Traffic-Aware Multiple Mix Zone Placement for Protecting Location Privacy

Xinxin Liu, Han Zhao, Miao Pan, Hao Yue, Xiaolin Li and Yuguang Fang
University of Florida, Gainesville, FL, 32611, USA
Email:{xinxin,han}@cise.ufl.edu,{miaopan,hyue}@ufl.edu,{andyli,fang}@ece.ufl.edu

***Abstract*—Privacy protection is of critical concern to Location-Based Service (LBS) users in mobile networks. Long-term pseudonyms, although appear to be anonymous, in fact empower third-party service providers to continuously track users' movements. Researchers have proposed the mix zone model to allow pseudonym changes in protected areas. In this paper, we investigate a new form of privacy attack to the LBS system that an adversary reveals a user's true identity and complete moving trajectory with the aid of side information. We propose a new metric to quantify the system's resilience to such attacks, and suggest using multiple mix zones to tackle this problem. A mathematical model is presented that treats the deployment of multiple mix zones as a cost constrained optimization problem. Furthermore, the influence of traffic density is also taken into account to enhance the protection effectiveness. The placement optimization problem is NP-hard. We therefore design two heuristic algorithms as practical and effective means to strategically select mix zone locations, and consequently reduce the privacy risks of mobile users trajectories. The effectiveness of our proposed solutions is demonstrated through extensive simulations on real-world mobile user data traces.**

I. Introduction

The rapid development of mobile devices and positioning technologies has led to flourish of personalized mobile services based on users' locations, known as Location-Based Service (LBS). Utilizing the underlying network infrastructure, mobile applications are capable of tracking a user's movement and delivering information based on his current location directly to his mobile device. A wide range of mobile LBS applications have been developed to aid people's daily activities, including but not limited to work, entertainment, health, and navigation. According to a recent study conducted by Strategy Analytic, with the increasing consumer demands such as search, maps or navigation, LBS is envisioned to become an over $10-billion-per-year business by year 2016 [1].

Although LBS significantly benefits mobile users, privacy issues arise during the process of collecting, storing, and sharing of users' location information. Users who subscribe to LBS may not realize the *extent* to which their location information is revealed or *with whom* the service providers (smartphone companies, app companies, and etc.) are sharing this information. We use the following example to illustrate the necessity of privacy protection. User Alice may use LBS at a shopping plaza to check out nearby restaurants and may not mind others discovering her current location. However, privacy becomes important when Alice later enters a special hospital and does not want to share this information with anyone else, especially when Alice is a well-known public figure. Even though pseudonyms instead of true identities are commonly used to camouflage the location trace files, previous works [2], [3] have pointed out that such pseudonym protected trajectories are vulnerable to inferential attacks, i.e., with the aid of side information, the adversary can discover true identities of many users and furthermore obtain an extended view of their whereabouts. Serious consequences such as physical crimes may happen due to the revelation of a user's complete moving trajectory.

Location privacy protection in mobile networking environments is challenging for two reasons. First, wireless communications in mobile networks are easy to intercept, e.g., an eavesdropper can collect transmitted information of mobile users at certain public place. Besides, since people are publicly observable, context information can easily be obtained from their conversations or behaviors. As a result, partial trajectory information associated with a user's true identity is inevitably exposed to the eavesdropper. Second, the limited resources of mobile devices greatly restrict **P**rivacy-**E**nhancing **T**echnologies (PET) one could apply and deploy in the network. Consequently, current PET solutions rest on simple schemes to hide the true identity of a mobile user from a passive adversary, rather than complex cryptographic technologies commonly used in wired network.

One common model for privacy protection is the mix zone model originally proposed by Beresford and Stajano [4]. A mix zone refers to a region where users can change their pseudonyms without being observed by the adversaries. It effectively breaks the continuity of a user's location exposure such that the user's future locations can be protected. Previous mix zone solutions mainly focus on single mix zone construction to achieve k-anonymity (a privacy metric denoting a state that the information of each individual cannot be distinguished from at least $k-1$ others) for location privacy protection. However, using a single mix zone is insufficient to handle the aforementioned inferential attack using side information, since side information may correspond to any part of a user's trajectory. In order to achieve a desired level of protection, multiple mix zones are needed for a certain region to minimize the identity correlation over all point-of-interests recorded in

The work presented in this paper is supported in part by National Science Foundation under Grant No. CNS-0709329, CNS-0721744, CNS-0923238, CCF-0953371, CNS-0916391, CNS-0721744, and CNS-0940805 (BBN subcontract).

978-1-4673-0775-8/12/$31.00 ©2012 IEEE

a user's trajectory. However, the deployment of mix zones also comes with a cost of impaired service availability which limits the number of mix zones one could deploy. The traffic density at each location also affects the effectiveness of mix zone deployment, e.g., mix zone works better at busy road intersections.

In this paper, we address the problem of optimal multiple mix zones placement to enhance the effectiveness of privacy protection. Using graph theory, we characterize properties and constraints of the optimization problem, and build a formal mathematical model with the objective of minimizing pairwise information correlation (measured by pairwise node connectivity) over all possible mix zone placement locations. Our contributions can be summarized as follows.

- We propose a new metric to quantify the system's resilience to the side information based attack model [3].
- An optimization formulation with cost and traffic constraints is presented to model the multiple mix zones placement problem. Since this problem is NP-hard, we propose two heuristic algorithms for practically finding a solution to the optimization problem.
- We verify the effectiveness of our solution using real-world mobile user traces.

The rest of the paper is organized as follows. In Section II we present the system model and description of mix zones. In Section III the adversary model is stated. The formulation of the mix-zone placement problem is presented in the ILP form in Section IV. Heuristic algorithms are proposed in Section V. Section VII summarizes related research in the literature. Section VIII concludes the paper.

