On the Optimal Placement of Mix Zones

Julien Freudiger, Reza Shokri, and Jean-Pierre Hubaux

LCA1, EPFL, Switzerland
firstname.lastname@epfl.ch

Abstract. In mobile wireless networks, third parties can track the location of mobile nodes by monitoring the pseudonyms used for identification. A frequently proposed solution to protect the location privacy of mobile nodes suggests changing pseudonyms in regions called mix zones. In this paper, we propose a novel metric based on the mobility profiles of mobile nodes in order to evaluate the mixing effectiveness of possible mix zone locations. Then, as the location privacy achieved with mix zones depends on their placement in the network, we analyze the optimal placement of mix zones with combinatorial optimization techniques. The proposed algorithm maximizes the achieved location privacy in the system and takes into account the cost induced by mix zones to mobile nodes. By means of simulations, we show that the placement recommended by our algorithm significantly reduces the tracking success of the adversary.

1 Introduction

Modern mobile devices are increasingly equipped with peer-to-peer communication technologies, such as WiFi or Bluetooth, thus allowing them to directly exchange information with other devices in proximity. Such peer-to-peer communications enable *context-aware* applications. For example, vehicular networks provide safer and more efficient road transportation [23,47]. Similarly, mobile social networks allow users to automatically detect and exchange information with their friends [1,2,3,4]. In practice, mobile nodes detect each others' presence by periodically broadcasting messages and use *pseudonyms* instead of their actual identity (i.e., MAC/IP address, public key) to identify/authenticate each other.

However, much to the detriment of privacy, external parties eavesdropping on communications can monitor pseudonyms to learn mobile nodes' locations. Previous works [7,27,34] show that if the *spatial* and *temporal* correlation between successive locations of mobile nodes is not carefully eliminated, an external party (i.e., an adversary) can compromise the *location privacy* of mobile nodes and obtain the real identity of mobile nodes' owners. For example, using location traces collected in an office environment from the Active Bat system, Beresford and Stajano [7] correctly identified all participants by simply examining where the participants spent most of their time. Similarly, using GPS traces from vehicles, two studies by Hoh *et al.* [27] and Krumm [34] found the home addresses (and thus the identity) of most drivers. Hence, pseudonyms are not sufficient to protect the location privacy of mobile nodes.

I. Goldberg and M. Atallah (Eds.): PETS 2009, LNCS 5672, pp. 216–234, 2009.

One popular technique for achieving location privacy consists in using *multiple pseudonyms* [7,22,46] that are changed over time to impede traceability. As a pseudonym changed by an isolated node can be trivially guessed by an external party, pseudonym changes are coordinated among mobile nodes in regions called *mix zones* [8]. But even if location traces of mobile nodes are completely anonymized (i.e., do not contain any identifier), Hoh and Gruteser [25] were able to reconstruct the tracks of mobile nodes using a multiple target tracking (MTT) algorithm. Hence, to protect against the spatial correlation of location traces, location traces should also be altered *spatially*. To do this, mix zones can also conceal the trajectory of mobile nodes to the external adversary by using: (i) Silent/encrypted periods [17,28,37], (ii) a mobile proxy [42], or (iii) regions where the adversary has no coverage [12]. The effectiveness of a mix zone, in terms of the location privacy it provides, depends on the adversary's ability to relate mobile nodes that enter and exit the mix zone [7]. Hence, mix zones should be placed in locations with high node density and unpredictable mobility [8,29].

While traversing a given area, mobile nodes go through a *sequence of mix zones* and "accumulate" untraceability [12,30]. Unlike wired mix networks such as Tor [16] where packets can be freely routed, the sequence of mix zones traversed by mobile nodes depends on the mobility of each node. In other words, the flow of mobile nodes cannot be controlled to maximize location privacy. Instead, we propose to control the *placement* of mix zones to impede the adversary from tracking the nodes' location. However, similarly to the delay introduced by mix nodes on packets, mix zones induce a cost for mobile nodes: With silent mix zones, mobile nodes cannot communicate while they are in the mix zone, and with a mobile proxy, all messages have to transit through the same mobile node. Hence, the number of mix zones to be deployed over a given area must be kept small.

We consider a trusted central authority that is responsible for the establishment of security and privacy in the network (e.g., in vehicular networks, the vehicle registration authority [23]). This authority deploys a limited number of mix zones in a given area to protect the location privacy of mobiles nodes. In order to help the authority evaluate the mixing effectiveness of mix zones prior to network operation, we first propose a metric based on mobility profiles. To do so, we model the strategy of the adversary in assigning exiting to entering flows as a decision problem [9]. We propose to use the Jensen-Shannon divergence [38] to measure the probability of error of the adversary. Then, we model the problem of placing mix zones as an optimization problem: We propose an algorithm to find the optimal placement of mix zones by maximizing the mixing effectiveness of the system at an acceptable cost for mobile nodes. The algorithm offers minimum location privacy guarantees by enforcing a maximum distance between traversed mix zones. Finally, we compare the optimal mix zones deployment to other deployments by using a realistic mobility simulator [33]. To the best of our knowledge, this paper is the first to investigate deployment strategies of mix zones in mobile networks.

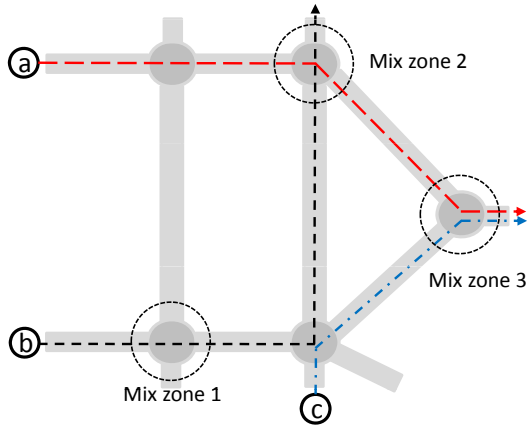

Fig. 1. Example of system model. Nodes move on plane (x, y) according to trajectories defined by flows a, b, and c. To achieve location privacy, nodes change pseudonyms in mix zones.

2 Preliminaries

2.1 System Model

We study a network where mobile nodes are autonomous entities equipped with WiFi or Bluetooth-enabled devices that communicate with each other upon coming in radio range. In other words, we consider a mobile wireless system such as a vehicular network or a network of directly communicating hand-held devices. Without loss of generality, we assume that each user in the system has a single mobile device and thus corresponds to a single node in the network.

