Attack-resilient Mix-zones over Road Networks: Architecture and Algorithms

Balaji Palanisamy, *Member, IEEE,* Ling Liu, *Senior Member, IEEE*

Abstract—Continuous exposure of location information, even with spatially cloaked resolution, may lead to breaches of location privacy due to statistics-based inference attacks. An alternative and complementary approach to spatial cloaking based location anonymization is to break the continuity of location exposure by introducing techniques, such as mix-zones, where no application can trace user movements. Several factors impact on the effectiveness of mix-zone approach, such as user population, mix-zone geometry, location sensing rate and spatial resolution, as well as spatial and temporal constraints on user movement patterns. However, most of the existing mix-zone proposals fail to provide effective mix-zone construction and placement algorithms that are resilient to timing and transition attacks. This paper presents MobiMix, a road network based mix-zone framework to protect location privacy of mobile users traveling on road networks. It makes three original contributions. First, we provide the formal analysis on the vulnerabilities of directly applying theoretical rectangle mix-zones to road networks in terms of anonymization effectiveness and resilience to timing and transition attacks. Second, we develop a suite of road network mix-zone construction methods that effectively consider the above mentioned factors to provide higher level of resilience to timing and transition attacks, and yield a specified lower-bound on the level of anonymity. Third, we present a set of mix-zone placement algorithms that identify the best set of road intersections for mix-zone placement considering the road network topology, user mobility patterns and road characteristics. We evaluate the MobiMix approach through extensive experiments conducted on traces produced by GTMobiSim on different scales of geographic maps. Our experiments show that MobiMix offers high level of anonymity and high level of resilience to timing and transition attacks, compared to existing mix-zone approaches.

Index Terms—Location Privacy, Mix-zone, Location-based Applications, k-anonymity.

1 Introduction

We are entering an era where people and vehicles are being connected and tracked continuously and automatically. Such location tracking, on one hand, can offer many life-enriching experiences and services to mobile users, and on the other hand, open doors to exposure of enormous amount of potentially sensitive information, leading to the intrusion of location privacy. We can classify location privacy research into three broad categories.

The first category is policy-based solutions that restrict access through privacy policies. Such policies typically provide an option for users to turn off location based services or to refuse being tracked [2].

The second category is represented by location *k-anonymization* techniques which compute a spatially cloaked location region that has k mobile users inside it. This approach degrades the resolution of location information in a controlled fashion to ensure location *k-anonymity*. A subject is considered *k-anonymous* if its location is indistinguishable from that of $k-1$ other users [4], [14], [21]. Location $k-anonymization$ approaches are targeted at the applications that do not require true identity or pseudo-identity of mobile users, such as finding nearby gas-stations or restaurants, and notifying the sale price of items of interest when a user passes a shopping mall. However, location *k-anonymization* techniques are ineffective when the location based services require identity or pseudo-identity of users, such as accessing subscribed content (songs, audios) or sending a printing request while on the move. This is because when identity or pseudo-identity of users are associated with the publication of spatially cloaked regions, the continuous exposure of location information combined with the persistent identity or pseudo-identity can lead to the breach of location privacy due to statistics-based inference attacks [22].

The third category of location privacy research is embodied by mix-zone based approaches. Mix-zones are regions in space where a set of users enter, change pseudonyms and exit in a way such that the mapping between their old and new pseudonyms is not revealed [5], [11], [12], [13]. In contrast to controlling the resolutions of locations used in spatial cloaking based location privacy solutions, mix-zones protect location privacy by changing pseudonyms at selective locations such that it is very hard to link new pseudonyms with old pseudonyms. Thus, the frequent changing of users' pseudonyms through setting up mix-zones in selected locations can protect location privacy by effectively breaking the association of users' pseudonym with a sequence of location exposures [5]. Mix-zones are location privacy solutions that are effective for the LBSs that require identity or pseudonym of users.

The research presented in this paper falls into the third category. Most of the existing mix-zone proposals are straightforward application of theoretical mix-zones [5] to road network environments. We argue that these approaches are vulnerable to both timing and transition attacks. Concretely, theoretic mix-zones are constructed independently of the spatially constrained road networks with the assump-

• *Balaji Palanisamy is with the School of Information Sciences, University of Pittsburgh and Ling Liu is with the College of Computing, Georgia Institute of Technology.*
E-mail: bpalan@pitt.edu, lingliu@cc.gatech.edu

tion that there are infinite entry points and exit points in a mix-zone, thus it is hard to link old pseudonym to new pseudonym as long as k users enter and exit the mix-zone at the same time. However, such assumption is no longer true for mobile users in the real world, because in reality people travel in spatially constrained networks or walk-paths. Thus the number of entry points and exit points for a given mix-zone is finite and often limited. An adversary can utilize the timing information of users' entry into and exit from a mix-zone and the non-uniformity in the transitions taken at the road intersections to guess the mapping between the old and new pseudonyms [11].

In this paper we present MobiMix, a road network based Mix-zone framework to protect location privacy of mobile users. Compared to the existing approaches, the MobiMix mix-zones have a number of unique features. First, the MobiMix mix-zones are developed based on a formal study of the assumptions of the theoretic mix-zone model and the detrimental effect on pseudonym anonymity when certain assumptions are violated. We argue that effective mix-zones should be constructed and placed by taking into consideration of both road network characteristics and motion behavior of mobile users. Second, we introduce an adversary model that launches attacks based on the road network characteristics and associated motion behavior and present the MobiMix Mix-zone model for constructing road network aware mix-zones that are robust against timing attacks and transition attacks. Third, we develop a suite of attack resilient mix-zone construction and placement techniques that guarantee unlinkability between the old and new pseudonyms. Our algorithms take into account multiple factors in constructing mix-zones, such as the road network characteristics, the timing and the transitioning probability of users in terms of their movement trajectory. We formally analyze and experimentally validate the robustness of our MobiMix approach against timing attacks and transition attacks.

The rest of the paper is organized as follows: Section 2 describes the constraints and anonymity of ideal mix-zones. Section 3 discusses the characteristics of road networks, the challenges of constructing mix-zones on road networks and the MobiMix road network mix-zone model. We introduce our attack-resilient mix-zone construction techniques and present a detailed analysis of the timing and transition attacks in Section 4. Section 5 presents the MobiMix placement algorithms for deploying mix-zones on a road network. Section 6 evaluates MobiMix and its algorithms through extensive experiments conducted on traces from GTMobiSim [19] using different scales of geographic maps. We review the related work in Section 7 and conclude the paper in Section 8.

2 Analysis of Theoretical Mix-zones

Theoretical mix-zones are ideal mix-zones that provide the maximum possible anonymity to the participating users by ensuring a set of properties. Informally, a mix-zone of k participants refers to as a k-anonymization region in which a set of k users enter in some order and change pseudonyms but none leave before all k users enter the mix-zone. These k users exit the mix-zone in an order different from their order of arrival, providing unlinkability between their entering and exiting events. Formally, a theoretical mix-zone is defined as follows:

Definition 1: A mix-zone Z is said to offer k-anonymity for a set A of users iff

1) The set A has k or more members, i.e., $|A| \geq k$.
2) All users in A must enter the mix-zone Z before any user $i \in A$ exits. Thus, there exists a point in time where all k users of A are inside the zone.
3) Each user $i \in A$, entering the mix-zone Z through an entry point $e_i \in E$ and leaving at an exit point $o_i \in O$, spends a completely random duration of time inside.
4) The probability of transition between any point of entry to any point of exit follows a uniform distribution. i.e., an user entering through an entry point, $e \in E$, is equally likely to exit in any of the exit points, $o \in O$.

Figure 1(a) shows a mix-zone of three participants, a, b and c exiting with new pseudonyms p, q and r.

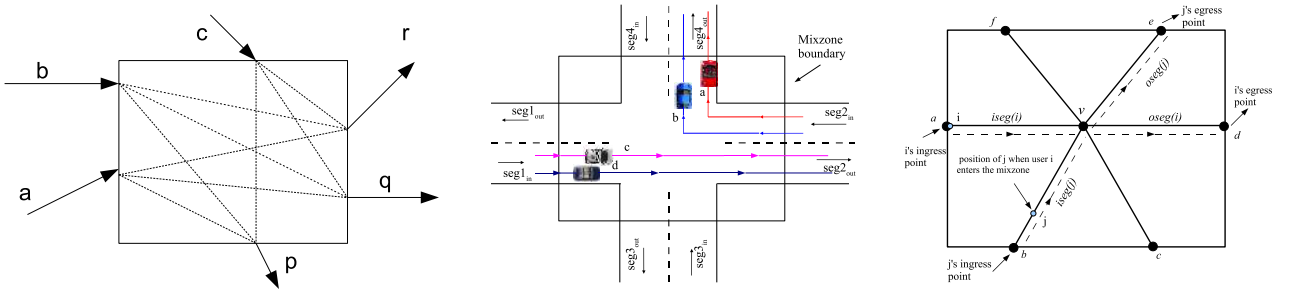
In the theoretical mix-zone model, the anonymity is measured in terms of the unlinkability between the old and new pseudonyms. For user i, exiting with a new pseudonym, i', let $p_{i' \to j}$ denote the probability of mapping i' to j, where $j \in A$. According to Definition 1, the theoretical mix-zone ensures an equi-probable distribution of mapping i' to $j \in A$. In other words, for every outgoing user, i', it is equally probable for i' to be any of the k users in the anonymity set A, having $p_{i' \to j} = \frac{1}{|A|}$. In other words, in an ideal mix-zone, the new pseudonym of user, i is indistinguishable from that of $|A_i|$ other users. Therefore, the entropy, $H(i')$ of each outgoing user i' is computed as follows [6], [7]

$$H(i') = -\sum_{j \in A} p_{i' \to j} \times log_2(p_{i' \to j})$$

The Entropy is a measure of the amount of information required to break the anonymity provided by the system.

Next, we discuss the significance of the two important assumptions in the mix-zone model namely (i) users stay for a random amount of time inside, and (ii) users follow uniform transition probability when entering and exiting a mix-zone.

When the users inside the mix-zone spend random amount of time, it ensures a random reordering between the entry and exit orders providing a strong unlinkability between their old and new pseudonyms. However, a mix-zone that does not ensure random duration of time inside for its users usually leaks information [5], [11]. Such leakage may aid attackers to infer the mapping between the old and new pseudonyms of users. For example, when all users spend a constant time inside, the system would simply function in a FIFO (first-in-first-out) style, with the first exit event corresponding to the first entry event and so on. In that case, even though the users might have changed pseudonyms inside, their mapping from the old and new



(a) Mix-zone Model (b) Road Network Mix-zone (c) Road Network Model

Fig. 1: Mix-zone Models

pseudonyms can still be inferred. A good mix-zone should therefore ensure sufficient randomness in the time spent inside it in order to obtain a high anonymity in terms of unlinkability after the pseudonym change process.

Similarly in a theoretical mix-zone, the probability of transition between an entry point and an exit point follows a uniform distribution. By relaxing this assumption, some transitions between entry and exit points may be more probable than the others. The attacker can use such knowledge to infer the mapping between the old and new pseudonyms. For example, if some transitions are less probable, the attacker may eliminate the pseudonym mappings corresponding to those transitions and thereby improve the success rate of his inference.

3 MobiMix: Overview

In this section, we present an overview of the MobiMix framework. We begin by introducing the challenges imposed by road networks for the construction of mix-zones.