II. Background

A. System settings

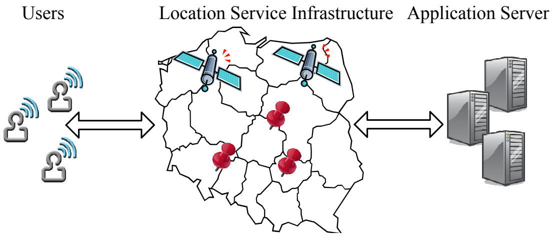


Fig. 1. System model: Users, Location Service Infrastructure, and Application server

Our target Location-Based Service (LBS) system is comprised of three major components: *Users*, *Location Service Infrastructure (LSI)*, and third-party *Applications*, as depicted in Figure 1. The location information about users are actively or passively updated through wireless communications between users' hand-held devices and LSI. [1] Some specific third-party application is interested in a set of locations on the map, and these locations are referred as Point-Of-Interests (POIs). For example, consider a gas-price application designed to provide gas station information tailored to each user's preference around the user's current location in Chicago. The application will register all gas stations as POIs at LSI. Now suppose a subscribed user Alice in Chicago is driving from home to her work place, the gas-price application will continuously display up-to-date gas prices at neighborhood gas stations along Alice's route. To implement this functionality, the gas-price application will record Alice's preferences and require LSI to periodically send Alice's location information as callbacks so that the price information at the nearest gas station can be retrieved.

The physical positions of the registered POIs may be located at road side or road intersections. At any time, the users use *true identity* (most likely the identifier associated with their devices) to exchange messages with LSI. For communication between LSI and third-party applications, user identities are camouflaged by *pseudonyms*. Users cannot bypass LSI and talk directly with third-party applications; otherwise, applications can trivially obtain the true user identities as well as the complete user trajectories. Because an application only knows the pseudonym of a user, it needs to send service information through LSI back to user. Therefore, pseudonyms play an important role to prevent identity exposure in the scenarios that some form of identity is required for the functioning of the third-party application.

The communications occurred during the service period result in a trajectory file recording a user's footprints. Each entry in the trajectory file is a 3-tuple: <pseudonym, timestamp, location>. Based on the trajectory record of Alice, one can approximate the time when Alice arrives at each POI along her complete trajectory. Using a single long-term pseudonym is vulnerable to privacy attacks, since one accidental true identity leakage will result in a user's whole trajectory being compromised. For better privacy protection, researchers have proposed the mix zone model to break the continuity of a user's trajectory.

The following notations are listed to ease the presentation in the later sections.

- $\{\mathcal{P}_i\}, (i = 1, 2, \ldots, n)$: the set of registered locations within certain range, e.g., POIs in a city.
- u_j: user j's pseudonym in the system.
- v_j: user j's true identity present in the side information.
- $\mathcal{T}_{u_j}(t_i), (i = 1, 2, \ldots, m; j = 1, 2, \ldots, n)$: per-user time-based function used to describe the location traces collected by an adversary. Here, t_i indicates the time when u_j's location is reported by LSI.
- $\mathcal{S}_{v_j}(t_i^{'}), (i = 1, 2, \ldots, \kappa; j = 1, 2, \ldots, \pi)$: the side information obtained by an adversary. It is also a function of time and records the set of users $\mathcal{V}$'s location information within the same city territory as the trace file.

B. Mix zone model

The concept of mix zone [4] refers to a service restricted area where mobile users can change their pseudonyms so that the mapping between their old pseudonyms and new pseudonyms are not revealed. For example, in Figure 2, five

[1] In this paper we do not elaborate the details of the location update procedure.

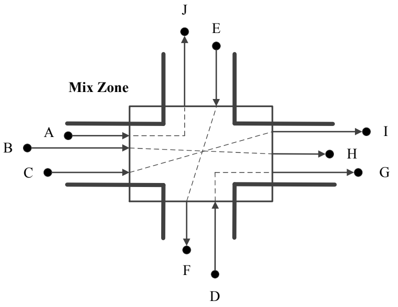


Fig. 2. A mix zone example. Rectangular area: a mix zone deployed at road intersection. Arrows begin or end with dots: observable user moving trajectory. Dashed lines: user moving trajectories not observable by LBS applications.

users with pseudonyms A-E enter the mix zone from different entrances and exit with a different set of pseudonyms F-J at approximately the same time. The links between old and new pseudonyms are not observable by any outsider. This change effectively "mixes" the identities of all users to achieve privacy protection. A significant amount of research [5]–[7] has been devoted to investigating the optimal size and shape of a single mix zone deployment. Targeting at vehicular networks, existing mix zone construction methods are not suitable for the LBS system model, as LBS users can be pedestrians that are not confined to vehicle moving patterns. In the system model presented in II-A, a mix zone is established by LSI at the software level. A mix zone is selected by LSI from the set of registered POIs, $\{\mathcal{P}_i\}, (i = 1, 2, \ldots, n)$. Once $\mathcal{P}_i$ is chosen as a mix zone, LSI will assign a set of new pseudonyms to the users leaving $\mathcal{P}_i$. Such a software level mix zone establishment approach has considerable flexibility over physical deployment of mix zones, because the location and the size of the mix zones are not constrained by terrestrial borders and can be easily adjusted.

If multiple mix zones are deployed alongside a user's routes, the user's continuous trajectory is broken into a set of discrete segments, where each segment is associated with a unique pseudonym. This causes an adversary to lose the tracking target. Each single mix zone lowers the privacy risk in the user's next trajectory segment. To quantify this protection effectiveness, a common metric for evaluating an adversary's uncertainty in finding out the link between a user's old and new pseudonym in a mix zone is information entropy given by:

$$H_m = -\sum_u p_u \log p_u \tag{1}$$

where p_u stands for the probability of mapping an old pseudonym to a new pseudonym when leaving the mix zone area. Another important characteristic of mix zone is that its effectiveness is greatly affected by traffic condition. For example, mix zones deployed at locations with higher traffic and more outlets have higher entropy than those placed at locations with less or barely no traffic. Therefore, when selecting mix zone locations, traffic density should be carefully considered.