As commonly assumed in such networks, we consider an offline central authority (CA) run by an independent trusted third party that manages, among other things, the security and privacy of the network. In vehicular networks for example, the vehicle registration authority could take this role. In line with the multiple pseudonym approach, we assume that prior to joining the network, every mobile node s registers with the CA that preloads a finite set of *pseudonyms* [40] (e.g., certified public/private key pairs, MAC addresses). Mobile nodes change pseudonyms in mix zones in order to achieve location privacy (Fig. 1). Upon changing pseudonyms, we consider for simplicity that the old pseudonym expires and is removed from the node's memory. Once a mobile node has used all its pseudonyms, it contacts the CA to obtain a new set of pseudonyms.

We assume that mobile nodes automatically exchange information (unbeknownst to their users) as soon as they are in communication range of each other. Note that our evaluation is independent of the communication protocol. Without loss of generality, we assume that mobile nodes advertise their presence by periodically broadcasting proximity beacons containing the node's identifying/authenticating information (i.e., the sender attaches its pseudonym to its

messages). Due to the broadcast nature of wireless communications, beacons enable mobile nodes to discover their neighbors. For example, when a node s receives an authenticated beacon, it controls the legitimacy of the sender by checking the certificate of the public key of the sender. After that, s verifies the signature of the beacon message.

We consider a discrete time system with initial time $t = 0$. At each time step t_s, mobile nodes can move on a plane (Fig. 1) in the considered area. As shown by Gonzalez, Hidalgo and Barabasi [19], mobile users tend to return regularly to certain locations (e.g., home and workplace), indicating that despite the diversity of their travel locations, humans follow simple reproducible patterns. Hence, we consider a *flow-based* mobility model [33]: Based on real trajectories of mobile nodes in the network (e.g., pedestrian or vehicular), we construct $f \in F$ flows of nodes in the network between the few highly frequented locations of mobile nodes, where F is the set of all flows. In practice, such real trajectories could be provided, for example, by city authorities in charge of road traffic optimization. Thus, each flow f defines a trajectory shared by several mobile nodes in the network during a period of time. For example in Fig. 1, each node is assigned to one of the three flows a, b, or c and follows the trajectory defined by the flow during the traversal of the plane. In stationary regime, a flow is characterized by its average number of nodes, λ. Note that during the course of the day, flows usually vary. For simplicity, we consider one of the possible stationary regimes of the system. Flows are defined over the road segments in the considered area. The mobility of the nodes is thus bound to the road segments.

2.2 Threat Model

An adversary $\mathcal{A}$ aims at tracking the location of some mobile nodes. In practice, the adversary can be a rogue individual, a set of malicious mobile nodes, or might even deploy its own infrastructure (e.g., by placing eavesdropping devices in the considered area). We consider that the adversary is *passive* and simply eavesdrops on communications. In the worst case, $\mathcal{A}$ obtains complete coverage and tracks mobile nodes throughout the entire area. We characterize the latter type of adversary as *global*.

$\mathcal{A}$ collects identifying information (e.g., the MAC address or the public keys used to sign messages) from the entire network and obtains *location traces* that allow him to track the location of mobile nodes. Hence, the problem we tackle in this paper consists in protecting the *location privacy* of mobile nodes, that is, to prevent other parties from learning a node's past and current location [8]. It must be noted that, at the physical layer, the wireless transceiver has a wireless fingerprint that the adversary could use to identify it [41]. However, because this requires a costly installation for the adversary and stringent conditions on the wireless medium, it remains unclear how much identifying information can be extracted in practice from the physical layer and we do not consider this threat.

3 Mix Zones

As described in the Introduction, location privacy is achieved by changing pseudonyms in regions called *mix zones* [7]. Mix zones are *effective* in anonymizing the trajectory of mobile nodes if the adversary is unable to predict with high certainty the relation between mobile nodes entering and exiting mix zones. In this section, we first give a description of mix zones and then evaluate their effectiveness using an information-theoretic divergence measure.

3.1 Mix Zones Description

A mix zone $i \in Z$ is defined by a triplet (x_i, y_i, R_i), where Z is the set of all mix zones in the considered area. The x_i and y_i coordinates are the center of the mix zone i and determine the location of the mix zone in the network. R_i is the radius of mix zone i, which we assume constant over all mix zones, $R_i = R$. In other words, a mix zone is a region of pre-determined shape and size that can be established anywhere in the considered area. We consider that the location of mix zones is determined centrally and communicated to the mobile nodes prior to their joining the network.

Each mix zone i is traversed by flows $f_j \in F_i \subseteq F$ of mobile nodes. Mobile nodes traversing a mix zone create entering and exiting *events* of the mix zone. Each node in a flow takes a certain amount of time, called the *sojourn time*, to traverse the mix zone. The sojourn time models the speed diversity of mobile nodes traversing mix zones. Speed differences are caused, for example, by a higher density of nodes on specific flows or by traffic lights. Each mix zone i has a set of entry/exit points L_i typically corresponding to the road network. Consider the example in Fig. 1: Mix zone 3 has three entry/exit points that are all traversed by some flows. Based on the flows traversing a mix zone, we can evaluate the different *trajectories* of mobile nodes in each mix zone. The *mobility profile* of a mix zone captures the typical behavior of mobile nodes traversing the mix zone (i.e., their sojourn time and trajectory). In practice, city authorities in charge of traffic lights optimization could provide the measured sojourn time distributions as well as typical trajectories over the course of the day.

There are several techniques for obtaining a mix zone: (i) Turning off the transceiver of mobile nodes [28,31,37], (ii) encrypting messages [17], (iii) relaying all wireless communications through a proxy [42], or (iv) exploiting regions where the adversary has no coverage [12]. In all cases, the adversary cannot observe the movements of the nodes within the mix zone. For example in Fig. 1, three mix zones have been established encompassing the entire intersection.

3.2 Mix Zones Effectiveness

In order to efficiently place mix zones in the network, we need to know - prior to their deployment - their mixing effectiveness. As the previously proposed entropy metric [7] depends on entering/exiting events of mix zones (after deployment), we propose a new metric based exclusively on the mobility profile of mix zones (before deployment).

Event-Based Metric. As presented by Beresford and Stajano [7] for mobile networks and by Diaz *et al.* [15] and Serjantov and Danezis [43] for mix networks, the uncertainty of the adversary (i.e. entropy) is a measure of the location privacy/anonymity achieved by a node. Assuming that $\mathcal{A}$ knows the *mobility profile* of the nodes within each mix zone, the adversary can predict their future direction from their past behavior. Consider a sequence of entering/exiting nodes traversing a mix zone i over a period of T time steps, the uncertainty of the adversary is:

$$H_T(i) = -\sum_{v}^{I} p_v \log_2(p_v) \tag{1}$$

where p_v is the probability of different assignments of entering nodes to exiting nodes and I is the total number of such hypothesized assignments. Each value p_v depends on the entering/exiting nodes and the mobility profile. In other words, the anonymity provided by mix zones mostly depends on factors beyond the control of the nodes. It is thus interesting to compute the average location privacy provided by a mix zone to evaluate its *mixing effectiveness*. The entropy measure is bound to the set of events happening in an interval of T time steps and does not capture the average mixing of a mix zone. The average mixing effectiveness of a mix zone i can be computed by taking the average entropy over n successive periods of T time steps: $E[H(i)] = \frac{1}{n}\sum_{v=1}^{n} H_{T_v}(i)$.