3.1 Problems with Theoretical Mix-zones

Theoretical mix-zones assume mobile users move in an Euclidian space without any spatial constraints. In the real world, mobile users always move on a spatially constrained space, such as road networks or walk paths. Each road network mix-zone corresponds to a road intersection on a road network. Mix-zones constructed at road intersections have a limited number of ingress and egress points corresponding to the incoming and outgoing road segments of the intersection. Furthermore, users in a road network mix-zone are also constrained by the limited trajectory paths and speed of travel that are limited by the underlying road segments and the travel speed designated by their road class category [3]. Thus, users are not able to stay for random amount of time inside a road network mix-zone and the assumption that users follow uniform transition probability when entering and exiting the mix-zone is no longer true.

For example, in Figure 1(b), users a and b enter the road intersection from segment 2 and turn on to segment 4. Users c and d enter from segment 1 and leave on segment 2. When user a and b exit the mix-zone on segment 1 with their new pseudonyms, say α and β, the attacker tries to map their new pseudonyms α and β to some of the old pseudonyms a,b, c, and d of the same users. The new pseudonym α is more likely to be mapped to two of the old pseudonyms, a or b, than the other pseudonyms because users a and b entered the mix-zone well ahead of users c and d and it is thus less probable for c and d to leave the mix-zone before users a and b given the speed and trajectory of travel. Here, the limited randomness on the time spent inside a road network mix-zone introduces more challenges to construct efficient mix-zones. Similarly, in Figure 1(b), in order for the attacker to map α and β to c and d, the old pseudonyms, users c and d should have taken a left turn from segment 1 to segment 4 and users a and b should have taken an U-turn on segment 2. Based on common knowledge of inference, the attacker knows that the transition probability of an $U-turn$ is small and the mapping of α and β to c and d is very less probable. Hence, an efficient road network mix-zone should be resilient to such transition and timing attacks. Next, we introduce the attack models and the anonymity measures for road network mix-zones.

3.2 Adversary Model

We assume that an adversary associated with an untrusted location based service provider may obtain a time series pseudonyms used by the mobile clients. The adversary is considered successful if he can utilize timing and transition based inference to infer the correct linkage between a pseudonym observed from the service requests sent before entering a mix-zone and a pseudonym observed after exiting the mix-zone. Thus the overall goal of the adversary is to track the whereabouts of the user by tracking the mappings between the old and new pseudonyms at various mix-zones and by associating a user's pseudonym to user's actual identity through the association of sensitive locations such as home address or office building of the same user. The goal of users is to change pseudonyms periodically and achieve unlinkability between their old and new pseudonyms so that they can remain anonymous. We therefore consider an adversary is successful if the correct mapping between a new and old pseudonym can be established in a mix-zone.

We describe three types of attacks based on the characteristics of road networks: (1) Timing Attacks, (2) Transition Attacks and (3) Combined Timing and Transition Attacks. **Timing Attack:** In timing attack, the attacker observes the time of entry, $t_{in}(i)$ and time of exit $t_{out}(i)$ for each user entering and exiting the mix-zone. When the attacker sees an user i' exiting, he tries to map i' to one of the users of the anonymity set, A_i. The attacker assigns a

probability, $p_{i' \to j}$ that corresponds to the probability of mapping i' to j, where $j \in A$. The mapping probabilities are computed through inference based on the likelihoods of the rest of the users to exit at the exit time of i', denoted by $t_{out}(i')$. Once the mapping probabilities are computed, the attacker can utilize the skewness in the distribution of the mapping probabilities to eliminate some low probable mappings from consideration and narrow down his inference to only the high probable mappings. Consider an example anonymity set, $A = \{a, b, c\}$, let user a exit with a new pseudonym a' at $t_{out}(a')$ and let the likelihoods of a, b and c exiting at time $t_{out}(a')$ be 0.1, 0.09 and 0.05 respectively. In this case, we show that it is easy to compute the mapping probabilities based on these likelihoods: $p_{a' \to a} = \frac{0.1}{0.1+0.09+0.05} = 0.416$, $p_{a' \to b} = \frac{0.09}{0.1+0.09+0.05} = 0.375$ and $p_{a' \to c} = \frac{0.05}{0.1+0.09+0.05} = 0.208$. Thus, with the timing information, the attacker is able to find that $a' \to a$ is the most probable mapping and $a' \to c$ is least probable. Such timing attack can be detrimental if not handled appropriately in the mix-zone construction and usage model.

Transition Attack: In transition attack, the attacker estimates the transition probability for each possible turn in the intersection based on previous observations. On seeing an exiting user, i', the attacker assigns the mapping probability $p_{i' \to j}$ for each $j \in A$ based on the conditional transitional probabilities $T((iseg(i), oseg(i'))$ and $T((iseg(j), oseg(i'))$. Recall, $T((iseg(j), oseg(i'))$ denotes the conditional probability of an user i' entering through the entry segment, $iseg(j)$ given that the user exited at the segment, $oseg(i')$. The mapping probabilities, $p_{i' \to i}$ and $p_{i' \to j}$ under the transition attack are therefore given by

$$p_{i' \to i} = \frac{T(iseg(i), oseg(i'))}{T(iseg(i), oseg(i')) + T(iseg(j), oseg(i'))}$$

and

$$p_{i' \to j} = \frac{T(iseg(j), oseg(i'))}{T(iseg(i), oseg(i')) + T(iseg(j), oseg(i'))}$$

Transition attack can equally affect the effectiveness of road network mix-zones as timing attack if not handled with care.

Combined Timing and Transition Attack: In the combined timing and transition attack model, the attacker is aware of both the entry and exit timing of the users and as well the transition probabilities at the road intersection for a given road network mix-zone. The attacker can estimate the mapping probabilities $p_{i' \to j}$ for each $j \in A$ based on both the likelihoods of every user j exiting at time $t_{out}(i')$ and the conditional transition probabilities $T(iseg(j), oseg(i'))$.

3.3 Evaluation Metrics

In this subsection, we discuss the set of metrics used by MobiMix and their suitability for measuring the anonymity of road network mix-zones.

Anonymity set size: The size of the anonymity set is the most straight-forward measure of anonymity. However, this metric alone is insufficient given the mapping probabilities may not be uniform in a road network mix-zone. Unlike an ideal mix-zone, in a road network mix-zone the attacker can identify which members are low-probable. Here, the low probable mappings do not effectively count for the anonymity. When the mapping probability distribution is not uniform, there can be attacks based on probability analysis [6], [7], [10]. In other words, we can not say that a road intersection performs as a good mix-zone just by the mere fact that the anonymity set size is greater than k. A number of users in the anonymity set can become low probable under timing and transition attacks and will not effectively count towards anonymity.

Entropy: An alternate measure of anonymity would be based on Entropy that captures the attacker's uncertainty in guessing the mapping between a new and old pseudonym [8], [9], [6], [7], [10]. However, entropy of a user is a measure over all members of the anonymity set. Therefore it may not effectively capture the cases where there are a few skewed mapping probabilities and a large number of non-skewed mapping probabilities. In such cases, a few high probable mappings can significantly increase the attacker's success of guessing the correct pseudonym mapping even though the entropy value may be high. In such cases, a significant part of the entropy could be contributed by a large number of non-skewed mapping probabilities leading to a high value of entropy. Hence, we cannot consider that a mix-zone provides good anonymity for a user if its entropy is greater than a certain value. Two systems can be shown to have the same entropy but however provide different levels of anonymity [10]. Therefore, the entropy measure may not be used as an accurate estimation of the privacy when the mapping probabilities are non-uniform [10] as in our road network mix-zone case.

Normalized Entropy: Normalized entropy, also called Degree of Entropy, is defined as the ratio of the entropy obtained from the road network mix-zone to the entropy obtained from a theoretical mix-zone with the same anonymity set. In other words, it is a measure of how close is the entropy of the roadnet mix-zone compared to a theoretical mix-zone. As entropy itself is a measure over all members of the anonymity set, comparing the entropy of the realistic mix-zone with the theoretical mix-zone also may not accurately capture the non-uniformity in the mapping probability distribution. It can be shown that there are still cases, such as when the normalized entropy is close to 1 but the mapping probabilities significantly deviate from the others [10].

Pairwise Entropy: In order to ensure that the distribution of the mapping probabilities does not deviate much from the uniform distribution, we argue that it is important to measure the deviation of the mapping probabilities in a pairwise fashion. Pairwise entropy between two users i and j is the entropy obtained by considering i and j to be the only members of the anonymity set. In that case, we have two events: the event of i exiting as i' and the event of j exiting as j'. For the first event, we have only two mapping probabilities: $p_{i' \to i}$ and $p_{i' \to j}$. If the probabilities $p_{i' \to i}$ and

$p_{i' \to j}$ are equal, then i' is equally likely to be i or j. The attacker has the lowest certainty of linking the outgoing user i' to i or j (50%). However, if one of the probabilities is much larger than the other, then the new pseudonym i' is more likely to be associated with one of the two old pseudonyms with high certainty ($> 50\%$) by eliminating the low probable one. In comparison, by Definition 1, a theoretical mix-zone ensures a uniform distribution for all possible mappings between old and new pseudonyms and a high pairwise entropy of 1.0 for all pairs of users in the anonymity set. If the pairwise entropy, $H(i,j)$ between users i and j when i exits as i' is close to 1, it means that the attacker will have a high uncertainty similar to that of an ideal mix-zone in guessing the old pseudonym of i'. However, the attacker also has another event namely the exit of j as j'. If this event leaks information, with a low pairwise entropy, $H(j,i)$, for instance if one of the mapping probabilities, $p_{j' \to i}$ and $p_{j' \to j}$ is significantly different from the other, the attacker will be able to identify the old pseudonym of j'. Consequently the attacker can also guess the old pseudonym of i' as i' and j' are mutually exclusive events. Therefore, both the pairwise entropies, $H(i,j)$ and $H(j,i)$ need to be close to 1. Hence, the effective pairwise entropy between users i and j can be assumed as the minimum of the two pairwise entropies $H(i,j)$ and $H(j,i)$.

An ideal mix-zone provides a pairwise entropy of 1.0 for all pairs of users. We argue that an effective mix-zone should provide a pairwise entropy close to 1.0 for all possible pairs of users in the anonymity set. In general if there are k members in the anonymity set, then it requires that the pairwise entropy for all k^2 possible pairs of users in the anonymity set is close to 1.0.

Relative Anonymity: The relative anonymity level is a measure of the level of anonymity provided by the mix-zones, normalized by the level of anonymity required by the users. Higher relative anonymity levels mean that, on the average, users get anonymized with larger k values than the system-specified minimum *k-anonymity* levels.

Success Rate: The success rate measures the ratio of the number of times users obtain anonymity equal or greater than the system-specified minimum *k-anonymity* levels. A good mix-zone should provide anonymization with a success rate close to 100%.

3.4 Road Network Mix-zone Model

In this section, we present the MobiMix model for road network mix-zones and discuss the level of anonymity offered in terms of pairwise entropy and the anonymity set size, k. We model the road network as a directed graph $G = (V_G, E_G)$ where the node set V_G represents the road junctions and the edge set E_G represents the road segments connecting the junctions. In this work, we consider only the road junctions that connect three or more road segments as candidate junctions for mix-zones. Consider a mix-zone constructed at a road intersection v as shown in Figure 1(c). Assume that each user i enters the mix-zone at time $t_{in}(i)$ and exits at time $t_{out}(i)$ with a new pseudonym i'. Let $iseg(i)$ denote the incoming segment of user i through which i enters the mix-zone, $oseg(i)$ denote the outgoing road segment of user i through which i leaves the mix-zone. The speed followed by the users in a road segment follows a Gaussian distribution as empirically verified in [24], [25], [26] with a mean μ and standard deviation σ, where μ and σ are specific to each road class category. For user i, the set of all other users who had entered the mix-zone during the time window defined by $t_{in}(i) - \tau$ to $t_{in}(i) + \tau$, forms the anonymity set of i, denoted as A_i where τ is a small value.