III. Threat Model

In our threat model, we consider LSI to be trustworthy for two reasons. First, a service provider who operates LSI in general has no incentive to become adversarial. This is because such a service provider who can afford the expensive equipments in LSI is more likely to be an established major player on the market. The opportunity cost for acting against its customers is too high to afford, e.g., facing expensive law suit and devastating reputation damage. Second, a majority of localization services offered by LSI rely on message exchange between users and LSI. In both cellular and wireless networks, the true identifier of a user's hand-held device is necessary for communication purpose. Therefore, if LSI is not trustful, we need to consider how to localize a mobile user under the current infrastructure, without exposing any ID information to LSI. This leads to another set of problems that are out of the scope of this paper.

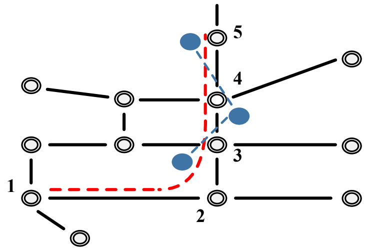


Fig. 3. Example of side information and user traces in an abstracted POI graph. Concentric circles: POIs in a graph. Edges: road segments connecting POIs. Shaded vertex: a mix zone. Dashed line: user trajectory. Solid dots connected by dashed line: side information.

The third-party LBS applications are considered not trustworthy. They may directly attack a mobile user's privacy, or secretely sell information to other individuals or organizations. An adversary $\mathcal{A}$ refers to any entity formed by one or more malicious parties (by colluding) who aim at learning the locations associated with mobile users' true identities. We do not consider the case that $\mathcal{A}$ actively stalks a particular user. Since an adversary has the complete trajectory profiles camouflaged by pseudonyms, it is often characterized as a *global passive eavesdropper* that becomes the major threat in the literature [8].

Besides the trajectory profile, $\mathcal{T}_{u_j}(t_i), (i = 1, 2, \ldots, m; j = 1, 2, \ldots, n)$, a new weapon was brought into sight recently to aid the adversary [3]. Because mobile users are publicly observable, partial trajectory information may be revealed when they travel in public places. For example, information such as "Alice was witnessed to pass by XYZ cafeteria at 3pm" becomes valuable auxiliary knowledge to track the mobile target. Such gathered occasional location information forms partial traces of the tracking targets, and becomes *side information* to $\mathcal{A}$, denoted by $\mathcal{S}_{v_j}(t_i^{'}), (i = 1, 2, \ldots, \kappa; j = 1, 2, \ldots, \pi)$. The goal of $\mathcal{A}$ is to ***identify the target mobile user in the trajectory file based on side information matching, and learn the complete footprints left by the tracking target***. For example, in Figure 3, suppose $\mathcal{A}$ obtains user v_j's side information $\mathcal{S}_{v_j}(t_3) = \mathcal{P}_3, \mathcal{S}_{v_j}(t_4) = \mathcal{P}_4$, and $\mathcal{S}_{v_j}(t_5) = \mathcal{P}_5$. If $\mathcal{A}$ has already learnt the whole trajectory record from t_1 to t_5 at $\mathcal{P}_1$ through $\mathcal{P}_5$ belonging to some user u_x, by performing side information matching, $\mathcal{A}$ will immediately know that v_j is u_x, and $\mathcal{P}_1$ and $\mathcal{P}_2$ have also been visited by v_j. Therefore, the whole trajectory of v_j is compromised. It must be noted that while the trajectory files contain accurate location records for service purposes, the side information may be noisy or even incorrect. This is because the source of the side information is unreliable, e.g., personal encounter or context inference.

With this established adversary model, we are now able to present our privacy protection goal as follows: to prevent adversary $\mathcal{A}$ from learning the tracking target's complete trajectory associated with true identity, given partial trajectory is exposed to $\mathcal{A}$. In the next section, we will present how to quantify this protection goal and build the formal mathematical model to solve the problem.

IV. Traffic-aware multiple mix zone placement

We model the location map with POIs as an undirected graph $G(V, E)$, where V is the set of vertices representing the registered POIs, $\{\mathcal{P}_i\}, i = 1, 2, \ldots, n$, and E is the set of road segments that links consecutive POIs. All vertices in G are considered as potential mix zone deployment locations. A trajectory record belonging to u_x defines a path consisting of one or a sequence of possibly repeated vertices. Similarly, a piece of side information corresponding to v_y is a portion of some specific trajectory in G. $\mathcal{P}_i$ and index i are used interchangeably to refer to a POI in the following sections.

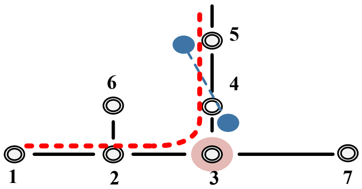


Fig. 4. User trajectory and pseudonym associated vertex pairs, e.g., 1 and 2, 4 and 5, and etc. Concentric circles: POIs in a graph. Edges: road segments connecting POIs. Shaded circle: a mix zone. Dashed line: user trajectory. Solid dots connected by dashed line: side information.