Flow-Based Metric. We propose a new method to theoretically evaluate the mixing effectiveness provided by mix zones. The proposed metric relies on the statistics of the mix zone, i.e., the mobility flows and the mobility profile, to compute the mixing effectiveness of the mix zone. The advantage of the proposed metric is that the mixing effectiveness can be computed prior to the operation of the mobile network as it does not rely on a particular set of events.

The metric is generic and independent of the nature of traffic. However, to simplify the treatment, we model each flow f_j as a homogeneous Poisson process with intensity λ_j. The distribution $Pois(b; \lambda_j)$ denotes the probability that b nodes enter the flow f_j during a time step t_s. Each flow f_j that traverses a mix zone i is subject to a sojourn time distribution $h_{i,j}(\Delta t)$, where Δt is the time spent in the mix zone. Observing the exit of a mix zone i, the adversary is confronted to a classical *decision-theory problem*: $\mathcal{A}$ must classify each exit event $x \in X$ happening at time t_x as coming from one of the F_i possible entering flows.

Let $m = |F_i|$ be the number of flows in mix zone i. Assume that $m = 2$ flows $\{f_1, f_2\}$ converge to the same mix zone exit l. The probability that the adversary misclassifies x depends on the number of nodes that can potentially correspond to it. This is related to the time spent in the mix zone and the inter-arrival time. We focus on a simple scenario where one mobile node from each flow enters the mix zone. Without loss of generality, we assume that the first mobile node arrives at time $t = 0$ from f_1 and that the second node arrives with a time difference δ from f_2. Figure 2 shows the exiting time probability distribution time for a given δ. We first compute the error probability with a fixed value of δ and then generalize our model by considering different values of δ.

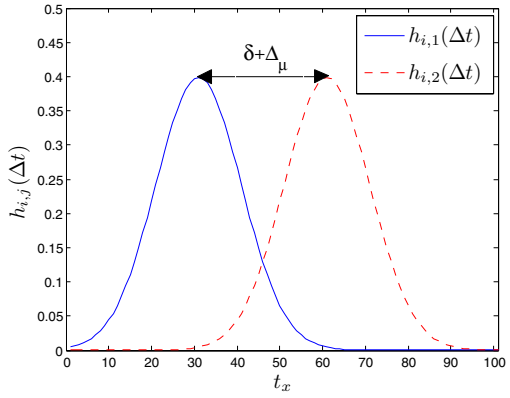

Fig. 2. Example of exiting time distribution of two flows with $h_{i,j}(\Delta t) \sim \mathcal{N}(\mu_j, \sigma_j)$, $j = 1, 2$. In this example, $(\mu_1, \sigma_1) = (2, 1)$, $(\mu_2, \sigma_2) = (4, 1)$, $\Delta_\mu = \mu_2 - \mu_1$, and δ is the arrival time difference between events of two flows (i.e., the first node arrives at time $t = 0$, and the second one arrives at time δ).

To compute the location privacy generated by a mix zone, we are interested in computing the probability that an adversary misclassifies an event. In other words, for one exit l, a successful mixing occurs whenever the adversary makes an error, i.e., assigns an exit event to the wrong flow. It is well known that the decision rule that minimizes the probability of error is the Bayes decision rule (i.e., choosing the hypothesis with the largest a posteriori probability). According to Bayes' theorem, the *a posteriori* probability that an observed event x belongs to flow f_j is

$$p(f_j|x) = \frac{p_j(x)\pi_j}{\sum_v p_v(x)\pi_v}, j = 1, 2 \tag{2}$$

where $p_j(x) = p(x|f_j)$ is the conditional probability of observing x knowing that x belongs to f_j and $\pi_j = p(f_j)$ is the *a priori* probability that an observed exit event belongs to flow f_j. The Bayes probability of error [24] is then given by:

$$p_e(p_1, p_2) = \sum_{x \in X} \min(\pi_1 p_1(x), \pi_2 p_2(x)) \tag{3}$$

The a priori probabilities depend on the intensity of the flows and are equal to: $\pi_j = \lambda_j / (\sum_{v:f_v \in F_i} \lambda_v)$. The conditional probabilities $p_1(x)$, $p_2(x)$ are equal to the probability that f_j generates an exit event at time t_x: $p_1(x) = \int_{t_x}^{t_x+t_s} h_{i,1}(t)dt$ and $p_2(x) = \int_{t_x}^{t_x+t_s} h_{i,2}(t-\delta)dt$.

A large body of research has focused on minimizing the probability of error. For example, the MTT algorithm minimizes the probability of error when tracking multiple moving objects. In the location privacy context, it is used to measure the effectiveness of path perturbation techniques by Hoh and Gruteser [25]. In our case, we evaluate the probability of error in order to find mix zones with

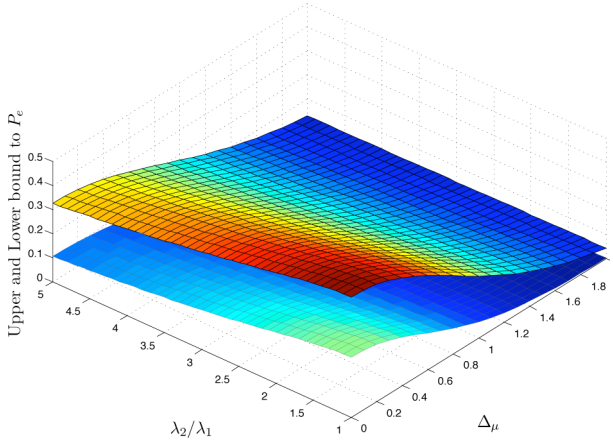

Fig. 3. Lower and upper bounds of the probability of error with $Pois(b;\lambda_j)$, $h_{i,j}(\Delta t) = \mathcal{N}(\mu_j, \sigma_j = 0.5)$, $j = 1, 2$, $\lambda_1 = 0.2$, $\lambda_2 \in [0.2, 2]$, $\mu_1 = 2s$, and $\mu_2 \in [2, 4]s$. As λ_2/λ_1 increases, the difference between the two probability functions increases as well and it becomes easier to classify the events (p_e becomes smaller). The decrease in p_e is faster if Δ_μ increases as well.