We first derive the pairwise entropy corresponding to user i and its anonymity set A_i under timing attack. Then, we discuss the anonymity obtained under transition attack. We define $d_i(i)$ as the distance travelled by i inside the mix-zone. It is the sum of the lengths of the mix-zone regions on the incoming and exiting segments, $iseg(i)$ and $oseg(i)$. $d_i(j)$ is defined as the distance that j needs to travel inside the mix-zone if it were to exit on the outgoing segment of i namely $oseg(i)$ instead of its actual outgoing segment, $oseg(j)$. $d_i(j)$ is the sum of the lengths of the mix-zone regions on the segments, $iseg(j)$ and $oseg(i)$. If $l_{iseg(i)}$ and $l_{oseg(i)}$ represent the lengths of the mix-zone on the incoming and outgoing segments of i, then $d_i(i)$ is given by

$$d_i(i) = l_{iseg(i)} + l_{oseg(i)}$$

Similarly,

$$d_i(j) = l_{iseg(j)} + l_{oseg(i)}$$

Let $speed_i$ and $speed_j$ denote the random variables of the speed of users i and j. As the speed is assumed to follow a Gaussian distribution, the variables $speed_i$ and $speed_j$ become Normal variables. We also assume that time is slotted and let t be the time of exit of user i, that is $t_{out}(i)$. Let $p_{i' \to j}$ be the probability that the exiting user i' is j and $p_{i' \to i}$ be the probability that the exiting user is i. Users i and j become anonymous from each other if the probability, $p_{i' \to j}$ is exactly equal to the probability, $p_{i' \to i}$ which happens when users i and j enter the mix-zone at the same time and travel the same distance to exit the mix-zone. In short, the more one of these probabilities differs from the other, the higher confidence the attacker will have in linking the old and new pseudonyms.

Let $P(j,t)$ denote the likelihood that user j exits the mix-zone in the time interval, t to $t+1$ where the pair (j,t) is a random variable and $P(j,t)$ numerically equals to the probability that user j takes time in the interval $(t - t_{in}(j))$ to $(t+1-t_{in}(j))$ to travel the distance $d_i(j)$. Accordingly, j needs to travel with an average speed in the range $s_1 = \frac{d_i(j)}{(t-t_{in}(j))}$ to $s_2 = \frac{d_i(j)}{(t+1-t_{in}(j))}$ in order to exit during the time interval between t to $t+1$. Therefore, we have

$$P(j,t) = \int_{s_2}^{s_1} P(speed_j = s) ds$$

Similarly,

$$P(i,t) = \int_{s_2}^{s_1} P(speed_i = s) ds$$

where $s_1 = \frac{d_i(i)}{(t-t_{in}(i))}$ to $s_2 = \frac{d_i(i)}{(t+1-t_{in}(i))}$ and $P(speed_j = s)$ denotes the probability that $speed_j = s$.

If $P(i',t)$ represents the likelihood that some user i' exits at time t to $t+1$, where i' can be either of i or j and the pair (i',t) is a random variable, we have

$$P(i',t) = P(i,t) + P(j,t)$$

Therefore, applying Bayes' Theorem, the probability of i' being j when i' exits at time t, denoted as $p_{i'\to j}(t)$ is given by

$$p_{i'\to j}(t) = P((j,t)|(i',t)) = \frac{P((i',t)|(j,t)) \times P(j,t)}{(i',t)}$$

Here $P((i',t)|(j,t)) = 1$, as $P(j,t)$ is contained in $P(i',t)$. Therefore

$$p_{i'\to j}(t) = \frac{P(j,t)}{P(i',t)}$$

Similarly, the probability of i' being i, $p_{i'\to i}(t)$ is given by

$$p_{i'\to i}(t) = P((i,t)|(i',t)) = \frac{P(i,t)}{P(i',t)}$$

The pair-wise entropy between users i and j when i exits as i' is given by

$$H_{pair}(i,j,t) = -(p_{i'\to i}(t) log p_{i'\to i}(t) + p_{i'\to j}(t) log p_{i'\to j}(t))$$

Similarly, the pair-wise entropy between users i and j when j exits as j' is given by

$$H_{pair}(j,i,t) = -(p_{j'\to i}(t) log p_{j'\to i}(t) + p_{j'\to j}(t) log p_{j'\to j}(t))$$

Here, we notice that even though when i' exits, it might resemble both i and j with a closely equal probability and a high pairwise entropy, $H_{pair}(i,j,t)$, when user j' exits, it might reveal that j' is more likely to be one of i and j than the other as these are mutually exclusive events. Therefore, although the pair-wise entropy between i and j, $H_{pair}(i,j,t)$ may be close to 1 when i' exits, it may happen that the pair-wise entropy of j, $H_{pair}(j,i,t_{out(j')})$ when j' exits is well below 1. Hence, it is important that both of the two pair-wise entropies are high enough to make the attacker harder to guess the mapping. Therefore, the effective pairwise entropy of users i and j is given by the minimum of the two pairwise entropies, $H_{pair}(i,j,t_{out(i')})$ and $H_{pair}(j,i,t_{out(j')})$

$$H_{pair}(i,j) = \min\{H_{pair}(i,j,t_{out(i')}), H_{pair}(j,i,t_{out(j')})\}$$

Also, we find that the pairwise entropy is a function of the exit time, t of i'. As the exit time depends on the time spent inside the mix-zone which is inversely proportional to the speed of the user inside the mix-zone, the pairwise entropy becomes a function of the speed of the user inside the mix-zone. A good mix-zone should offer high pairwise entropy for a wide range of user speeds, for example, say 0 to 90 mph on a highway road and 0 to 40 mph on a residential road. The lowest pairwise entropy offered by the mix-zone within this speed range would define the lowerbound pairwise entropy of the mix-zone. A good mix-zone should therefore offer a high lowerbound, α on the pairwise entropy for a wide range of user speeds.

We now extend our discussion with the pairwise entropy under transition attack. Based on the transition probabilities of the road junction, let $T(seg_l, seg_m)$ be the conditional transition probability computed by the attacker on exit of i'. $T(seg_l, seg_m)$ represents the conditional probability of user i' entering through an incoming segment seg_l given that i' exited on the outgoing segment seg_m. The mapping probabilities, $p_{i'\to i}$ and $p_{i'\to j}$ under the transition attack are therefore given by

$$p_{i'\to i} = \frac{T(iseg(i), oseg(i'))}{T(iseg(i), oseg(i')) + T(iseg(j), oseg(i'))}$$

and

$$p_{i'\to j} = \frac{T(iseg(j), oseg(i'))}{T(iseg(i), oseg(i')) + T(iseg(j), oseg(i'))}$$

Hence, the pairwise entropy under transition attack will be

$$H_{pair}(i,j) = -(p_{i'\to i} log p_{i'\to i} + p_{i'\to j} log p_{i'\to j})$$

In order for the mix-zone to be resilient to transition attacks, the mix-zone should offer a high lowerbound, β on the pairwise entropy after transition attack for all pairs of users in the anonymity set.

Next, we define the criteria for a roadnet mix-zone to function as an effective mix-zone based on the lowerbounds α and β on the pairwise entropies after timing and transition attacks.

Definition 2: A road network mix-zone offers k-anonymity to a set A of users if and only if the following conditions are met:

1) There are k or more users in the anonymity set A.
2) Given any two users $i, j \in A$ and assuming i exiting at time t, the pairwise entropy after timing attack should satisfy the condition: $H_{pair}(i,j,t) \geq \alpha$.
3) For any two users $i, j \in A$, the pairwise entropy after transition attack should meet the condition: $H_{pair}(i,j) \geq \beta$.

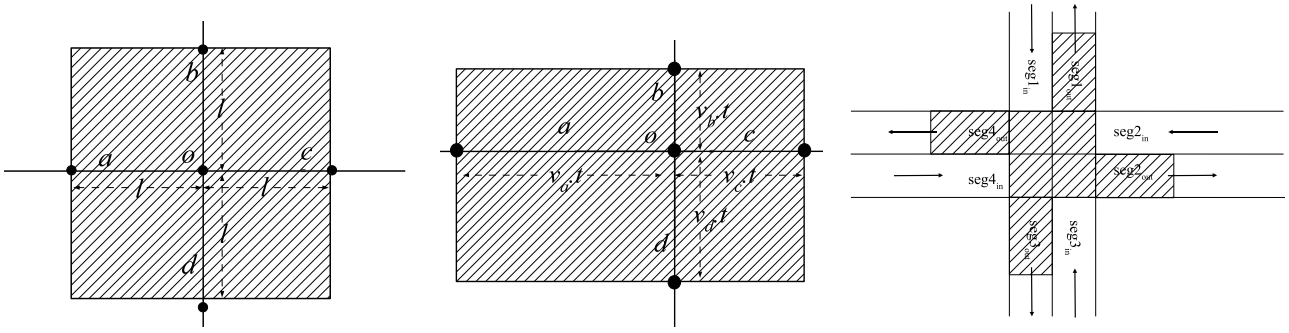
In the next section, we present our proposed techniques and approaches to construct road network mix-zones that effectively satisfy the above conditions.

4 Mix-zone Construction

We compare and analyze the effectiveness of the MobiMix mix-zone construction approaches against timing attack and discuss how the mix-zone geometry and road characteristics impact on the attack-resilience.

4.1 Construction Approaches

We first describe the weaknesses of the naive rectangular mix-zone approach and then propose three MobiMix mix-zone construction techniques taking into consideration the geometry of the zones and their impact on the resilience to timing attack. We propose: (i) Time Window Bounded(TWB) Rectangular, (ii) Time Window Bounded(TWB) Shifted Rectangular and (iii) Time Window



(a) Rectangular Mix-zones (b) Shifted Rectangular Mix-zones (c) Non-rectangular Mix-zones

Fig. 2: Mix-zone Shapes

Bounded(TWB) Non-rectangular mix-zones. All perform better than the naive Rectangular mix-zones under timing attack.

4.1.1 Naive Rectangular Mix-zones

A straight forward approach to construct mix-zones around the road junction is to define a rectangular region centered at the road junction as shown in Figure 2(a). The rectangle is defined based on some default size. For each exiting user i', the set of users that were inside the mix-zone at any given time during user i''s presence in the mix-zone forms its anonymity set, A_i. Here, any two users that were present together at any same given time, become members of each other's anonymity sets.

4.1.2 TWB Rectangular Mix-zones

In the time window bounded approach, the rectangle is constructed in the same way as in naive rectangular mix-zone, however, the anonymity set for each user, i is assumed to comprise of users who had entered within a time window in the interval, $|t_{in}(i) - \tau_1|$ to $|t_{in}(i) + \tau_2|$. Here, $t_{in}(i)$ is the arrival time of user i and τ_1 and τ_2 are chosen to be small values so that the time window ensures that the anonymity set of i comprises only of the users entering the mix-zone with a closely similar arrival time as that of i. The goal of the mix-zone construction is to ensure high pairwise entropy for every pair of users entering within the time window. We would like to note that the anonymity guarantee made by the mix-zone is by design a lower bound on the anonymity observed by the adversary for two reasons. First, we argue that a good anonymity system should anonymize users in such a way that there is similar probability of mapping the actual subject to all the other users in the anonymity set. Thus, by discarding the low probable mappings and the corresponding users from the guaranteed anonymity set, we get an estimate of the number of users whose mapping probability distribution closely resembles a uniform distribution. Thus we get a measure of the number of users to belong to the anonymity set in such a way that they get anonymized in a way very similar to that of an ideal system. For road intersections that have segments with the same speed distribution, we can precisely guarantee a lowerbound on the pairwise entropy for the members of the anonymity set by constructing the anonymity set with the right value of time window based on our MobiMix road network model. Although, the notion of mix-zone time window has been adopted in existing mix-zone proposals [11], [13] where a default value of time window is assumed for the junctions, the TWB rectangular approach decides the right size of the time window based on the arrival rate of users so that k or more users enter within the time window. Also as mentioned earlier, for road intersections that have road segments with same speed distribution, we can guarantee a lowerbound pairwise entropy based on the Mobimix model for each pair of users entering with the time bound window. But for the sake of our experimental comparisons, we consider TWB rectangular mix-zones as the candidate mix-zone for comparison with the existing mix-zones proposed in [11].