A. Privacy metric

In a graph G, two vertices are pairwise connected when there is at least one path connecting them. In a LBS system, if a user using one pseudonym from $\mathcal{P}_x$ can travel to $\mathcal{P}_y$ *without going through a mix zone* and changing pseudonym, $\mathcal{P}_x$ and $\mathcal{P}_y$ are *pairwise associated*. Using a binary variable $\psi_{ij} \in \{0, 1\}$ to indicate the association status of two POIs, if $\mathcal{P}_x$ and $\mathcal{P}_y$ are pairwise associated $\psi_{xy} = 1$; otherwise, $\psi_{xy} = 0$. Take Figure 4 as an example, suppose Alice travels from $\mathcal{P}_1$ to $\mathcal{P}_5$ using the pseudonym u_x, without any mix zone deployed in between, we say $\mathcal{P}_1$ and $\mathcal{P}_5$ are pairwise associated. Similarly, $\mathcal{P}_1$ and $\mathcal{P}_4$, $\mathcal{P}_3$ and $\mathcal{P}_4$, are also pairwise associated. An important implication of the pairwise association is that, if u_x appears at $\mathcal{P}_1$, u_x can only appear at locations that are pairwise associated to $\mathcal{P}_1$. Furthermore, once the adversary discovers Alice's pseudonym at $\mathcal{P}_1$, locations that are pairwise associated to $\mathcal{P}_1$ will definitely be compromised if u_x visited them. Given that users change pseudonyms in mix zones, and pseudonyms are unique, placing a mix zone at $\mathcal{P}_3$ will break the pairwise association and protect Alice's future locations, $\mathcal{P}_4$ and $\mathcal{P}_5$, even if her identity is revealed at $\mathcal{P}_1$. We use the *total number of pairwise associations* in the graph as a privacy metric to quantify the system's privacy protection level. It is given by:

$$\Phi = \sum_{i,j \in \{\mathcal{P}_i\}} \psi_{ij} \tag{2}$$

Our goal in case of the side information based attack is to deploy multiple mix zones to minimize Φ to ensure the maximum protection level for mobile users. Hence, when a user exposes his identity at some point, only limited trajectory can be disclosed by an adversary. Note that there might be multiple paths connecting two vertices in G. The two vertices are dissociated only when all paths in between are blocked by mix zones.

B. Mix zone deployment constraints

The maximum protection level is achieved when mix zones are deployed at all vertices in G. By doing so, when adversary $\mathcal{A}$ discovers Alice's partial trajectory using side information, an immediate pseudonym change can prevent $\mathcal{A}$ from learning Alice's future locations. However, deploying mix zones adds certain cost to LSI, e.g., pseudonym transformation for every user in the mix zone area, saving state information, and informing application servers of newly arrived users. Moreover, mix zones also result in Quality-of-Experience (QoE) degradation perceived by users. When Alice passes by a mix zone area, she might lose services temporarily due to pseudonym changes and synchronization. For these reasons, deploying mix zones at all POIs is both expensive and inefficient. We need to strategically plan mix zone placement locations in the system to achieve the maximum location privacy protection subject to cost and service constraint.

Following the mapping from location privacy protection to graph model, the multiple mix zone placement problem is formulated as an optimization problem, in which the objective function is to *minimize the overall number of associated vertex pairs*. The assumption behind this objective function is that side information may include true identity exposures at any POI. We do not make any probability assumption for side information exposure in advance and the objective function quantifies the global protection effectiveness of deploying multiple mix zones in G.

Cost and service constraints. As mentioned before, ψ_{ij} is a binary variable indicating whether there is a path association between vertex i and j in G. Let d_i be another binary variable associated with each vertex i in G. $d_i = 1$ indicates vertex i is selected to be a mix zone; otherwise, $d_i = 0$. Considering cost and service constraints posed on LSI, we limit the number of mix zones to be deployed to be at most K. The constraint is expressed as:

$$\sum_{i \in V} d_i \leq K \tag{3}$$

Graph related constraints. The first graph constraint considers two vertices connected by an edge in G. If there is an edge connecting i and j, then there will be a pairwise association between i and j; otherwise, at least one of them should be deployed as a mix zone. That is:

$$\psi_{ij} + d_i + d_j \geq 1 \ \ \forall (i,j) \in E \tag{4}$$

The second graph constraint concerns all vertex triplets. Specifically, the pairwise association is transitive for all vertices in V. If vertex i and j are pairwise associated, and j and k are pairwise associated, then i and j are pairwise associated, meaning there must be some path $i \rightsquigarrow j \rightsquigarrow k$ that a user can travel through without entering into a mix zone. This constraint is described as:

$$\psi_{ij} + \psi_{jk} + \psi_{ki} \neq 2 \ \ \forall (i,j,k) \in V \tag{5}$$

Traffic related constraints: When traffics are not uniformly distributed around the service coverage area, the difficulty of inferential attack conducted by adversary $\mathcal{A}$ varies significantly. For example, suppose $\mathcal{A}$ observes Alice drives on Main Street at 9:50am, and only one location update belonging to user u_x was recorded in the trajectory profile. Then $\mathcal{A}$ will easily associate u_x with Alice. We use entropy to represent the uncertainty for $\mathcal{A}$ to guess which pseudonym belongs to Alice. It quantifies the inherent attacking resilience for each element in graph G. First, the entropy for a road segment is defined as follows:

$$H_r = -\sum_{u} p_u \log p_u \tag{6}$$

where p_u corresponds to the probability that the identity contained in the side information matches to a particular pseudonym on the road segment.

In addition to road segment entropy, pairwise entropy is useful to describe the uncertainty that an adversary finds out a user has visited both POIs of an associated POI pair. Before defining pairwise entropy, we first clarify the concept of path entropy. A path τ consists of consecutive intermediate vertices between two associated vertices and it has no cycle. The entropy for τ is the expected uncertainty in determining if a user has traveled this path or not. Denote p_{r_i} as the probability that the user's side information is leaked on the ith road segment with road segment entropy H_{r_i}, we have:

$$H_\tau = \sum_{i} H_{r_i} \times p_{r_i} \tag{7}$$

Since there may be multiple paths connecting two vertices, we denote p_{τ_i} as the probability that the user's side information is leaked on the ith path. The pairwise entropy between two vertices is then calculated as:

$$H_p = \sum_{i} H_{\tau_i} \times p_{\tau_i} \tag{8}$$

If two vertices have very low pairwise entropy, i.e., they are highly correlated, then we should consider deploying a mix zone to isolate them from other POIs. By doing so, when a user Alice exposes her identity at these two POIs, she can change pseudonym immediately to prevent further location information exposure. A mix zone deployment is considered to be effective only when it satisfies the minimum pairwise entropy requirement. Our proposed model for optimal mix zone placement is traffic-aware because it takes traffic density and entropy into consideration when examining the graph. Specifically, two constraints are defined to ensure the effectiveness of mix zone deployment. First, a mix zone deployed at each vertex on the graph should exceed the predefined entropy threshold ξ_d:

$$(1 - d_i) \times M > \xi_d - e_i \ \ \forall i \in V \tag{9}$$

where M is a very large constant, and e_i is the entropy for location i. In addition to the vertex entropy constraint, we define the following pairwise entropy constraint in our model:

$$(1 - \psi_{ij}) \times M > \xi_p - \vartheta_{ij} \ \ \forall i,j \in V \tag{10}$$

where ξ_p is a predefined threshold, and ϑ_{ij} is the pairwise entropy for i and j.