high mixing effectiveness, i.e., that maximize the probability of error. Because computing the probability of error is most of the time impractical [32] (when $m > 2$), we consider the *distance* between the two probability distributions p_1, p_2 to compute bounds on the error probability. Intuitively, the further apart these two distributions are, the smaller the probability of mistaking one for the other should be. The *Jensen-Shannon divergence* [38] (JS) is an information-theoretic distance measure that is particularly suitable for the study of decision problems as the one considered here. It provides both a lower and an upper bound for the Bayes probability of error.

$$JS_\pi(p_1, p_2) = H(\pi_1 p_1(x) + \pi_2 p_2(x)) - \pi_1 H(p_1(x)) - \pi_2 H(p_2(x)) \tag{4}$$

The JS divergence (4) provides a simple way to estimate the misclassification error of the adversary over a mix zone. The Bayes probability of error is lower/upper bounded as follows [38]:

$$\frac{1}{4}(H(\pi_1, \pi_2) - JS_\pi(p_1, p_2))^2 \leq p_e(p_1, p_2) \leq \frac{1}{2}(H(\pi_1, \pi_2) - JS_\pi(p_1, p_2)) \tag{5}$$

where $H(\pi_1, \pi_2)$ is the entropy of the a priori probabilities. The JS divergence is thus particularly useful in order to select mix zones with a high mixing effectiveness. In addition, the JS divergence can be extended to a larger number of flows [38]:

$$JS_\pi(p_1, ..., p_m) = H(\sum_{i=1}^{m} \pi_i p_i(x)) - \sum_{i=1}^{m} \pi_i H(p_i(x)) \tag{6}$$

Consider the following example: Two flows f_1, f_2 with equal input Poisson intensities $\lambda_j = 0.2$ share an exit l of mix zone i. The sojourn times are distributed according to a Normal distribution $h_{i,j}(\Delta t) = \mathcal{N}(\mu_j = 2, \sigma_j = 0.5)$, $j = 1, 2$, and $\delta = 0$. Figure 3 shows how the lower and upper bounds on the probability of error are influenced by a difference Δ_μ of the sojourn time distributions ($\Delta_\mu = \mu_2 - \mu_1$) and by the ratio λ_2/λ_1 of flows' intensities. We observe that if Δ_μ increases and $\lambda_2/\lambda_1 = 1$, p_e decreases, showing that, with a fixed δ, a difference in the sojourn time distributions alone helps distinguish between the two distributions. We also observe that if λ_2/λ_1 increases and $\Delta_\mu = 0$, the probability of error decreases. The intuition is that as the difference between the flows' intensities increases, the flow with higher intensity dominates the exit of the considered mix zone. In addition, we observe that if both λ_2/λ_1 and Δ_μ increase, p_e decreases faster. The mixing effectiveness is maximal when both flows have the same intensity and sojourn time distribution.

Until now, we focused on scenarios with one mobile node entering from each flow, and a fixed δ. We generalize our model by considering the average difference in arrival time of nodes in flows. More specifically, based on the average arrival rate λ_j, we compute the average difference in arrival time between flows and the average number of nodes that can potentially correspond to an exit event x. The average difference in arrival time between any two flows depends on the flow intensities. The average number of nodes that can be confused with an event x depends on the maximum sojourn time window $\omega_{i,l} = \max_{f_j \in F_{i,l}}(\Delta t_{f_j})$, where Δt_{f_j} is the time spent in the mix zone by nodes in flow f_j and $F_{i,l}$ is the set of flows in F_i that exit at l. For each flow $f_j \in F_{i,l}$, there is a set of possible entering events with average arrival time differences in a time window $\omega_{i,l}$ with respect to beginning of the window: $\zeta^i_{j,l} = \{\delta_{j,v}\colon v/\lambda_j \leq \omega_{i,l}, v \in \mathbb{N}\}$, where $\delta_{j,v} = v/\lambda_j$. We compute the probability of error of the adversary at exit l as follows:

$$p^i_{e,l} = \frac{\sum\limits_{f_j \in F_{i,l}} p_e\Big(p_j(x,0), p_{\kappa_1}(x, \delta_{\kappa_1,v_1}), p_{\kappa_1}(x, \delta_{\kappa_1,v_2}), ..., p_{\kappa_2}(x, \delta_{\kappa_2,v_1}), ...\Big)}{|F_{i,l}|} \quad (7)$$

where $p_j(x,0)$ is the conditional probability $p_j(x)$ with $\delta = 0$, $p_{\kappa_1}(x, \delta_{\kappa_1,v_1})$ corresponds to the conditional probability $p_{\kappa_1}(x)$ with $\delta_{\kappa_1,v_1} \in \zeta^i_{\kappa_1,l}$, and κ_1, κ_2, ..., κ_{m-1} are not equal to j. In other words, we evaluate the confusion of the adversary for each flow with respect to other flows. Finally, we compute the average probability of error caused by a mix zone i by considering the error created by each exit $l \in L_i$ of mix zone i:

$$\bar{p}^i_e = \frac{\sum_{L_i} p^i_{e,l}}{|L_i|} \quad (8)$$

With this model, we consider the average arrival rate of the nodes and can thus compute the mixing effectiveness prior to network operation. Note that we assumed for simplicity that the sojourn time distribution is independent of the flows' intensity. The model can be extended to capture the interactions between nodes in the mix zone and their effect on the sojourn time distributions [18].

4 Placement of Mix Zones

In principle, mix zones can be placed anywhere in the considered area. Their placement determines the accumulated location privacy provided by each mix zone. Thus, the optimal solution consists in placing mix zones on the entire surface of the considered area. However, mix zones have a cost because they impose limits on the services available to mobile users and require a pseudonym change. Hence, the total number of mix zones deployed in the network should be limited to minimize the disruptions caused to mobile nodes. We assume that a central authority, responsible for the establishment of security and privacy in the system, is confronted with the problem of organizing mix zones in the network. Thus, users must trust that the central authority will protect their privacy. We propose a solution based on combinatorial optimization techniques that relies on the divergence metric introduced in Sect. 3 to select appropriate mix zones. Our paper, by making a possible algorithm public, increases the trustworthiness of the authority as it provides a basis for comparison.

4.1 Mix Zones Placement

After Chaum's seminal work on *mixes* [13], there have been multiple proposals on the way mixes should be connected and organized to maximize the provided anonymity [11]. This led to a classification of different organization concepts. For example, the choice of the sequence of mixes is either distributed (i.e., *mix networks*) or centrally controlled (i.e., *mix cascades*).

The system considered in this paper, namely mix zones deployed over a considered area, presents three different characteristics: (i) The organization of mixes depends on the placement of mix zones in the area, (ii) mobile nodes move in the considered area according to flows constrained by the underlying road network, and (iii) the road network is a connected network with a restricted number of routes. Hence, we must characterize mix zones placements that maximize the achievable location privacy.