4.1.3 TWB Shifted Rectangular Mix-zones

In the Time window bounded shifted rectangular approach, the rectangle is not centered at the centre of the junction, instead it is shifted in such a way that from any point of entry into the mix-zone, it takes the same amount of time to reach the centre of the road junction when travelled at the mean speed as shown in Figure 2(b). In the same way, from the centre of the junction, it takes the same time to reach any exit point when travelling at the mean speed of the road segments. Here, a set of users entering within the short time window, $|t_{in}(i) - \tau_1|$ to $|t_{in}(i) + \tau_2|$ are likely to exit the mix-zone at the same time. Hence, when user i exits as i' the attacker would find that i' is likely to be any of the members of the anonymity set, A_i. If t represents the average time to reach the centre of the road junction from an entry point which is the same as the average time to reach an exit point from the junction center, then the mix-zone lengths on the segments would be given by the product of their mean speed, say v and the average time, t as shown in 2(b). Compared to naive rectangular and time window bounded rectangular mix-zones, shifted rectangular mix-zones provide good pairwise entropy for many cases, however, they do leak information when the speed of the users deviate from the mean speed resulting in a weaker anonymity system [6], [7], [10]. Another limitation of this approach is that it may not be possible to satisfy the shifted rectangle property if the road segments are not orthogonal. Hence, this approach is limited to only road junctions with orthogonal segments.

4.1.4 TWB Non-Rectangular mix-zones

A more effective way to construct mix-zones would be to have the mix-zone region start from the centre of the junction only on the outgoing road segments as shown in Figure 2(c). We refer to this technique as non-rectangular approach. The non-rectangular approach is free from timing attacks caused by the heterogeneity in the speed distribution on the road segments. As in the rectangular approaches, the anonymity set for each user, i comprises of users who had entered the mix-zone within a time window in the interval, $|t_{in}(i) - \tau_1|$ to $|t_{in}(i) + \tau_2|$. The length of the mix-zone along each outgoing segment is chosen based on the mean speed of the road segment, the size of the chosen time window and the minimum pairwise entropy required. We discuss details on computing the mix-zone size and time window in section 4.4.

4.2 Timing Attack Analysis

In this sub-section, we analyze the privacy strengths of the proposed mix-zone approaches under timing attack and compare their attack-resilience.

4.2.1 Naive Rectangular Mix-zones

Timing attacks are highly effective in Naive rectangular Mix-zones. In Naive Rectangular mix-zones, although the anonymity set size is typically large, a large number of members of the anonymity set become low probable under the timing attack. For instance, in Figure 2(a), consider two users i and j entering from the segments a into the mix-zone. Let user i exit with a new pseudonym i' on segment c and let us assume the four road segments in the mix-zone, a, b, c and d have the same speed distribution. If the arrival times of i and j differ by a large value, then although users i and j might have been present together in the mix-zone for some amount of time, the attacker might infer that the user who entered first is more likely to exit first and that it is unlikely for j to have overtaken i before i exits the mix-zone. Therefore, the pairwise entropy of the naive rectangular mix-zones is low under timing attack, leaking more information to aid the attacker.

4.2.2 TWB Rectangular Mix-zones

TWB rectangular mix-zones have high resilience to timing attack in road junctions that have segments with the same speed distribution as the members of its anonymity set have similar time of arrival into the mix-zone. However, when the segments of the road intersection have different mean speeds, for instance if they belong to different road classes, the attacker may be able to eliminate some mappings based on the timing information. For example, in Figure 2(a), let us assume a mix-zone of size 0.5 miles $\times$ 0.5 miles with segments a and c of residential road category having a mean speed of 20 mph and segments b and d of highway roads with a mean speed 60 mph. Consider two users i and j entering the mix-zone at the same time. Let user i enter through the highway segment b and exit through the highway segment d and let user j enter though the residential segment a and exit through the residential segment c. If both i and j travel around the mean speed of their respective road segments, then i and j would exit approximately in 30 seconds and 90 seconds respectively. When user i exits out with a changed pseudonym i' in 30 seconds, the attacker can infer that i' is more likely to be i than j. Thus, even though the anonymity set consists of users entering with closely similar arrival time, the differences in the speed distribution on the roads leaks information to aid the timing attack.

4.2.3 TWB Shifted Rectangular Mix-zones

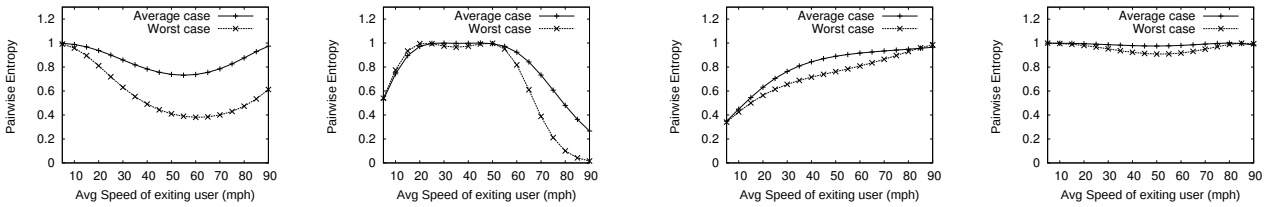
TWB shifted rectangular mix-zones are resilient to timing attacks even on road junctions that have segments with different mean speeds provided the users travel at the mean speed of the segments. However, they are also prone to timing attack when the speed of the users deviate from the mean speed of the road segments. For example, in Figure 2(b), consider a mix-zone of size 0.5 miles X 0.5 miles in a road intersection with a slow residential road segment, a having mean speed 20 mph and three other highway segments, b, c, and d having mean speed 60 mph. Let all road segments have a standard deviation of 10 mph from their mean speed. The computation would yield $v_a.t = 0.375$ miles and $v_b.t = v_c.t = v_d.t = 0.125$ miles. Let users i and j enter the mix-zone at the same time. Let user i enter through the highway segment, b and exit through the highway segment, d and let j enter through the residential road segment, a and exit through the highway segment, c. Let us assume user j travels with a speed of 10 mph on segment a and travels at 60 mph on segment, c. In this case, the attacker would see j' exiting in 2 minutes, 32.5 seconds. With this timing information, the attacker can find that j' is more likely to be mapped to j than i because if j' is i, then i should have travelled really slow on the highway segments b and c, with an average speed of 5.9 mph in order to exit after 2 minutes, 32.5 seconds. However, if j' is j, then j needs to have travelled only at 10 mph on the residential road segment, a which is more likely to happen. Thus, the attacker can guess that j' is j with high confidence. In general, the shifted rectangular approach performs badly when the user's speed deviate from the mean speed of the road segments.

4.2.4 TWB Non-rectangular Mix-zones

The TWB non-rectangular mix-zone is most resilient to timing attacks as it does not encounter any disparity in the speed distributions. Here, as long as a pair of users enter within each other's time window, the attacker can not infer the correct pseudonym mappings if the length of the mix-zone is sufficiently large for the chosen time window. In the next subsection, we compare the effectiveness of these mix-zone approaches.

4.2.5 Pairwise Analysis

In order to better understand the effect of timing attack on guessing the mapping between the old and new pseudonyms, we perform a pairwise analysis considering only two users in the mix-zones. We compare the effectiveness of the different approaches in Figure 3. As an example, we consider a mix-zone of length 400 meter in a road junction that has two highway road segments where



(a) Naive Rectangular (b) TWB Rectangular (c) TWB Shifted Rectangular (d) TWB non-rectangular

Fig. 3: Effectiveness of Mix-zones against timing attack.

the speed is normally distributed with 60 mph mean and 20 mph standard deviation and 2 residential road segments where the speed is distributed with 25 mph mean and 10 mph standard deviation. For a rectangular and shifted rectangular mix-zone, the mix-zone length corresponds to the longer side of the rectangle and for the non-rectangular mix-zone, the mix-zone length refers to the length of the longest mix-zone region on the outgoing road segments. In this pairwise analysis, for the rectangular mix-zones, the breadth is also taken as 400 meter. We consider two users i and j and measure the worst case and average case pairwise entropies. User i travels on the fast highway segments and user j travels on the slow residential segments. The worst case typically represents the arrival times of i and j separated by the maximum possible value defined by the mix-zone time window. Here the mix-zone time window is taken as 4 sec for the example mix-zone considered. The average case represents the case where the arrival times of i and j are separated by half the size of the time window, namely 2 sec. User i changes its pseudonym to i' and the X-axis shows the average speed followed by the exiting user, i' inside the mix-zone and the Y-axis shows the worst case and average case pairwise entropies. We find that both the naive rectangular approach and the time window bounded rectangular approach have low pairwise entropy for both the worst case and average case for speeds even close to 60 mph, the mean speed of the highway segments that i travelled. Interestingly, the TWB rectangular approach shows higher pairwise entropy when user i' travels slow on its highway segments. This is because, if i' travels slow on the highway segments, then its exit time would resemble that of j much better as j is travelling on a slow residential segment. Similarly, the shifted rectangular approach shows good pairwise entropy when the speed of i' is close to the mean speed, 60 mph. However, its pairwise entropy drops when the speed of i' deviates from its mean speed. Outperforming all these approaches, the TWB non-rectangular approach has a very steady high pairwise entropy for a wide range of speeds of i'. This is because, in this mix-zone geometry, users travel only on one segment in the mix-zone and thereby do not encounter any disparity in the speed distributions and therefore it is the most resilient geometry for timing attack.

4.3 Transition Attack Analysis

We now analyze the impact of transition attack that can be launched to guess the mapping between the pseudonyms. For each exiting user, i' the attacker observes the exiting

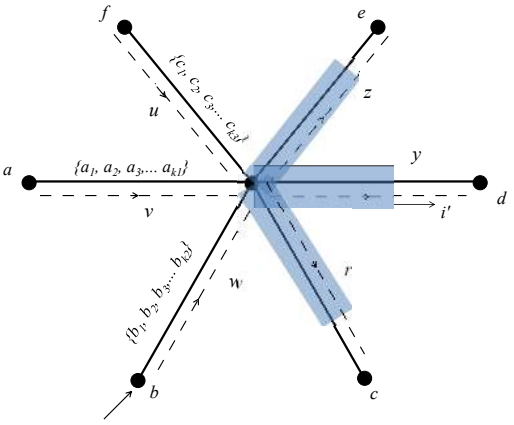


Fig. 4: Countering Transition Attack

segment of i' and tries to map i' to one of the users, j in the anonymity set based on the conditional transitional probability of exiting in the outgoing segment, $oseg(i')$ given that j entered from the incoming segment, $iseg(j)$. We refer the interested readers to Appendix D for an experimental study on the significance of the transition attack on the road map of the NW Atlanta region of Georgia.

In order to protect against transition attack in cases where the transition probability is skewed, the mix-zone time window should be chosen in such a way that for each outgoing segment, l, there are enough number of users (k or more) entering the mix-zone from the road segments that have similar transitioning probability to the outgoing segment, l, and hence have a higher pairwise entropy, say greater than or equal to β. Therefore, the attacker will have at least k users in the anonymity set that he cannot ignore from consideration.