C. Optimal placement of mix zones

Combining the objective function and all constraints, we derive a formal Integer Linear Programming (ILP) formulation for our Traffic-aware Multiple Mix Zone Placement (TMMP) problem. The complete formulation is described as follows:

$$\begin{array}{lll}
\text{Minimize} & \sum_{i,j \in V} \psi_{ij} & \\
 & & \\
\text{Subject to} & \psi_{ij} + d_i + d_j \geq 1 & \forall (i,j) \in E \\
 & \psi_{ij} + \psi_{jv} + \psi_{vi} \neq 2 & \forall (i,j,v) \in V \\
 & \sum_{i \in V} d_i \leq K & \\
 & (1 - d_i) \times M > \xi_d - e_i & \forall i \in V \\
 & (1 - \psi_{ij}) \times M > \xi_p - \vartheta_{ij} & \forall i,j \in V \\
 & \psi_{ij} \in \{0,1\} & \forall i,j \in V \\
 & d_i \in \{0,1\} & \forall i \in V
\end{array}$$

The ILP formulation of TMMP falls into the category of NP-hard problems [9]. We propose two heuristic algorithms in the next section to calculate the optimal mix zone placement for TMMP.

V. Heuristic algorithms

A common technique to solve the ILP formulation of TMMP is to relax the binary constraint $\psi_{ij}, d_i \in \{0,1\}$ to a pair of linear constraints $0 \leq \psi_{ij}, d_i \leq 1$. By doing so, the original NP-hard problem is transformed to a Linear Program (LP) that is solvable in polynomial time. In general, the optimal solution derived from solving LP does not have all variables either 0 or 1. It cannot be directly used to answer TMMP. In this section, we develop two heuristic algorithms to provide approximate solutions for TMMP. The first algorithm assumes uniform traffic pattern over the network, and solves TMMP formulation without constraints (9) and (10). We denote this heuristic algorithm as **U**niform **T**raffic **M**ix Zone **P**lacement (UTMP). It provides an estimation of achievable privacy level when no knowledge about traffics is available. The second heuristic algorithm aims at solving the complete TMMP when LSI obtains enough traffic information over the target area. We name it as **N**on-**U**niform **T**raffic **M**ix Zone **P**lacement (NUTMP). Both UTMP and NUTMP share the same set of inputs including the abstracted POI graph $G = (V, E)$ and the maximum mix zone number K, which is typically less than the number of vertices in G. NUTMP requires additional input of entropy information to take traffic into account. Their output is a set Ω containing mix zone placement locations in G.

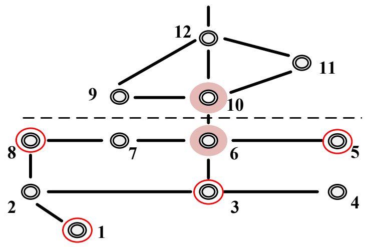


Fig. 5. An execution snapshot of our heuristic algorithms. Concentric circles: POIs in a graph. Edges: road segments connecting POIs. Shaded vertices: articulation points. Circled vertices: maximal independent set in bottom part of the graph.

A. *Uniform traffic mix zone placement (UTMP)*

In real world, any POI should be reachable from any other POI in the target area. Thus the area graph is connected without isolated points. We can see that the total number of possible pairwise connections in such a graph of n vertices is $O(n^2)$. The first step in UTMP is built on the observation that partitioning G into several disconnected components is helpful to eliminate the pairwise connections across these components. Therefore, we are seeking for vertices whose removal disconnect the graph. Such vertices are typically referred to as *articulation points* in graph theory. Take the area graph in Figure 5 as an example. Any route from 1 to 9 or from 1 to 12 needs to go through vertices 6 and 10. Therefore, 6 and 10 are articulation points in this graph. If a mix zone is deployed at vertex 6 or 10, a pseudonym appears at any vertex in the bottom part of the graph cannot appear at vertices 9, 12, and 11. Hence, the total number of pairwise associations is reduced.

After G is partitioned into disconnected components, the mix zone deployment in each component is further refined to improve the solution quality. In graph theory, an independent set refers to a set of vertices that are not adjacent to each other. Hence, if all vertices that are *not* in an independent set are selected as mix zones, there will be no pairwise association between the vertices in the independent set. Again, refer to the bottom part of Figure 5 as an example. Circle highlighted vertices, $\{1, 8, 3, 5\}$, form a maximal independent set for the lower part of the graph. If vertices $\{2, 4, 6, 7\}$ are selected as mix zones, a user Alice's pseudonym u_x appears at vertex 1 will not appear at any other vertex in the independent set. As a result, Alice's past and future locations on her trajectory are protected, even though her identity get exposed at vertex 1. Finally, we need to control the number of mix zones to meet the cost and service constraint. At the last step of our algorithm, we iteratively remove the vertex that introduces the least number of pairwise association increment from the mix zone candidate set selected by previous steps until constraint (3) is met. Algorithm 1 summarizes the proposed UTMP algorithm.