In order to evaluate the location privacy provided by mix zones deployed over a mobile network, one solution consists in computing the uncertainty accumulated by the adversary with the joint entropy [43]. However, the complexity of the formulation increases as the number of mix zones increases, making it hard to evaluate. Instead, to compute the overall location privacy, we maximize the total probability of error of the adversary by considering the sum of error probabilities over each deployed mix zone and we guarantee that the distance over which the adversary can successfully track mobile nodes is upper-bounded, i.e., the average *distance-to-confusion* (dtc). A mix zone is a *confusion point* if the error probability of the adversary is larger than a given threshold θ [26].

However, mix zones induce a cost on mobile nodes that must be taken into account in the mix zone deployment phase. The cost associated to each mix zone depends on the considered application. For example, with silent periods, the cost is typically directly proportional to the duration of the imposed silent period (i.e., the size of the mix zone). Similarly, the cost also depends on the

number of used pseudonyms. Pseudonyms are costly to use because they are a limited resource that requires contacting the CA for refill.

4.2 Placement Optimization

In this section, we model the problem of mix zones placement as an optimization problem. Formally, consider a finite set Z of all possible mix zones' locations, a set F of mobility flows in the system, and a mobility profile for each potential mix zone in the considered area. The goal is to optimize the placement of mix zones to maximize the overall probability of error of an adversary tracking mobile nodes in the considered area while respecting the cost and distance-to-confusion constraints. We select a subset $\widehat{Z} \subseteq Z$ of *active* mix zones, which is a solution of the following combinatorial optimization problem:

$$\max_{\widehat{Z}} \sum_{i \in Z} \bar{p}_e^i \cdot z_i \tag{9}$$

$$\text{subject to } \sum_{i \in f_j} w_i z_i \leq W_{max}, \forall f_j \tag{10}$$

$$E[dtc(f_j, \widehat{Z})] \leq C_{max}, \forall f_j \tag{11}$$

where $z_i \in \{0,1\}, \forall i \in Z$ indicates if a mix zone is active (i.e., $z_i = 1$), $\widehat{Z}$ is the set of active mix zones, $\bar{p}_e^i$ captures the error introduced by mix zone i, w_i is the cost associated with mix zone i, W_{max} is the maximum tolerable cost, $E[dtc(f_j, \widehat{Z})]$ is the average distance-to-confusion of flow f_j with the set of active mix zones $\widehat{Z}$, and C_{max} is the maximum tolerable distance-to-confusion. We compute the probability of error $\bar{p}_e^i$ by using the lower bound obtained with the Jensen-Shannon divergence in the previous section. The first constraint limits the number of mix zones that can be deployed per flow by taking into account the cost associated with each mix zone. The second constraint ensures that the average distance-to-confusion is upper bounded, i.e., C_{max} defines a maximal distance over which mobile nodes can be tracked on average.

5 Application Example

To test the relevance of our approach, we implemented a simulator in Java that evaluates the tracking efficiency of the adversary.[1] The simulator takes as input a mobility trace on a map and a set of locations for mix zones. It first computes the mobility profile of mix zones and then attempts to predict the trajectory of mobile nodes.

5.1 Simulation Setup

We simulate mobility traces with Sumo [33], a urban mobility simulator, over a cropped map [5] of Manhattan of 6 km^2. Sumo features the creation of routes

[1] The code is available at: http://icapeople.epfl.ch/freudiger

for mobile nodes using mobility flows: Each flow is defined by a source, a destination and a traffic intensity. Each mobile node belongs to a single flow and is routed from source to destination over the shortest path. Roads have one lane in each direction, and road intersections are modeled with yields. Some roads (e.g., highways) have higher priority and do not have to yield.

In this application example, the constraints of the optimization algorithm are defined as follows. The cost of mix zones w_i is proportional to the cost of a pseudonym change γ. We assume that the cost of a pseudonym change is fixed and the same for all nodes, $\gamma = 1$. We set $W_{max} = 3$, meaning that each node can traverse a maximum of three mix zones. Similarly, we set $C_{max} = 2000$m, i.e., the adversary cannot track nodes over more than two kilometers. A total of 40 flows were deployed over the area, generating 1210 nodes in a fluid scenario ($\lambda_j \sim 0.02$) and 2000 nodes in a congested scenario ($\lambda_j \sim 0.04$). The radius of mix zones is a constant $R = 100$m. We simulate a mobile network for 20 minutes with nodes moving at a maximum speed of 50km/h and with an average trip time of 6 minutes. Finally, a mix zone is considered as a confusion point if the introduced error is larger than zero, i.e., $\theta = 0$.

Mobility Profiles. We consider a powerful (worst-case) adversary that can construct a mobility profile of each mix zone i by measuring the time at which nodes enter/exit mix zones. We denote with Q the measuring precision of the adversary, and assume $Q = 1$ second. Hence, $\mathcal{A}$ knows for each mix zone: (i) The distribution of nodes' trajectories, and (ii) the sojourn time distributions. The distribution of nodes' trajectories is captured in a matrix of directions D_i: For each entering/exiting points (k, l), the matrix contains the probability of the trajectory: $D_i^{k,l} = Pr(\text{“ Enter at } k \text{ and exit at } l \text{ ”})$. The sojourn time distribution is captured in a matrix of sojourn times J_i: For each entering/exiting points (k, l), the matrix contains the probability distribution of the sojourn time: $J_i^{k,l}(\Delta t) = Pr(\text{“Enter at } k \text{ and spend } \Delta t \text{ before exiting at } l\text{”})$.

Attack. Based on the mobility profiles, the adversary $\mathcal{A}$ predicts the most probable assignment of entering/exiting mobile nodes for each mix zone. To do so, the attacker can model entering/exiting events with a weighted bipartite graph as suggested by Beresford in [6]. Each edge is weighted according to the a priori probability of linking an exiting event at l to an entering event at k: $D_i^{k,l} \cdot J_i^{k,l}(\Delta t)$. Then, the maximum weight matching of the bipartite graph corresponds to the optimal guess of the adversary. As discussed in [45], a more elaborate attack consists in computing all perfect matchings of the bipartite graph to weight edges, according to the a posteriori probability of linking entering/exiting events. However, this attack has a large complexity, increasing exponentially with the number of entering/exiting pairs and its scalability remains an open problem.