Figure 4 shows a TWB non-rectangular mix-zone with 3 incoming segments, u, v, w and three outgoing segments, r, y, z. Let $T(u, y)$ be the conditional probability of an user entering the junction through segment u given that the user exited on segment y. The attacker assigns probability $p_{i' \to j}$ to each of the users $\{a_1, a_2, a_3, ...a_{k1}, b_1, b_2, b_3, ...b_{k2}, c_1, c_2, c_3, ...c_{k3}\}$ based on the conditional transition probabilities $T(u, y)$, $T(v, y)$, $T(w, y)$. Assume the conditional transition probability $T(u, y)$ is too small compared to $T(v, y)$ and $T(w, y)$ and let the probabilities $T(v, y)$ and $T(w, y)$ be similar. Let us assume an user i enters from segment w and exits in segment y as i'. Here, the attacker may be able to ignore $\{c_1, c_2, c_3, ...c_{k3}\}$ from the anonymity set of i'. However, i' would have a higher pairwise entropy with

$\{a_1, a_2, a_3, ...a_{k1}, b_1, b_2, b_3, ...b_{k2}\}$. Thus, for outgoing segment y, if we can ensure that there are always k or more users entering from the segments v and w, then for any user, i' exiting on segment y, the attacker would be confused to differentiate i' from at least k other users that forms the effective anonymity set, A'_i. Here it should be also noted that even though the exit of i' on segment y does not leak information, the exit of some user, say a_2 along segment z may leak some information if the transition probability, $T(v, z)$ is much smaller than $T(w, z)$. Therefore, the effective anonymity should not contain those members that exit in a segment where user i's probability of exiting is lower as these are mutually exclusive events.

In the next sub-section, we discuss how to determine the time window and size of the mix-zone so as to make it resilient to both timing and transition attacks, yielding a high lowerbound, α and β on the pairwise entropies after timing and transition attacks respectively.

4.4 Combination of Timing and Transition attacks

The mix-zone time window directly impacts the number of users arriving from the various segments and therefore decides the mix-zone's resilience to transition attack. Once the right size of the mix-zone time window is determined for a specified level of resilience to transition attack in terms of a high lowerbound, β on the pairwise entropy after transition attack, we need to determine the length of the mix-zone for the given time window so as to ensure a high lowerbound on the pairwise entropy after timing attack.

According to empirical research on road traffic modeling [37], [38], [39], the user arrival on the road segments follows a Poisson process. Let λ_l denote the mean arrival rate on each incoming segment l, $\lambda_L^{x,y}$ represent the cumulative mean arrival rate of the users that effectively count towards the anonymity set of an user and i' exiting along segment y that entered through segment x. If $M^{x,y}$ is a subset of the road segments in the mix-zone, we have $\lambda_L^{x,y} = \sum_{l \in M^{x,y} | H^y_{pair(l,x)} > \beta} (\lambda_l - \sum_{z | \exists m \in M^{x,y}, H^z_{pair(m,l)} < \beta} T(l, z) \times \lambda_l)$. It is the sum of the arrival rate of the segments such that the members have high pairwise entropy with each other and with i' during the exit of i' in segment y. Note that it excludes among the users who entered from segment, l, those that would exit in some segment, z where the conditional probability of exiting in z is significantly different. Here $M^{x,y}$ is chosen as that subset of the road segments that maximizes $\lambda_L^{x,y}$. If $N(t)$ represents the number of users who had entered the mix-zone at time t since the beginning, then the probability of having n users enter during a short time window, $\tau^{x,y}$ is given by

$$P[N(t + \tau^{x,y}) - N(t) = n] = \frac{e^{-\lambda_L^{x,y}\tau^{x,y}}(\lambda_L^{x,y}\tau^{x,y})^n}{n!}$$

$N(t + \tau^{x,y}) - N(t)$ would represent the number of users arrived within the short time interval, $\tau^{x,y}$. The probability that k or more users enter the mix-zone in the time window, $\tau^{x,y}$ is

$$P[(N(t+\tau^{x,y}) - N(t) \geq k] = 1 - \sum_{1 \leq n \leq k} \frac{e^{-\lambda_L^{x,y}\tau^{x,y}}(\lambda_L^{x,y}\tau^{x,y})^n}{n!}$$

By adjusting the size of the time window, $\tau^{x,y}$, we can lowerbound the number of users arriving from the segments whose conditional probability of exiting in segment y is similar to that of users from segment x. For instance, we may choose the time window, $\tau^{x,y}$ such that there are $k = 5$ or more users entering with a high probability, say $p = 0.9$. The overall time window, τ of the mix-zone is given by the maximum value of $\tau^{x,y}$ among the various segments, y in the road junction.

$$\tau = \max_y \tau^{x,y}$$

Once the value of τ is decided, we determine the length of the mix-zone so that the mix-zone provides a high lowerbound, α on the pairwise entropy after timing attack for a wide range of user speeds. For example, we might want a lowerbound pairwise entropy of $\alpha = 0.9$ for a wide range of users' speed, say 0 mph to 90 mph. Our algorithm iteratively increments the length of the mix-zone till the expected lowerbound on the pairwise entropy is met for the chosen time window, τ. In this context, we note that except for the TWB non-rectangular mix-zones, the other approaches suffer from timing attacks and hence it is not possible to have a time window and mix-zone length for them to ensure a high lowerbound on the pairwise entropy. However, the TWB non-rectangular mix-zones offer high lowerbounds even for small mix-zone lengths. As we have a lower bound on the pair-wise entropy and a lower bound on k, the number of users, the mix-zone can now make probabilistic guarantees on the anonymity provided.

5 Mix-zone placement

In this section, we present the mix-zone placement algorithms that find the best set of road intersections to function as mix-zones based on the user arrival rates, statistics of user movements, road network topology and road characteristics in terms of mean user speeds and the temporal and spatial resolution of location exposure. Although individual mix-zones are efficient with respect to providing the required level of anonymity, careful deployment on the road is crucial to ensure good cumulative anonymity for users as they traverse through multiple mix-zones on their trajectories. Mix-zones placed too far from each other may lead to longer distances between adjacent mix-zones in users' trajectories. On the other hand, if mix-zones are placed too close to one another, users may go through mix-zones more frequently than necessary. An optimal solution to the mix-zone placement problem is NP-complete for even small road networks [12]. Thus we use a heuristic-based placement approach in MobiMix. A good placement algorithm should (i) provide sufficient anonymity in each of the mix-zones (ii) ensure that users go through sufficient number of mix-zones along their path to the destination and (iii) minimize the total number of mix-zones in the

system, thereby minimizing the overall cost of the privacy protection. A naive placement strategy is to randomly select a subset of road junctions with three or more road segments. A better strategy is to place mix-zones at intersections that have high density of traffic and low skewness in the transition probability distribution. We call this approach the road-aware top n placement. An alternative approach is the grid-based quadtree placement strategy, which divides a road-network into grid cells using quadtree index partition and maximizes the average distance between any pair of mix-zones within each quadrant (grid cell). Due to space constraint, we omit the design detail of these mix-zone placement algorithms in the paper and refer readers to Appendix B for further detail.

6 Experimental Evaluation

We divide the experimental evaluation of MobiMix into three components: (i) the effectiveness of our mix-zone construction approaches in terms of their resilience to timing and transition attacks and comparison with existing naive mix-zone approaches (ii) their performance in terms of success rate and relative anonymity levels and (iii) the effectiveness of the mix-zone placement algorithms in terms of overall cumulative anonymity, mix-zone size and spatial uniformity of placement. Before reporting our experimental results, we first briefly describe the experimental setup.

6.1 Experimental setup

We use the GT Mobile simulator [19] to generate a trace of 10000 cars moving on a real-world road network, obtained from maps available at the National Mapping Division of the USGS [3]. By default we use the map of Northwest Atlanta region of Georgia that has 6831 road intersections with 10000 mobile users. We refer the readers to Appendix C for a detailed description of the experimental setup including the realistic mobility model used in the experiments.

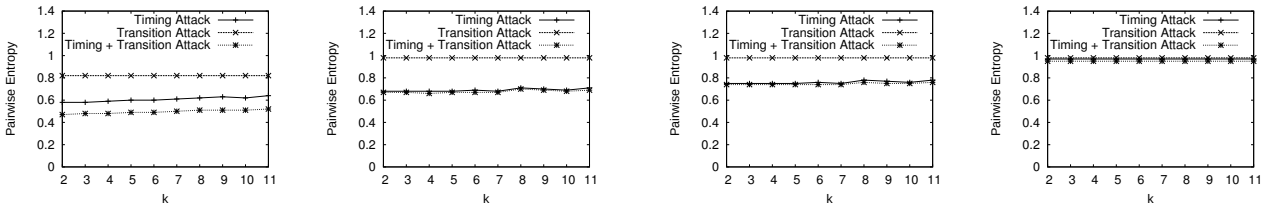
6.2 Experimental results

Our experimental evaluation consists of three parts. First, we evaluate the effectiveness of the mix-zone construction algorithms by measuring their attack resilience to timing and transition attacks. We then evaluate the effectiveness of the mix-zones in terms of the success rate in providing the desired value of k and study the relative anonymity level which is defined as the ratio of the obtained value of k to the expected value of k. We observe how these parameters behave when we vary the settings of a number of parameters, such as the expected value of k, the expected probability of success, p. Our final set of experiments evaluates the performance of the mix-zone placement algorithms in terms of the overall cumulative anonymity of the users, average mix-zone size and spatial uniformity of mix-zone placement. Our results show that the MobiMix construction techniques are effective, fast and scalable and outperform the basic construction methods by a large extent.

6.2.1 Resilience to Timing and Transition Attacks

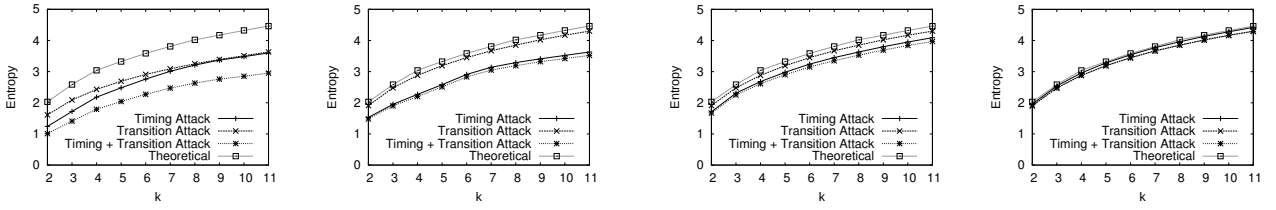
In our first set of experiments, we analyze the effectiveness of the mix-zones against timing, transition and combined attacks. The description of the road map used for this set of experiments is described in appendix C. Out of the 6831 road junctions in the map, more than 2000 candidate junctions were chosen to build mix-zones based on their user arrival rate and the number of road segments that connect to them. Figure 5 shows the average pairwise entropy of the mix-zones for various values of k, the size of the anonymity set. We observe that the pairwise entropy after transition attack is low in the naive rectangular mix-zone compared to the other MobiMix approaches as the MobiMix mix-zones are protected for transition attack with their anonymity sets consisting of only members that have high pairwise entropy to each other. The effect of timing attack is different across various approaches: we find that the TWB non-rectangular mix-zones perform the best under timing attack with the average pairwise entropy close to 1.0. Here, the length of the non-rectangular mix-zone is computed so as to ensure a lowerbound pairwise entropy of $\alpha = 0.9$ for the chosen time window size, τ which is computed based on the user arrival rate in the road junction to ensure the expected value of k with a high probability of p = 0.9. However, as discussed in section 4.2.5, it is not possible to lowerbound the pairwise entropy for the other mix-zone approaches. Hence, in order to compare the effectiveness of these approaches with the TWB non-rectangular approach, we construct the TWB rectangular and TWB shifted rectangular mix-zones with the same length and time window as used by the non-rectangular mix-zone. Similarly, the size of the naive rectangular mix-zone is fixed in such a way that the mean time to cross the mix-zone equals the time window of the TWB non-rectangular mix-zone. In Figure 5, we also find that the naive rectangular and time window bounded rectangular mix-zones have low pairwise entropies after timing attack but the pairwise entropy of the TWB shifted rectangular approach is relatively higher, close to 0.8 as it's geometry is more resilient to timing attack. However, a high pairwise entropy of 0.9 or higher may be often required to ensure strong anonymity. In such cases, the time window bounded rectangular approach becomes the most efficient approach. Additionally, in the figure, we find that the effect of combined timing and transition attack is at least as severe as either of these attacks in isolation and it gets worse in naive rectangular mix-zones which is least resilient to both timing and transition attacks.