Algorithm 1: Uniform traffic mix zone placement (UTMP)

```
input  : A graph $G = (V, E)$ and $K$
output: A set $\Omega$ of at most $K$ selected mix zone
        positions
/* --Step #1: Find articulation points--  */
Depth first search for $G$ to find discover time $i.d$ for
each vertex;
for each vertex $i$ in $G$ do
    $i.\nu \leftarrow \min\{i.d, \min_{backedge\ i \to w}\{w.d\}\}$ ;
end
Initialize articulation points set $\Lambda \leftarrow \varnothing$;
for each vertex $i$ in $G$ do
    if $i.\nu \geq i.d$ then
        $\Lambda \leftarrow \Lambda \cup \{i\}$;
    end
end
$\Omega \leftarrow \Omega \cup \Lambda$;
/* --Step #2: Maximal independent set--   */
Find maximal independent set $\mathcal{I}_{\mathcal{C}_j}$ for each connected
component $\mathcal{C}_j$ by iteratively adding non-adjacent vertices;
$\mathcal{I} \leftarrow \cup \mathcal{I}_{\mathcal{C}_j}$;
$\Omega \leftarrow \Omega \cup \{V \setminus \mathcal{I} \setminus \Lambda\}$;
/* --Step #3: Maintain cost constraint--  */
while $|\Omega| > K$ do
    Find vertex $x \in \Omega$ that contributes the least pairwise
    associations to $V \setminus \Omega$, and remove it from $\Omega$;
end
Return $\Omega$;
```

B. *Non-uniform traffic mix zone placement (NUTMP)*

Algorithm 2 summarizes the proposed NUTMP algorithm that further considers the impact of traffic conditions on mix zone deployment effectiveness. Specifically, NUTMP incorporates two filtering procedures in addition to UTMP to guarantee the final solution meets the traffic-related constraints (9) and (10). First, in the articulation point selection step, only those articulation points with entropy values higher than ξ_d are considered as mix zone candidates and put into set Ω. Second, unlike UTMP that selects a maximal independent set as the starting point, in NUTMP, we first choose all vertices that have lower entropy values than ξ_p into a set Ψ so that they cannot be used as mix zones. Then, the vertices that are not articulation points and are not adjacent to any vertex in from

Ψ are put into Ψ. The reason for this step is similar to the maximal independent set selection in UTMP. By adding non-adjacent vertices to Ψ, no pairwise association is introduced (if all others are mix zones). It is possible that the vertices not qualified to become mix zones are adjacent to each other. If the threshold values are set appropriately, the pairwise entropy constraint should be satisfied in this step. Let Ω become $(V \setminus \Psi)$. By iterating through all mix zone candidates in Ω, we remove those vertices that satisfy the pairwise entropy constraint and incur the least number of pairwise association increment until mix zone cost constraint (3) is met.

Algorithm 2: Non-uniform traffic mix zone placement (NUTMP)

```
input : A graph $G = (V, E)$, $K$, mix zone entropies,
        and entropy matrix for vertex pairs
output: A set $\Omega$ of at most $K$ selected mix zone
        positions
/* --Step #1: Find articulation point--   */
Find articulation points set $\Lambda$ as in Algorithm 1;
Remove the articulation points that have entropy value
less than $\xi_d$ from $\Lambda$;
/* --Step #2: Non-mix-zone vertices
   selection--                            */
Put all vertices with entropy values less than $\xi_d$ into $\Psi$;
Select vertices from $V \setminus \Lambda \setminus \Psi$ that are not adjacent to
any vertex in $\Psi$, and put them into $\Psi$;
$\Omega \leftarrow V \setminus \Psi$;
/* --Step #3: Maintain cost constraint--  */
while $|\Omega| > K$ do
    Find vertex $x \in \Omega$ that satisfies the pairwise entropy
    constraint and incurs the least pairwise association
    increase, and remove it from $\Omega$;
end
Return $\Omega$;
```

C. Complexity analysis

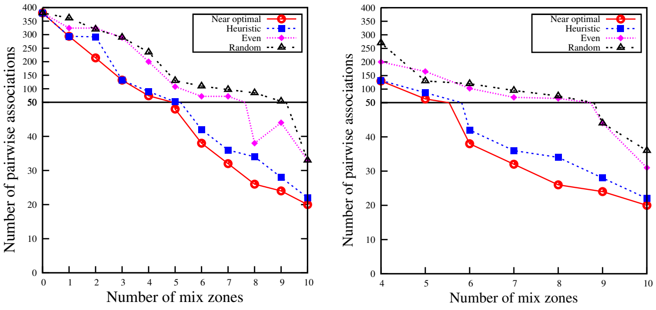
The complexity of both UTMP and NUTMP algorithms consists of mainly three components. First, the method for finding all articulation points in G is an algorithm suggested by and analyzed in [9]. Its complexity is $O(E)$. Second, finding a maximal independent set by iteratively adding vertices that are not adjacent to current selected vertices requires only linear time in both heuristic algorithms. Step 3 in both UTMP and NUTMP are similar to the critical node detection algorithm proposed in [10], which has complexity $O(|V|^2|E|)$. As a result, the overall complexity for UTMP and NUTMP are both $O(|V|^2|E|)$.

VI. Performance Evaluation

In this section, we present the simulation results of the proposed UTMP and NUTMP algorithms. Both algorithms are implemented in C++. Due to the differences in privacy metrics used and problem formulation, it is difficult to conduct direct performance comparison with some existing works, e.g., [8], [11], [12]. To evaluate the solution quality of UTMP and NUTMP, we compare the results with the near optimal solution obtained from CPLEX$^{\text{TM}}$ [13] using standard techniques, e.g., branch-and-bound algorithm. For trajectory generation, we adopt the real world mobility trace of San Francisco Bay area cabs from CRAWDAD [14]. The partial road map of the same area is abstracted as our input graph. We select 20 POIs from the map covering a diverse location types, e.g., road intersections, hospitals, and bars/coffee shops.