Metrics. Assume that Z_s is the set of mix zones traversed by node s and let $G_s \subseteq Z_s$ be the set of mix zones successfully matched by the adversary. $\mathcal{A}$ is *successful* in tracking the location of node s in a mix zone if the real trajectory

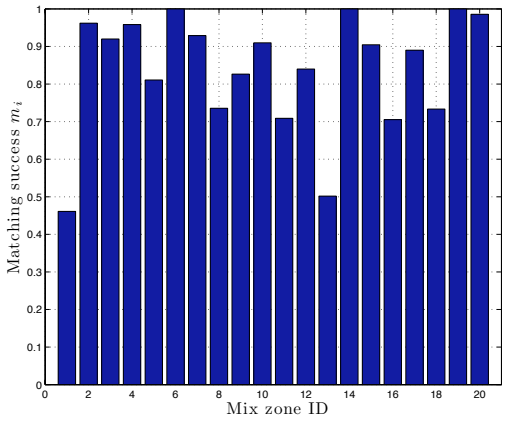

Fig. 4. Matching success m_i of the 20 potential mix zone locations

of node s is correctly guessed. For example, $G_s = \{z_3, z_5, z_{10}\}$ means that node s was successfully tracked in three mix zones.

For each mix zone i, the mixing effectiveness is $m_i = \frac{u_i}{N_i}$ where u_i is the number of successful matches in mix zone i and N_i is the total number of nodes that entered mix zone i over the course of the simulation. This metric reflects the mixing effectiveness of mix zones. The tracking success of the adversary is defined as the percentage of nodes that can be tracked over k consecutive mix zones: $ts(k) = \frac{N_{suc}(k)}{N(k)}$, where $N_{suc}(k)$ is the number of nodes successfully tracked over k consecutive mix zones, and $N(k)$ is the total number of nodes traversing k consecutive mix zones. This metric reflects the distance over which nodes can be tracked before confusing the adversary.

5.2 Results

Mix Zone Performance. Figure 4 shows the histogram of mixing effectiveness for the 20 potential mix zone locations. We observe that the mixing effectiveness can vary significantly across mix zones and hence some nodes might experience a poor mixing while traversing a mix zone. This affects the optimal deployment, because mix zones with a low mixing effectiveness are sometimes chosen to fulfill the distance-to-confusion constraint. Other than that, the optimization algorithm will tend to choose mix zones that offer the lowest tracking success to the adversary, e.g., mix zones 1 and 13 are particularly effective.

Mix Zone Placement. We consider a total of 20 possible mix zone locations and test four deployments of mix zones: (i) The *optimal* mix zone deployment computed according to Sect. 4.2 resulting in 6 deployed mix zones, (ii) a *random* mix zone deployment of 10 mix zones selected uniformly at random, (iii) a *bad* mix zone deployment of 6 mix zones with poor mixing effectiveness, and (iv)

Table 1. Percentage of mobile nodes traversing a certain number of mix zones for various mix zone deployments. The avg column gives the average number of traversed mix zones. The last column gives the percentage of nodes that were successfully tracked over all mix zones in the considered area.

# of traversed mix zones	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**avg**	**Tracked (%)**
Bad (6 mix zones)	68	20	7	5	0	0	0	0	0	0.48	98
Random (10 mix zones)	14	43	24	10	9	0	0	0	0	1.56	78
Optimal (6 mix zones)	14	33	37	16	0	0	0	0	0	1.55	53
Full (20 mix zones)	0	8	24	24	16	14	8	4	2	3.56	48

a *full* mix zone deployment where the 20 mix zones are in use. We observe in Table 1 that in the optimal deployment, the majority of the nodes traverses at least one mix zone and none exceeds the tolerable cost of three mix zones. The random and optimal deployment perform relatively close in terms of the number of traversed mix zones, but with the optimal deployment, less nodes are tracked (53%) approaching the performance of the full deployment (48%). As expected, the bad mix zone deployment performs the worst.

The average number of traversed mix zones in Table 1 also reflects the total cost. We observe that the optimal deployment has a higher cost than the bad deployment for the same number of deployed mix zones. However, compared to the full deployment, the optimal deployment achieves a tolerable cost and approaches the same mixing effectiveness.

Tracking Success. We compare the tracking success of the adversary for the optimal, random, bad and full deployment of mix zones. We observe in Fig. 5 (a) that in general the probability of success of the adversary decreases as mobile nodes traverse more mix zones. The optimal deployment of mix zones is more effective at anonymizing flows than other deployments and complies with the cost constraint. In particular, the optimal deployment is superior to the full deployment because it avoids the bad placement of mix zones.

Note that in the case of the full deployment, traversing more mix zones does not necessarily increase (and actually decreases) the location privacy. The reason is that the majority of the flows traversing more than five mix zones actually go through a sequence of ineffective mix zones. Hence, all flows are not equal in terms of the achievable location privacy.

In Fig. 5 (b), we observe the effect of an increase in the flow intensity λ_j (leading to a congested scenario). The optimal deployment is not affected by the change of intensity because it places mix zones in regions with high traffic density anyway. The random deployment significantly improves its mixing effectiveness and approaches the performance of the optimal deployment.

In Fig. 5 (c), we observe that as the tracking precision Q of the adversary diminishes, so does its ability to track nodes. A reduction of the tracking precision of the adversary reflects scenarios where the knowledge of the adversary about mobility profiles is noisy.

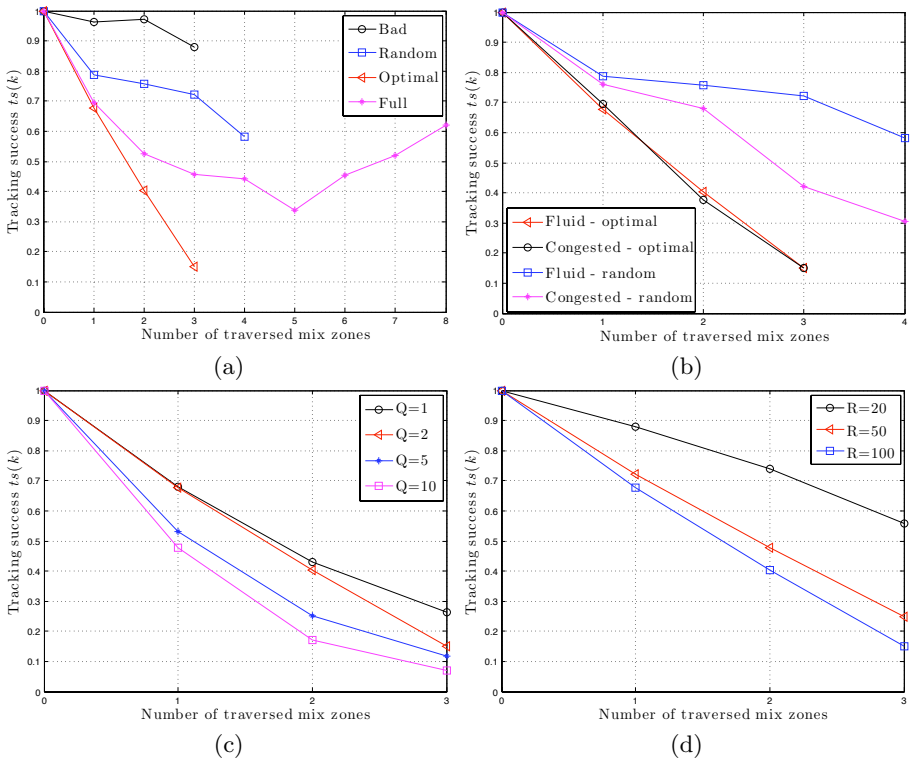

Fig. 5. Tracking success of adversary $ts(k)$, i.e., the fraction of nodes that can be tracked over k consecutive mix zones. (a) Tracking success for various mix zones' deployments. (b) Tracking success in a fluid and congested scenario. (c) Tracking success with various adversary's precision. (d) Tracking success for various sizes of mix zone.