Similarly, Figure 6 shows the comparison of the worst case pairwise entropy after timing attack for various mix-zones. The worst case pairwise entropy represents the lowest possible pairwise entropy obtained by the users after timing attack. Here also, only the TWB non-rectangular approach offers a high value for the worst case pairwise entropy. The other approaches in their bad cases leak a lot information to aid the attacker. We also compare the overall entropy under attacks for various values of k in



(a) Naive Rectangular (b) TWB Rectangular (c) TWB Shifted Rectangular (d) TWB non-rectangular

Fig. 5: Average Pairwise Entropy after Attacks



(a) Naive Rectangular (b) TWB Rectangular (c) TWB Shifted Rectangular (d) TWB non-rectangular

Fig. 7: Comparison of Entropy after attacks

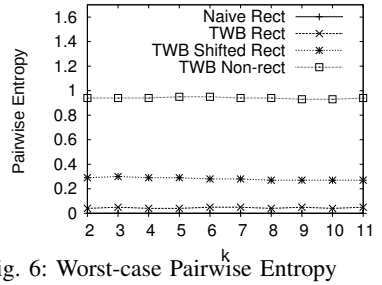


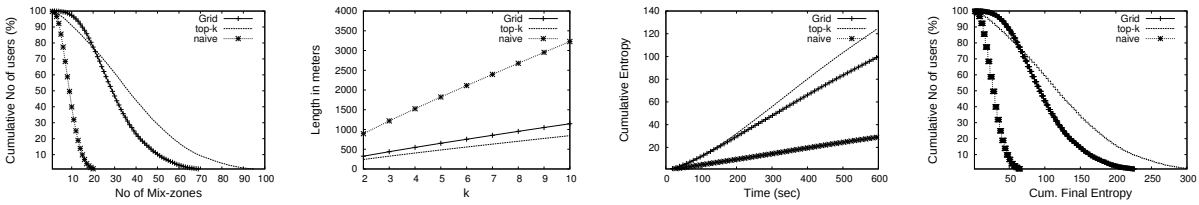
Fig. 6: Worst-case Pairwise Entropy

Figure 7 for the same experimental setting and road map described in appendix C. The overall entropy is computed by assigning the probability distribution, $P_{i'->j}$ for each user $j \in A_i$ based on the likelihood of user j to exit at the exit time of i. The line showing the theoretical value of entropy corresponds to the actual entropy obtained from an ideal mix-zone for the same anonymity set as the realistic mix-zones. We find that the TWB non-rectangular approach has the highest overall entropy after timing, transition and combined attacks closely resembling that of a theoretical mix-zone. We discuss additional results on success rate and relative anonymity in Appendix D.

6.3 Performance of Placement Techniques

We now study the performance of the various mix-zone placement algorithms in terms of the mix-zone size, spatial uniformity of placement, the average number of mix-zones traversed by the mobile clients and the entropy obtained during user's travel with the three mix-zone placement algorithms namely (i) Naive placement (ii) top-n (user and road characteristics-aware) placement and (iii) Grid (Quadree) based network-aware placement. The experiment uses the NW atlanta region map that contains 6831 road junctions, out of which the placement algorithms chooses 7% of the road intersections for deploying mix-zones that corresponds to 478 road junctions. The experiment uses a 10 minute simulation period. Figure 8(a) shows the cumulative distribution function (CDF) of the users in percentage for various number of mix-zones traversed during their trip. We find that users traverse less number of mix-zones in the naive mix-zone deployment scheme. We find more than 60% of the users traverse less than 10 mix-zones during their entire 10 minute travel. The top-n (user and road characteristics-aware) placement scheme enables users to pass through higher number of mix-zones as it basically finds all the intersections that have dense traffic. Here, users go through more number of mix-zones in short intervals of distance which may not be necessary. Such unnecessary traversal of mix-zones may deteriorate the quality of service for the mobile clients. In Figure 8(a), we also find that there is a significant percentage of users traversing less number of mix-zones. For example, more than 9% of the users traverse only less than 10 mix-zones during the 10 minute trajectory. This is due to the non-uniformity in the spatial distribution of the mix-zones. Hence, users traversing some part of the road networks go through few mix-zones while users travelling in other parts unnecessarily go through many mix-zones. The Grid (Quadree-based network-aware) deployment ensures a higher level of spatial uniformity in the distribution of mix-zones. In the Grid approach, we find that almost all users traverse at least 10 mix-zones during the 10 minute interval. Also, we find that users do not unnecessarily traverse many mix-zones, only few users travel a large number of mix-zones as compared to the top-n placement scheme.

Figure 8(b) shows the average size of the mix-zone in meters for values of k. We find that the naive placement approach leads to larger mix-zone sizes for even small values of k as it lacks knowledge of the user arrival rate and user transition probability. Such large mix-zone size would significantly impact the service quality of the mobile users. The top-n scheme has the lowest mix-zone length among the three approaches as it identifies the most densely populated road junctions where even small mix-zone sizes yield higher k. However, the Grid placement scheme is also able to achieve almost similar mix-zone lengths as the top-n placement as it considers the road characteristics and user population factors in addition to the



(a) Mix-zones traversed (b) Mix-zone size (c) Entropy Timeline (d) Final Entropy

Fig. 8: Mix-zone Placement

road network-aware spatial uniformity. Figure 8(c) shows the time line of cumulative entropy. The x-axis shows the time in seconds and the Y-axis represents the average cumulative entropy obtained. The naive placement shows low cumulative entropy, particularly in the beginning of the timeline (0 to 200 sec). Also, we find that both the top-n and Grid placements show similar average cumulative entropy in the beginning of the timeline although the top-n scheme has higher cumulative entropy at the later part of the timeline as users go through a large number of mix-zones with the top-n placement. In order to better understand the impact of the spatial uniformity of the mix-zone deployment on the cumulative entropy, in Figure 8(d) we study the cumulative distribution of the users in percentage for various values of average final cumulative entropy at the end of the 10 minute interval. It shows a very similar trend as in Figure 8(a). We find the naive placement scheme does not achieve high final cumulative entropy for all users. The top-n scheme has overall higher final cumulative entropy but has a significantly higher percentage of users having low final cumulative entropy. In the Grid approach, almost all users obtain higher final cumulative entropy and therefore the distribution has low skewness. Thus, the Grid placement scheme becomes the most effective choice in deploying mix-zones.

7 Related Work

Anonymization based location privacy research can be broadly categorized into spatial cloaking with location k-anonymity guarantee and mix-zones with unlinkability of old and new pseudonyms.

Spatial cloaking with location k-anonymity has evolved from uniform k for all mobile users [16] to personalized k-anonymity [14], [20], [4]. Most recent work on location k-anonymity have focused more on travelers on road networks [21], [23]. *XStar* [21] performs spatial cloaking based on road-network-specific privacy and QoS requirements, striking a balance between the attack resilience of the performed protection and the processing cost of the anonymous queries. *Cachecloak* [23] uses cache prefetching to hide the exact location of the user by requesting the location based data along an entire predicted path. [33] proposes a collaborating strategy where users can have their LBS queries answered by nearby peers and thereby minimize the exposure of location information to the untrusted LBS. As discussed before, the approaches based on location cloaking do not work for applications that require identity or pseudo-identity of mobile users. Also the existing methods [23], [33] are not suitable for continuous location query services.

The mix-zone based location privacy research is targeted at protecting location privacy for users who request continuous location services or LBSs that require pseudo-identity, such as tracking a taxi cab within 5 miles of my location. The idea of using mix-zones for location privacy was introduced in [5] and the idea of building mix-zones at road intersections were proposed in [11], [13]. [36] proposes the idea of changing pseudonyms at social spots (similar to mix-zones) so that users can remain anonymous. A formulation for optimal placement of mix-zones on a road map is discussed in [12], which showed that such optimal placement is NP-hard for even small road networks. Similarly, [34] presents an optimal solution to the mix-zone placement problem which is NP-hard and presents approximations by assuming every road segment has at least one mix-zone and by relaxing the assumption of non-uniform traffic and by ignoring road junctions that yield lower entropy. [35] proposes a game-theoretic approach to mix-zone placement with the assumption that at least one end of each road segment has a mix-zone. We note that most of the existing mix-zone techniques are straight forward by using rectangular or circular shaped zones and their construction methodologies do not take into account the effect of timing attacks and transition attacks.

To the best of our knowledge, MobiMix mix-zones are the only solutions to date that take into consideration of timing and transition attacks in its mix-zone constructions. Thus MobiMix makes a number of original contributions: First, its mix-zone construction algorithms minimize the effect of timing and transition attacks based on the characteristics of the underlying road network and guarantee an expected value of anonymity by incorporating the statistics of both road network topology and road network traffics. Second, unlike previous mix-zone placement techniques, such as [34], [35], which leads to having 50% road junction as mix-zones, the MobiMix mix-zone placement techniques are closely integrated with its attack-resilient mix-zone construction methodologies and thereby achieves good privacy even with as few as 10% mix-zones on the road network.

8 Conclusions

We have presented MobiMix, a framework for building attack resilient road network mix-zones for protecting the location privacy of mobile clients. We highlight that road network mix-zone construction and placement techniques should take into consideration a number of factors such

as the mix-zone geometry, the statistics of the user population, and the spatial and velocity constraints on the movement patterns of the users. We show analytically and experimentally that the MobiMix construction and placement techniques are efficient and more resilient to timing and transition attacks than the existing mix-zone approaches. Our research on MobiMix continues along several directions, including considering more sophisticated attack models based on background knowledge about the users' trajectory patterns and travel behavior.

9 Acknowledgments

This research is partially supported by grants from NSF CISE NetSE program, SaTC program, I/UCRC and a grant from Intel ICST on Cloud Computing.