A. Protection Effectiveness

First, we compare the solution quality of both UTMP and NUTMP to the near optimal solution derived by CPLEX$^{\text{TM}}$ (marked as "near-optimal"). Besides, we also include the simulation results for randomly selected mix zone locations (marked as "random"), and selecting representative mix zones from K evenly partitioned components in G (marked as "even"). The input graph is shown in Figure 7, where all POIs are potential mix zone deployment locations. We evaluate the protection effectiveness for K ranging from 0 to 10. For the NUTMP algorithm, 20% of the edges and 10% of the vertices are randomly selected as low-traffic locations. Their entropy values are drawn from the normal distribution of $\mathcal{N}(1, 0.5)$, and the entropy values for the rest are drawn from the normal distribution of $\mathcal{N}(4, 0.5)$. Figure 6 shows the reduction in total number of pairwise associations when different number of mix zones are deployed in the system. As expected, the number of pairwise associations decreases with the increased number of mix zones in all four methods, under both uniform and non-uniform traffic assumptions. We observe that both UTMP and NUTMP perform very close to the near optimal solution. When the number of the selected mix zones is larger than 4, the average difference of pairwise associations between our heuristic algorithms and the near optimal solutions provided by CPLEX$^{\text{TM}}$ is less than 10%. Because entropy constraints for both vertex and incident edges are taken into account in NUTMP, its outcome is in general different from UTMP. Mostly the value derived from NUTMP is higher than that in UTMP. A possible explanation for this phenomenon is that the ideal locations for minimizing pairwise associations in UTMP may not be qualified in NUTMP because of the traffic-related constraints. Finally, when K becomes larger, the possibility

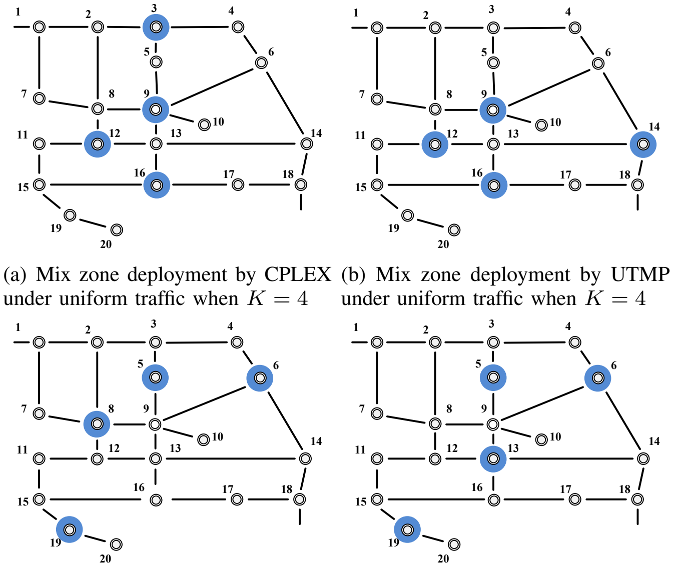


(a) Uniform traffic condition

(b) Non-uniform traffic condition

Fig. 6. Total number of pairwise associations

of selection overlapping increases for all methods. Hence, we observe that both "random" and "even" approach performs fairly well when K is large. Figure 7 presents an example mix zone selection result to compare the near-optimal solution and our heuristic algorithms. We can see that, the majority of the locations are overlapped. Since the assigned entropy values are low for edges $3 \leftrightarrow 5$ and $19 \leftrightarrow 20$, and for vertices 3 and 20, vertex 3 is not selected in Figure 7(c) and Figure 7(d). Instead, vertex 19 is selected in Figure 7(c) and Figure 7(d) to satisfy the traffic constraint. When the number of mix zones becomes larger, the selected location sets exhibit more overlap. This is the same trend exhibited in Figure 6(a) and Figure 6(b), where the number of pairwise associations between optimal and heuristic become very close.



(a) Mix zone deployment by CPLEX under uniform traffic when $K = 4$

(b) Mix zone deployment by UTMP under uniform traffic when $K = 4$

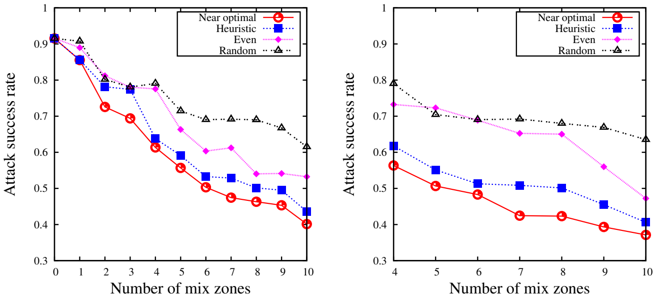
(c) Mix zone deployment by CPLEX under non-uniform traffic when $K = 4$

(d) Mix zone deployment by NUTMP under non-uniform traffic when $K = 4$

Fig. 7. Comparison of mix zone locations between CPLEX's solution and heuristic algorithms

B. Resilience to side information based attack

Utilizing the mix zone placement selection results presented in the last section, we conduct another set of simulations to investigate the systems' resilience to side information based attacks. We randomly select 500 partial mobility traces from the San Francisco Bay area cab's mobility traces in CRAWDAD [14]. Each of them is recorded with a distinct pseudonym. These mobility traces simulate users' trajectories in the input graph. Since the trace file is recorded in <time,coordinates> format, we consider a user stepping onto the corresponding vertex in G, when his trace appears within a certain range of one of the marked POIs. Similarly, the coordinates of a user's trace between two POIs are interpolated and mapped to the closest edge in G. We randomly select some portion of the selected user mobility traces to generate 100 shorter trajectories as side information. Each side information belongs to a particular ID that serves as the true identity of a user. Since real world side information often contains noises [3], we obfuscate the generated side information to better simulate this effect. The maximum likelihood estimation approach for adversary $\mathcal{A}$ is implemented as described in [3] to simulate the side information based inferential attack. An attack is successful if the adversary finds out the corresponding pseudonym used by a user in the side information. The success rate of an adversary is the ratio of number of successful attacks over total number of attacks. Figure 8 shows the attack success rate when different number of mix zones are deployed in the target area. According to [3], this type of inferential attack has high success rate when no mix zone is deployed. Using our mix zone deployment algorithms, we observe that the attack success rate can be reduced to about 50% of original value. Moreover, the difference between our heuristic algorithms and the near optimal solution provided by CPLEX^{TM} is only about 10% on average. The reason is that, previously, a piece of side information may be able to be matched back to its original mobility trace with high probability. When more mix zones are deployed, this mobility trace may be broken into more pieces of shorter trajectories. It is difficult to find the best match because the side information now faces many possibilities with these broken trajectories under different pseudonyms. From Figure 8 we can see that both UTMP and NUTMP achieve nearly the same protection effect as the near-optimal solutions, and result in lower attack rate than the other two approaches. Moreover, from Figure 8(b) we can see that when traffic intensity is considered, better protection effectiveness is achieved. The reason is that when a road segment has high traffic intensity, it is hard to distinguish users on the road with or without the help of side information. Therefore, the traffic-related constraint provides another level of protection to privacy attack.