In Fig. 5 (d), we observe that increasing the mix zone radius R from 50 to 100 does not increase much the mixing effectiveness, whereas a small radius $R = 20$ dramatically reduces the achieved location privacy. One reason is that changes in speed and direction occur mostly at the center of mix zones. Another reason is that with $R = 20$, the size of mix zones tends to be smaller than the size of crossroads of the considered map. On one hand, it is thus important to choose mix zones that are not too small. On another hand, large mix zones are inappropriate because they do not significantly increase location privacy and have a high cost.

We also vary the parameters of the optimization problem. The cost w_i, associated with mix zones, changes the optimal placement of mix zones. As we increase the cost, fewer mix zones are deployed and the achievable location privacy decreases compared to the full deployment. Instead, if the tolerable cost increases, the optimal deployment performs closer to the full deployment in terms of the

achieved location privacy. Finally, if the tolerable distance-to-confusion is lowered, the optimization problem might not have a solution. If there is a solution, it will require more mix zones and will increase the cost per node.

Discussion. Our results show the limitations of mix zones, but also exhibit the importance of optimizing their placement. In particular, the optimal deployment prevents bad placement of mix zones. Another interesting result is that traversing more mix zones is not necessarily an advantage. It must be noted that the relatively high success rate of the adversary is also due to the application example. First, we consider a worst-case adversary with global coverage and access to precise mobility profiles ($Q = 1$). Second, we consider a relatively small map with a simple road intersection model.

6 Related Work

There are several techniques for achieving location privacy besides the multiple pseudonyms approach [36]. Mobile nodes can also intentionally add noise to their location [21], or report their location as a region instead of a point [44]. However, in mobile wireless networks, the peer-to-peer wireless communications between mobile nodes unveil their locations. Hence, obfuscating the location data contained in messages is insufficient to protect the location privacy of mobile nodes. In other words, the use of multiple pseudonyms is required for achieving location privacy in such networks. To anonymize pseudonyms such as the MAC address, one approach [22] consists in changing the MAC address over time between connections with WiFi access points. Another possibility [20] is to obscure the MAC address and use an identifier-free link layer protocol. However, in peer-to-peer wireless networks, mobile nodes continuously broadcast messages and cannot be anonymized only with respect to WiFi access points. Similarly, mobile nodes must be identifiable on several layers of the protocol stack. Hence, we propose to change pseudonyms in optimally placed mix zones.

Huang *et al.* suggest in [30] the use of cascading mix zones. Mix zones are created by repeatedly turning off the transceivers of mobile nodes. They evaluate the quality of service implications on real-time applications of users traversing several mix zones, but do not evaluate strategies of mix zones deployments. In [12], Buttyan *et al.* evaluate the performance of sequences of mix zones for vehicular networks. The locations of mix zones correspond to regions where the adversary has no coverage. In their system, the adversary has a high tracking success because of the insufficient mixing of vehicles. In this paper, we provide a theoretical framework for the analysis of the mixing effectiveness of mix zones and of their optimal placement in a considered area.

Note that in wired mix networks, the disadvantages of free routes were studied in [10,14] showing the importance of route selection and network connectivity. In this paper, we study an equivalent problem for mobile networks considering the optimal positioning of mix zones and its effect on the achievable location privacy.

7 Conclusion

We have considered the problem of constructing a network of mix zones in a mobile network. We first showed how to evaluate the mixing effectiveness of mix zones prior to network operation by using the Jensen-Shannon divergence measure. The proposed metric relies on statistical information about the mobility of nodes in mix zones. Then, we modeled the problem of placing mix zones as an optimization problem by taking into account the distance-to-confusion and the cost induced by mix zones on mobile nodes. By means of simulations, we investigated the importance of the mix zone deployment strategy and observed that the optimal algorithm prevents bad placement of mix zones. In addition, we measured the benefit brought by the optimal placement of mix zones, i.e., a 30% increase of location privacy compared to a random deployment of mix zones, in our considered example. We also noticed that the optimal mix zone placement performs comparatively well to the full deployment scenario, but at a lower cost. This work is a first step towards a deeper understanding of the advantages and limitations of mix zones.

Future Work. We intend to extend the simulations by using real mobility traces. In order to allow for location privacy at specific locations (i.e., nodes might want to hide the fact that they traversed a particular location), we also plan to weigh the importance of specific locations in the placement strategy. Finally, it would be interesting to consider other attacks [35,39] and how an active adversary would affect the performance of the system.

Acknowledgements

We would like to thank Mathias Humbert, Maxim Raya, Marcin Poturalski, and Michal Piorkowski for their insights and suggestions on earlier versions of this work, and the anonymous reviewers for their helpful feedback. Special thanks go to Carmela Troncoso and Claudia Diaz for shepherding the paper.

References

1. http://en.wikipedia.org/wiki/Bluedating
2. http://www.aka-aki.com
3. http://csg.ethz.ch/research/projects/Blue_star
4. http://reality.media.mit/serendipity.php
5. TIGER maps, http://www.census.gov/geo/www/tiger/
6. Beresford, A.R.: Location privacy in ubiquitous computing. In: Ph.D. Thesis (2005)
7. Beresford, A.R., Stajano, F.: Location privacy in pervasive computing. IEEE Pervasive Computing 2(1), 46–55 (2003)
8. Beresford, A.R., Stajano, F.: Mix zones: user privacy in location-aware services. In: Pervasive Computing and Communications Workshops, pp. 127–131 (2004)
9. Berger, J.O.: Statistical Decision Theory and Bayesian Analysis. Springer, Heidelberg (1993)