References

[1] J.R. Cuellar, J.B. Morris, D.K. Mulligan, J. Peterson and J. Polk. Geopriv requirements. *IETF Internet Draft*, 2003.
[2] U. Hengartner and P. Steenkiste. Protecting access to people location information. In *Security in Pervasive Computing*, 2003.
[3] U.S. Geological Survey. http://www.usgs.gov.
[4] B. Bamba, L. Liu, P. Pesti, and T. Wang. Supporting Anonymous Location Queries in Mobile Environments with PrivacyGrid. In *WWW*, 2008.
[5] A. Beresford and F. Stajano. Location Privacy in Pervasive Computing. *Pervasive Computing, IEEE*, 2003.
[6] C. D?az, S. Seys, J. Claessens, B. Preneel. Towards Measuring Anonymity. *PETS*, 2002.
[7] A. Serjantov and G. Danezis. Towards an Information Theoretic Metric for Anonymity. *PETS*, 2002.
[8] C. Troncoso, B. Gierlichs, B. Preneel and I. Verbauwhede. Perfect Matching Disclosure Attacks *PETS*,2008.
[9] G. Danezis and C. Troncoso. Vida: How to use Bayesian inference to de-anonymize persistent communications Privacy Enhancing Technologies Symposium (PETS 2009)
[10] G. Toth, Z. Hornak and F. Vajda. Measuring Anonymity Revisited. In Norsec, 2004.
[11] J. Freudiger, M. Raya, M. F?legyhazi, P. Papadimitratos, and J.-P. Hubaux. Mix-Zones for Location Privacy in Vehicular Networks. In WiN-ITS, 2007.
[12] J. Freudiger, R. Shokri and J.-P. Hubaux. On the Optimal Placement of Mix Zones. In PETS, 2009.
[13] L. Buttyan and T. Holczer and I. Vajda. On the effectiveness of changing pseudonyms to provide location privacy in VANETs In ESAS 2007
[14] B. Gedik and L. Liu. Location Privacy in Mobile Systems: A Personalized Anonymization Model. In *ICDCS*, 2005.
[15] G. Ghinita, P. Kalnis, and S. Skiadopoulos. PRIVE: Anonymous Location-Based Queries in Distributed Mobile Systems. In *WWW*, 2007.
[16] M. Gruteser and D. Grunwald. Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. In *MobiSys*, 2003.
[17] M. Gruteser and D. Grunwald. Enhancing Location Privacy in Wireless LAN Through Disposable Interface Identifiers: A Quantitative Analysis. *Mobile Networks and Applications*, 2005.
[18] J. Hong and J. Landay. An Architecture for Privacy-Sensitive Ubiquitous Computing. In *Mobisys*, pages 177–189, 2004.
[19] P. Pesti, B. Bamba, M. Doo, L. Liu, B. Palanisamy, M. Weber. GTMobiSIM: A Mobile Trace Generator for Road Networks. College of Computing, Georgia Institute of Technology, 2009, *http://code.google.com/p/gt-mobisim/*.
[20] M. Mokbel, C. Chow, and W. Aref. The New Casper: Query Processing for Location Services without Compromising Privacy. In *VLDB*, 2006.
[21] T. Wang and L. Liu. Privacy-Aware Mobile Services over Road Networks In *VLDB*, 2009
[22] J. Krumm. Inference Attacks on Location Tracks In *PERVASIVE*, 2007.
[23] J. Meyerowitz and R. Choudhury. Hiding Stars with Fireworks: Location Privacy through Camouflage In *MOBICOM 2009*
[24] D. Helbing Derivation and empirical validation of a refined traffic flow model In Physica A 1996, 223, 253-282.
[25] D. Helbing Empirical traffic data and their implications for traffic modeling In Physical Review E 1997, 55, R25-R28.
[26] D. Helbing Fundamentals of traffic flow In Physical Review E 1997, 55, 3735-3738.
[27] Freudiger, Julien, et al. Evaluating the privacy risk of location-based services In Financial Cryptography and Data Security. Springer Berlin Heidelberg, 2012. 31-46.
[28] [Srivatsa, Mudhakar, and Mike Hicks. Deanonymizing mobility traces: Using social network as a side-channel In 2012 ACM conference on Computer and communications security, 2012.
[29] Zang, Hui, and Jean Bolot Anonymization of location data does not work: A large-scale measurement study In 17th annual international conference on Mobile computing and networking, 2011.
[30] Ma, Chris YT, et al. Privacy vulnerability of published anonymous mobility traces In 16h annual international conference on Mobile computing and networking, 2010.
[31] Shokri, Reza, et al. Quantifying location privacy In 2011 IEEE Symposium on Security and Privacy (SP), 2011.
[32] Shokri, Reza, et al. Unraveling an old cloak: k-anonymity for location privacy In 9th annual ACM workshop on Privacy in the electronic society, 2010.
[33] Shokri, Reza, et al. Collaborative location privacy In2011 IEEE 8th International Conference on Mobile Adhoc and Sensor Systems (MASS), 2011.
[34] Liu, Xinxin, et al. Traffic-aware multiple mix zone placement for protecting location privacy In INFOCOM, 2012.
[35] Jadliwala, Murtuza, et al. Optimizing mixing in pervasive networks: a graph-theoretic perspective In Computer SecurityESORICS, 2011.
[36] Lu, Rongxing, et al. Pseudonym changing at social spots: An effective strategy for location privacy in vanets In IEEE TVT, 2012
[37] Darroch, J. N. On the Traffic-Light Queue In Ann.Math. Statist., 35, pp. 380-388, 1964.
[38] Haight, F. A. Overflow At A Traffic Flow. Biometrika. Vol. 46, Nos. 3 and 4, pp. 420-424, 1959
[39] Ohno, K. Computational Algorithm for a Fixed Cycle Traffic Signal and New Approximate Expressions for Average Delay In Transportation Science, 12(1), pp. 29-47, 1978
[40] G. Ghinita, P. Kalnis, M. Kantarcioglu, E. Bertino Approximate and exact hybrid algorithms for private nearest-neighbor queries with database protection. In*GeoInformatica*, 2011.
[41] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, K. Tan Private Queries in Location Based Services: Anonymizers are not Necessary. In *SIGMOD*, 2008.
[42] T. Wang and L. Liu. Execution Assurance for Massive Computing Tasks In *IEICE Transactions on Information and Systems*, Vol. E93-D, No. 6 (June 2010), Special session on Info-Plosion.
[43] P. Williams, R. Sion Usable PIR. In *NDSS*, 2008.

Balaji Palanisamy is an assistant professor in the School of Information Science in University of Pittsburgh. He received his M.S and Ph.D. degrees in Computer Science from the college of Computing at Georgia Tech in 2009 and 2013 respectively. His primary research interests lie in scalable and privacy-conscious resource management for large-scale Distributed and Mobile Systems. At University of Pittsburgh, he co-directs research in the Laboratory of Research and Education on Security Assured Information Systems (LERSAIS), which is one of the first group of NSA/DHS designated Centers of Academic Excellence in Information Assurance Education and Research (CAE CAE-R). He is a recipient of the Best Paper Award at the 5^{th} International Conference on Cloud Computing, IEEE CLOUD 2012. He is a member of the IEEE and he currently serves as the chair of the IEEE Communications Society Pittsburgh Chapter.

Ling Liu is a full Professor in Computer Science at Georgia Institute of Technology. She directs the research programs in Distributed Data Intensive Systems Lab (DiSL), examining various aspects of large scale data intensive systems. Prof. Ling Liu is an internationally recognized expert in the areas of Cloud Computing, Database Systems, Distributed Computing, Internet Systems, and Service oriented computing. Prof. Liu has published over 300 international journal and conference articles and is a co-recipient of the best paper award from a number of top venues, including ICDCS 2003, WWW 2004, 2005 Pat Goldberg Memorial Best Paper Award, IEEE Cloud 2012, IEEE ICWS 2013. Prof. Liu is also a recipient of IEEE Computer Society Technical Achievement Award in 2012 and an Outstanding Doctoral Thesis Advisor award in 2012 from Georgia Institute of Technology. In addition to services as general chair and PC chairs of numerous IEEE and ACM conferences in data engineering, very large databases and distributed computing fields, Prof. Liu has served on editorial board of over a dozen international journals. Currently Prof. Liu is the editor in chief of IEEE Transactions on Service Computing, and serves on the editorial board of ACM Transactions on Internet Technology (TOIT), ACM Transactions on Web (TWEB), Distributed and Parallel Databases (Springer), Journal of Parallel and Distributed Computing (JPDC).

APPENDIX A
MOBIMIX SYSTEM ARCHITECTURE

The system architecture of MobiMix consists of following components (1) MobiMix Anonymizer, (2) Road Network Monitor, (3) Mix-zone construction modules, (4) Mix-zone placement and (5) Computing Infrastructure. We describe each of them below:

A.0.1 Mix-zone Anonymizer

The Mix-zone anonymizer is responsible for anonymizing the raw location updates received from the mobile clients before releasing it to the Location Based Service provider for processing. The anonymizer stores two important information: (1) Mix-zone-junctions Map that stores which junctions are presently functioning as mix-zones and (2) User-pseudonyms Map that stores the mapping between the user's real identity and their current pseudonyms. Upon arrival of a location update from a client, the anonymizer checks to see if the present location of the client corresponds to a mix-zone region. If so, the anonymizer drops the location update from being sent to the Location-based service (LBS) provider and denies service to the mobile client. Also, the mobile user is assigned a new pseudonym and the corresponding entry is updated in the User-pseudonym Map. If the mobile user is not currently inside a mix-zone, then the anonymizer passes the location update to the LBS server by replacing the real identity of the user its the current pseudonym.

A.0.2 Road Network Monitor

The road network monitor works closely with the mix-zone anonymizer. It examines each location update of the mobile client and monitors the current behaviour of the road network in terms of the user speeds and their arrival patterns. It consists of the following sub-components:

Arrival Rate Monitor: The arrival rate monitor observes the user arrivals in each road junction along each road segment and identifies the user arrival process and the associated parameters. It provides the arrival rate parameter to the mix-zone construction module.

Transition Monitor The transition monitor observes the transitions taken by the users in each road junction and computes the transition probabilities for all possible transitions in the road intersections. This information is used to compute the conditional transition probabilty in the attack-resilient mix-zone construction phase.

Road Speed Monitor Based on the location updates received from the clients, the road speed monitor computes the current speed of the road segments in terms of the mean speed and standard deviation. Also, it is aware of the speed limits of the road segments based on the road category they belong to.

A.0.3 Mix-zone Construction

The mix-zone construction module consists of the implementation of the MobiMix attack-resilient mix-zone techniques. It has information about the user arrival rate, transition probability in the junctions and speed distribution in the road segments through the road network monitor. The mix-zone construction takes into account the effect of both timing and transition attacks and ensures an expected number of users in the mix-zone that directly corresponds to the level of anonymity obtained. The construction module outputs the mix-zone size and shape for each mix-zone and also assists the mix-zone placement module to determine the best set of road intersections to function as mix-zones based on the user arrival rate, the transition probabilities at the junctions and the speed characteristics of the road segments.

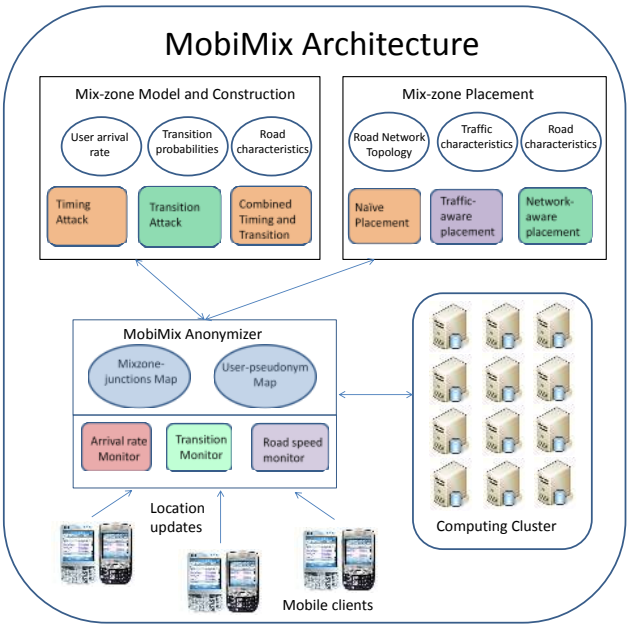


Fig. 1: MobiMix System Architecture

A.0.4 Mix-zone Placement

The mix-zone placement component is responsible for deploying the mix-zones in the road network. In a huge road network of several tens of thousands of road junctions, the critical decision of which road junctions function as mix-zones can significantly impact the anonymity of the users. Improper selection of road junctions may result in unacceptably large size of mix-zones due to low user arrival rate or skewed transition probability distribution in the junctions. The placement module has knowledge of the road network topology, road characteristics in terms of road segment speed and arrival rate and also the mobility profiles of the users in terms of the transitioning probabilities at the road junction. MobiMix implements three mix-zone placement techniques namely (i) Naive Placement, (ii) Road characteristics aware (top-n) placement and (iii) Quadtree based (Grid) Network-aware placement.

