. Wagner, "Measuring privacy in vehicular networks," in *Proc. 42nd IEEE Conf. Local Comput. Netw.*, Oct. 2017, pp. 183–186.
[15] I. Wagner, "Evaluating the strength of genomic privacy metrics," *ACM Trans. Privacy Secur.*, vol. 20, no. 1, pp. 2:1–2:34, Jan. 2017.
[16] S. J. Murdoch, "Quantifying and measuring anonymity," in *Proc. Data Privacy Manage. Auton. Spontaneous Secur.*, Jan. 2014, vol. 8247, pp. 3–13.
[17] R. W. Reeder, P. G. Kelley, A. M. McDonald, and L. F. Cranor, "A user study of the expandable grid applied to P3P privacy policy visualization," in *Proc. 7th ACM Workshop Privacy Electronic.*, Oct. 2008, pp. 45–54.
[18] Z. Ma, F. Kargl, and M. Weber, "Measuring long-term location privacy in vehicular communication systems," *Comput. Commun.*, vol. 33, no. 12, pp. 1414–1427, Jul. 2010.
[19] S. van der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy array: A structure for efficient numerical computation," *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 22–30, Mar. 2011.
[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.
[21] L. Dalcín, R. Paz, M. Storti, and J. D'Elía, "MPI for Python: Performance improvements and MPI-2 extensions," *J. Parallel Distrib. Comput.*, vol. 68, no. 5, pp. 655–662, May 2008.
[22] L. Bracciale, M. Bonola, P. Loreti, G. Bianchi, R. Amici, and A. Rabuffi, "CRAWDAD Dataset roma/taxi (v.2014–07-17)," Jul. 2014. [Online]. Available: https://crawdad.org/roma/taxi/20140717
[23] C. Celes, F. Silva, A. Boukerche, R. Andrade, and A. Loureiro, "Improving VANET simulation with calibrated vehicular mobility traces," *IEEE Trans. Mobile Comput.*, vol. 16, no. 12, pp. 3376–3389, Dec. 2017.
[24] S. Uppoor, O. Trullols-Cruces, M. Fiore, and J. M. Barcelo-Ordinas, "Generation and analysis of a large-scale urban vehicular mobility dataset," *IEEE Trans. Mobile Comput.*, vol. 13, no. 5, pp. 1061–1075, May 2014.
[25] L. Codeca, R. Frank, and T. Engel, "Luxembourg SUMO traffic (LuST) scenario: 24 hours of mobility for vehicular networking research," in *Proc. IEEE Veh. Netw. Conf.*, Dec. 2015, pp. 1–8.
[26] Federal Highway Administration (FHWA), "Next generation simulation (NGSIM) US route 101 dataset," U.S. Dept. Transp. Intell. Transp. Syst. Joint Program Office, Washington, D.C., USA, Tech. Rep. FHWA-HRT-07-030, 2007.
[27] M. Gramaglia, O. Trullols-Cruces, D. Naboulsi, M. Fiore, and M. Calderon, "Vehicular networks on two Madrid highways," in *Proc. 11th Annu. IEEE Int. Conf. Sensing Commun. Netw.*, Jun. 2014, pp. 423–431.
[28] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*. Boston, MA, USA: Artech House Publishers, Jun. 1999.
[29] K. Emara, W. Woerndl, and J. Schlichter, "On evaluation of location privacy preserving schemes for VANET safety applications," *Comput. Commun.*, vol. 63, pp. 11–23, Jun. 2015.
[30] A. Serjantov and G. Danezis, "Towards an information theoretic metric for anonymity," in *Proc. 2nd Int. Workshop Privacy Enhancing Technol.*, 2003, pp. 41–53.



Yuchen Zhao received the BEng and MSc degrees, both in information security from the Huazhong University of Science and Technology, in 2011 and Wuhan University, in 2013, respectively, and the PhD degree in computer science from the University of St Andrews, in 2017. He is a research fellow in computer science (cybersecurity) with De Montfort University, Leicester, United Kingdom. His research revolves around privacy protections including privacy metrics, usable privacy, and user experience in privacy recommender systems.

Isabel Wagner received the MSc degree in computer science (Dipl.-Inf. Univ.) and the PhD degree in engineering (Dr.-Ing.) from the University of Erlangen, in 2005 and 2010, respectively. She is a senior lecturer in computer science (cybersecurity) with De Montfort University, Leicester, United Kingdom. In 2011 she was a JSPS postdoctoral fellow in the research group of Masayuki Murata with Osaka University, Japan. Her research interests include privacy and privacy-enhancing technologies, particularly on metrics to quantify the effectiveness of privacy protection mechanisms, as well as on privacy-enhancing technologies in genomics, smart cities, vehicular networks, and smart grids. She is also interested in bio-inspired mechanisms for privacy. She is a member of the IEEE.