10. Berthold, O., Pfitzmann, A., Standtke, R.: The disadvantages of free MIX routes and how to overcome them. In: Federrath, H. (ed.) Designing Privacy Enhancing Technologies. LNCS, vol. 2009, pp. 30–45. Springer, Heidelberg (2001)
11. Bohme, R., Danezis, G., Diaz, C., Kopsell, S., Pfitzmann, A.: Mix cascades vs. peer-to-peer: Is one concept superior? In: PET (2004)
12. Buttyán, L., Holczer, T., Vajda, I.: On the effectiveness of changing pseudonyms to provide location privacy in VANETs. In: Stajano, F., Meadows, C., Capkun, S., Moore, T. (eds.) ESAS 2007. LNCS, vol. 4572, pp. 129–141. Springer, Heidelberg (2007)
13. Chaum, D.: Untraceable electronic mail, return addresses, and digital pseudonyms. Communications of the ACM 24(2), 84–90 (1981)
14. Danezis, G.: Mix-networks with restricted routes. In: Dingledine, R. (ed.) PET 2003. LNCS, vol. 2760, pp. 1–17. Springer, Heidelberg (2003)
15. Díaz, C., Seys, S., Claessens, J., Preneel, B.: Towards measuring anonymity. In: Dingledine, R., Syverson, P.F. (eds.) PET 2002. LNCS, vol. 2482, pp. 54–68. Springer, Heidelberg (2003)
16. Dingledine, R., Mathewson, N., Syverson, P.: Tor: the second-generation onion router. In: USENIX Security Symposium, pp. 21–21 (2004)
17. Freudiger, J., Raya, M., Felegyhazi, M., Papadimitratos, P., Hubaux, J.-P.: Mix zones for location privacy in vehicular networks. In: WiN-ITS (2007)
18. Gazis, D.C.: Traffic Theory. Kluwer Academic Publishers, Dordrecht (2002)
19. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.-L.: Understanding individual human mobility patterns. Nature 453(7196), 779–782 (2008)
20. Greenstein, B., McCoy, D., Pang, J., Kohno, T., Seshan, S., Wetherall, D.: Improving wireless privacy with an identifier-free link layer protocol. In: MobiSys, pp. 40–53 (2008)
21. Gruteser, M., Grunwald, D.: Anonymous usage of location-based services through spatial and temporal cloaking. In: MobiSys, pp. 31–42 (2003)
22. Gruteser, M., Grunwald, D.: Enhancing location privacy in wireless LAN through disposable interface identifiers: a quantitative analysis. Mobile Networks and Applications 10(3), 315–325 (2005)
23. Hartenstein, H., Laberteaux, K.: A tutorial survey on vehicular ad hoc networks. IEEE Communications Magazine 46(6) (June 2008)
24. Hellman, M., Raviv, J.: Probability of error, equivocation, and the Chernoff bound. IEEE Transactions on Information Theory 16(4), 368–372 (1970)
25. Hoh, B., Gruteser, M.: Protecting location privacy through path confusion. In: SECURECOMM, pp. 194–205 (2005)
26. Hoh, B., Gruteser, M., Herring, R., Ban, J., Work, D., Herrera, J.-C., Bayen, A.M., Annavaram, M., Jacobson, Q.: Virtual trip lines for distributed privacy-preserving traffic monitoring. In: MobiSys, pp. 15–28 (2008)
27. Hoh, B., Gruteser, M., Xiong, H., Alrabady, A.: Enhancing security and privacy in traffic-monitoring systems. IEEE Pervasive Computing 5(4), 38–46 (2006)
28. Huang, L., Matsuura, K., Yamane, H., Sezaki, K.: Enhancing wireless location privacy using silent period. In: WCNC, pp. 1187–1192 (2005)
29. Huang, L., Yamane, H., Matsuura, K., Sezaki, K.: Towards modeling wireless location privacy. In: Danezis, G., Martin, D. (eds.) PET 2005. LNCS, vol. 3856, pp. 59–77. Springer, Heidelberg (2006)
30. Huang, L., Yamane, H., Matsuura, K., Sezaki, K.: Silent cascade: Enhancing location privacy without communication QoS degradation. In: Clark, J.A., Paige, R.F., Polack, F.A.C., Brooke, P.J. (eds.) SPC 2006. LNCS, vol. 3934, pp. 165–180. Springer, Heidelberg (2006)

31. Jiang, T., Wang, H.J., Hu, Y.-C.: Preserving location privacy in wireless LANs. In: MobiSys, pp. 246–257 (2007)
32. Kailath, T.: The divergence and Bhattacharyya distance measures in signal selection. IEEE Transactions on Communication Technology 15(1), 52–60 (1967)
33. Krajzewicz, D., Hertkorn, G., Rossel, C., Wagner, P.: SUMO (Simulation of Urban MObility) - an open-source traffic simulation. In: MESM (2002)
34. Krumm, J.: Inference attacks on location tracks. In: LaMarca, A., Langheinrich, M., Truong, K.N. (eds.) Pervasive 2007. LNCS, vol. 4480, pp. 127–143. Springer, Heidelberg (2007)
35. Krumm, J.: A Markov model for driver route prediction. In: SAE World Congress (2008)
36. Krumm, J.: A survey of computational location privacy. In: Personal and Ubiquitous Computing (2008)
37. Li, M., Sampigethaya, K., Huang, L., Poovendran, R.: Swing & swap: user-centric approaches towards maximizing location privacy. In: WPES, pp. 19–28 (2006)
38. Lin, J.: Divergence measures based on the Shannon entropy. IEEE Transactions on Information theory 37, 145–151 (1991)
39. De Mulder, Y., Danezis, G., Batina, L., Preneel, B.: Identification via location-profiling in GSM networks. In: WPES, pp. 23–32 (2008)
40. Pfitzmann, A., Köhntopp, M.: Anonymity, unobservability, and pseudonymity – a proposal for terminology. In: Designing Privacy Enhancing Technologies, pp. 1–9 (2001)
41. Rasmussen, B., Capkun, S.: Implications of radio fingerprinting on the security of sensor networks. In: SECURECOMM, pp. 331–340 (2007)
42. Sampigethaya, K., Huang, L., Li, M., Poovendran, R., Matsuura, K., Sezaki, K.: CARAVAN: Providing location privacy for VANET. In: ESCAR (2005)
43. Serjantov, A., Danezis, G.: Towards an information theoretic metric for anonymity. In: Dingledine, R., Syverson, P.F. (eds.) PET 2002. LNCS, vol. 2482, pp. 41–53. Springer, Heidelberg (2003)
44. Sweeney, L.: k-anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, 557–570 (2002)
45. Tóth, G., Hornák, Z.: Measuring anonymity in a non-adaptive, real-time system. In: Martin, D., Serjantov, A. (eds.) PET 2004. LNCS, vol. 3424, pp. 226–241. Springer, Heidelberg (2005)
46. Wong, F.-L., Stajano, F.: Location privacy in Bluetooth. In: Molva, R., Tsudik, G., Westhoff, D. (eds.) ESAS 2005. LNCS, vol. 3813, pp. 176–188. Springer, Heidelberg (2005)
47. Xu, Q., Mak, T., Ko, J., Sengupta, R.: Vehicle-to-vehicle safety messaging in DSRC. In: VANET, pp. 19–28 (2004)