A.0.5 Computing Infrastructure

The anonymizer with its monitoring sub-components run in a computing infrastructure. This computing infrastructure

can be a dedicated infrastructure within the anonymizer's organization. Here, a set of servers would be reponsible for anonymizing users in one geographical area and each server gets to receive only the location updates corresponding to its geographic area thereby balancing the overall load in the system.

APPENDIX B
DESCRIPTION OF PLACEMENT ALGORITHMS

An optimal solution to the mix-zone placement problem may be obtained using a formulation similar to that discussed in [14], however such optimal solutions to the placement problem become NP-complete for even small road networks. In this appendix, we present three heuristic-based strategies for mix-zone placement. The mix-zone placement algorithms find the best set of road intersections to function as mix-zones based on the user arrival rates, statistics of user movements, road network topology and road characteristics in terms of mean user speeds and the temporal and spatial resolution of location exposure. We know that the anonymity strength of the mix-zone is directly proportional to the anonymity set size and the attack resilience of the mix-zone, however, for a given value of anonymity set size, k, the size of the mix-zone is directly proportional to the arrival rate of the users form various road segments connected to the road junction and the skewness in the transition probability distribution. Therefore, the cost of a mix-zone is directly proportional to the size of the mix-zone as it directly impacts the limits on the usage of the location based service. A good placement algorithm should provide sufficient anonymity in each of the mix-zones, should also ensure that users go through sufficient number of mix-zones along their path to the destination, while minimizing the total number of mix-zones maintained in the system.

B.1 Naive Placement

In the naive placement scheme, the mix-zones are chosen based on only the structure of the intersections, considering only those that connect to three or more road segments. This set of road intersections forms the candidate set of mix-zones. Among the candidate set of road intersections, the mix-zones are placed by choosing a random subset of the candidate set of mix-zones. Although this straight-foward approach of mix-zone placement is aware of the road intersection topology, the approach lacks knowledge of the user arrival rate and user travel characteristics and hence it does not make careful decisions to minimize the cost of the constructed mix-zones. For example, even road intersections having low user arrival rates and skewed transition probability distributions may get chosen for placing mix-zones. However, constructing mix-zones at them would lead to huge mix-zone sizes in order for them to be sufficiently resilient to timing and transition attacks. Hence, the overall cost of the mix-zone placement in the naive approach may not be minimal.

B.2 Road-aware top -n Placement

In this placement methodology, the mix-zones are placed at intersections that have high density of traffic and low skewness in the transition probability distribution. The mix-zones constructed at such intersections are small in size, incurring minimal cost in terms of limiting the service inside the mix-zones. All the mix-zones are constructed to yield a certain lower-bounded anonymity in terms of the anonymity set size, k, and resilience to timing and transition attacks. This is done by carefully choosing the time window, τ to ensure that sufficient number of vehicles arrive in the anonymizing time window and the size of the mix-zone in such a way that every member of the anonymity set has a high pairwise entropy after transition and timing attacks. In this approach, the top-n mix-zones are selected based on their average estimated anonymity levels of the road intersections, precisely in terms of the cost of the mix-zones. If $C(v)$ is the cost of the mix-zone constructed at road junction v for the privacy guarantees H_{min}. The selection algorithm sorts the road junctions in the increasing order of the cost of the mix-zones $C(v)$ and chooses the top-n candidates for the placement. Although, this approach minimizes the overall cost of the mix-zones in the road network, the distribution of the mix-zones may not be uniform across the road network. For example, while some parts of the network may be densely populated with mix-zones, some other parts may be very scarce in mix-zones. As a result, users following some trajectories will pass through unnecessarily more mix-zones, while some users may not be able to find sufficient mix-zones in their trajectories.

B.3 Quadtree/Grid Network-aware Placement

In the quadtree-based network-aware approach, the placement algorithm considers the topology of the road map in addition to the user and road characteristics. Similar to the top-n placement approach, this approach also considers only road intersections having low skewness in transition probability and high traffic arrival rates. However in order to ensure a uniform distribution of the mix-zones, the placement decision is made by closely considering the underlying road network topology. For instance, the placement of the mix-zones should ensure that the trajectories followed by the users have sufficient number of mix-zones at evenly separated distances. Hence, the mix-zone deployment in the road network has to ensure saptial uniformity while minimizing the overall cost of the mix-zones in terms of their size.

The Quadtree-based network-aware placement is a two phase algorithm. The first phase of the algorithm recursively divides the entire road map to construct a quadtree index. The quadtree construction divides the area based on the number of road junctions in it, the overall geographical area and the total length of the road segments and the number of candidate junctions for mix-zones. The algorithm dynamically decides and partitions if it needs to recursively partition the space further into four quadrants. At the end of the quadtree construction, each quadrant roughly consists

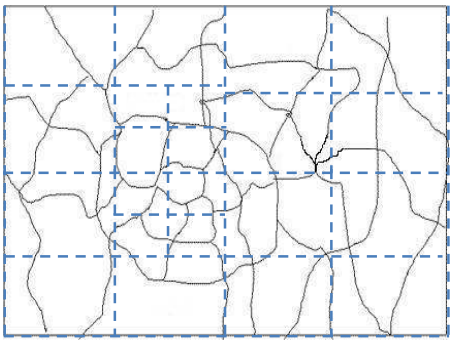


Fig. 2: Grid-based Placement

of the same number of road junctions, total segment length and number of candidate mix-zones as shown in figure 2.

The second phase of the algorithm deploys the mix-zones on a quadrant by quadrant basis. In each quadrant, the algorithm attempts to deploy the same number of mix-zones, however, the decision of which road junctions function as mix-zones is done to minimize the overall mix-zone cost while maximizing the average distance between any pair of mix-zones in a given quadrant. The objective is to maximize the pairwise distance between the mix-zones while not exceeding a certain specified maximum cost. This ensures that the mix-zones are uniformly distributed within each quadrant achieving higher spatial uniformity. Let Q represent the set of quadrants in the road network and let each quadrant, q have m mix-zones. If V_q and M_q respectively represent the set of all intersections in quadrant q and the set of intersections that functions as mix-zones in quadrant, $q \in Q$, then the objective function is given by

$$\min_{q \in Q} \sum_{v1,v2 \in M_q} dist(v1, v2)$$

subject to the constraints:

$$\sum_{v \in V_q} x_v = m$$

$$\sum_{v \in V_q} C(v) x_v \leq C_{max} \times m$$

where x_v is a boolean variable indicating if vertex $v \in V_q$ is a mix-zone and vertex v belongs to M_q if $x_v = 1$.

APPENDIX C
EXPERIMENTAL SETUP

The GTMobiSim mobile simulator extracts the road network based on three types of roads – *expressway*, *arterial* and *collector* roads. Our experimentation uses maps from three geographic regions namely that of Chamblee and Northwest Atlanta regions of Georgia and San Jose West region of California to generate traces for a two hour duration. We generate a set of 10,000 cars on the road network that are randomly placed on the road network according to a uniform distribution. The speed of the cars are distributed based on the road class categories as shown in Table 1. We use the Random Router mobility model in GTMobiSim where Cars generate random trips with source and destination chosen randomly and shortest path routing is used to route the cars for the random trips. This captures more realistic scenarios than the random walk model. For instance, unlike the random walk model, the highway roads and expressways are more populated than the small residential roads as these roads share more parts of the shortest paths used by the users. Also, the random router model gives more realistic transition probabilities at the junctions which is essential to our evaluation.

Road type	**Expressway**	**Arterial**	**Collector**
Mean speed(mph)	60	50	25
Std. dev.(mph)	20	15	10
Speed Distribution	Gaussian	Gaussian	Gaussian

TABLE 1: Motion Parameters

Parameter	**Value**
Map	Northwest Atlanta region
Mobility Model	Random Roadnet Router
Total number of vehicles	10000
Number of Road junctions	6831
Number of Road segments	9187

TABLE 2: Simulation Parameters and Setting

APPENDIX D
ADDITIONAL EXPERIMENTAL RESULTS

D.0.1 Significance of Transition Attack

We study the significance of protecting mix-zones against transition attack by measuring the distribution of the pairwise entropy among the road junctions based on the skewness in their transition probabilities. We show the distribution of worst case and average pairwise entropies after transition attack in table 3 for the Northwest Atlanta map of Georgia. The worst case refers to the least possible pairwise entropy obtained in the junction. We notice that most junctions have only reasonably high average pairwise entropy after transition attack, suggesting that the transition probabilities at these junction do not follow an uniform distribution. We find that only less than 12 % of the junctions have a high pairwise entropy in the range 0.9 to 1.0 after the transition attack. Also, the worst case entropy of many junctions (more than 90%) have a low value of 0, corresponding to the mappings that indicate an U-turn. Clearly, in these cases of low pairwise entropy, the attacker would able to eliminate the mappings if transition attack is not handled properly in the mix-zone construction.

(a) Average

$H_{(i,j)}$	% of junctions
0.0-0.1	0
0.1-0.2	0
0.2-0.3	0
0.3-0.4	0
0.4-0.5	0.25
0.5-0.6	1.33
0.6-0.7	7.75
0.7-0.8	37.75
0.8-0.9	41.33
0.9-1.0	11.58

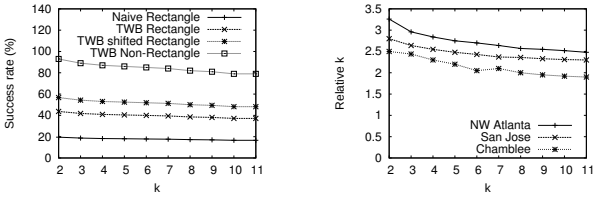
(b) Worst case

$H_{(i,j)}$	% of junctions
0.0-0.1	95.58
0.1-0.2	0.166
0.2-0.3	0.5
0.3-0.4	0.42
0.4-0.5	0.25
0.5-0.6	0.42
0.6-0.7	0.33
0.7-0.8	1.0
0.8-0.9	0.58
0.9-1.0	0.75

TABLE 3: Pairwise Entropy with Transition attack

D.0.2 Success Rate and Relative Anonymity

In order to measure the effectiveness of the mix-zones, we study the success rate of them in providing the expected value of k. Here, the expected probability of getting k or more users, p is taken to be 0.9 and the value of k is varied from 2 to 11. Figure 3(a) shows the comparison of the success rate among the mix-zone approaches. A mix-zone is considered successfull for an user if the user has at least k other users in its anonymity set with pairwise entropies greater than 0.9 under both timing and transition attacks. As evident from the figure, the TWB non-rectangular mix-zones have the highest success rate, the other mix-zones have low success rate due to their lack of resilience to timing attack. In order to compare the level of anonymity offered by the mix-zones with the anonymity expected from them, we measure relative anonymity which is defined as the ratio of the value of obtained k to the value of expected k. Figure 3(b) shows the variation of relative-k of TWB non-rectangular mix-zones with respect to the expected value of k for different geographic maps. The expected success rate is set to 90%. The graphs show that the value of relative k lies within the range of 2 to 3, meaning that the mix-zone on an average offers two to three times the anonymity requested by the users.



(a) Success Rate (b) Relative Anonymity

Fig. 3: Success rate and Relative-k