(a) Uniform traffic condition

(b) Non-uniform traffic condition

Fig. 8. Attack success rate under different traffic and mix zone deployment situations

VII. Related work

Location privacy issues in mobile computing environments have received significant attention in recent years. An early study [15] showed that location-tracking LBS (locations are tracked by other parties) generates more concerns of privacy leaking than position-aware LBS (device's self-awareness of its current location) for mobile users. Hence, most existing works focus on the location-tracking LBS model and assume the presence of a centralized trusted anonymization server. The most popular technique to achieve desired level of privacy protection is to degrade the resolution of location information in a controlled way. This has led to a large number of location perturbation and obfuscation schemes proposed in the last decade. For example, spatial cloaking [16], [17] allows obfuscation of a mobile user's exact location using cloaked spatial areas to meet anonymity constraints, e.g., k-anonymity. However, spatial cloaking may result in a severe degradation of service quality due to the large cloaked area over an extended time period [18], and is not suitable to protect privacy in the network-constrained mobile environments such as road networks [19].

An alternative approach for location perturbation and obfuscation is to restrict locating of mobile user position in certain areas, known as the mix zone model [15]. A mix zone often covers a small area, e.g., a road intersection, and allows users to change pseudonyms within the area. Due to its ability to reduce the linkability between identity and trajectory, mix zone deployment over road intersections has gained popularity in vehicular networks. Given the presence of a global passive adversary, Freudiger et al. [6] proposed the CMIX protocol to create cryptographic mix zones at road intersections. Dahl et al. [20] improved the cryptographic approach by fixing the key establishment protocol in CMIX. A more sophisticated protocol, MobiMix [7], improved attack resilience by considering various factors, e.g., traffic density, user moving patterns, and etc. All these approaches do not consider the optimal placement of multiple mix zones. Huang et al. [11] proposed the use of cascading mix zones. Their investigation focused on evaluating the QoS implication on real-time applications, rather than protection effectiveness of using multiple mix zones. Shin et al. [12] proposed a request partitioning method to increase the unlinkability of different requests over time. The most related research in [8] analyzed the optimal placement of multiple mix zones with combinatorial optimization techniques. Our work is significantly distinctive in the following aspects: (1) compared with the flow-based metric used in [8], the cumulated pairwise location associativity is more appropriate to capture the global placement effects; (2) using on this metric, our optimal placement strategy is capable of handling a recently emerging side information based attacking model [3] in addition to the simple passive adversary model; and (3) we consider the impact of traffic density at each mix location to enhance the attack resilience.

VIII. Conclusion

This paper investigated the optimal multiple mix zones placement problem for location privacy protection. We modeled the area covered by location-based services as a graph, where all vertices (POIs) are considered as candidates for mix zone deployment. In order to protect mobile users from side information based inferential attacks, we propose to use pairwise vertex association to characterize the linkability of the POIs along a user's trajectory on the map. To achieve maximum privacy protection, we formulated the optimization problem with the objective of maximizing the overall discontinuity of all possible trajectories on the road network and subject to deployment cost and traffic constraints. For each road segment and intersection, the traffic density effect in terms of entropy is also taken into account. We designed two heuristic algorithms for practical and efficient solutions to the NP-hard optimization problem. Simulation results based on realistic mobile user data traces show that our solution yields satisfactory performance in reducing the success rate of inferential attacks. The mathematical modeling and performance results presented in this paper offer both theoretical and practical guidance to multiple mix zones placement in mobile networks for protecting users' location privacy.

References

[1] "The $10 Billion Rule: Location, Location, Location," URL: http://www.strategyanalytics.com.

[2] M. Terrovitis and N. Mamoulis, "Privacy preservation in the publication of trajectories," in *Proc. of MDM'08*, 2008, pp. 65–72.

[3] C. Y. Ma, D. K. Yau, N. K. Yip, and N. S. Rao, "Privacy vulnerability of published anonymous mobility traces," in *Proc. of MobiCom'10*, 2010.

[4] A. Beresford and F. Stajano, "Location privacy in pervasive computing," *IEEE Pervasive Computing*, vol. 2, no. 1, pp. 46–55, 2003.

[5] ——, "Mix zones: User privacy in location-aware services," in *Proc. of the 2nd IEEE Annual Conference on Pervasive Computing and Communications Workshops (PERCOMW '04)*, 2004, pp. 127–131.

[6] J. Freudiger, M. Raya, M. Félegyházi, P. Papadimitratos, and J. Hubaux, "Mix-zones for location privacy in vehicular networks," in *Proc. of the 1st International Workshop on Wireless Networking for Intelligent Transportation Systems (WiN-ITS'07)*, 2007.

[7] B. Palanisamy and L. Liu, "MobiMix: Protecting Location Privacy with Mix-zones over Road Networks," in *Proc. of ICDE'11*, 2011, pp. 494–505.

[8] J. Freudiger, R. Shokri, and J.-P. Hubaux, "On the optimal placement of mix zones," in *Proc. of the 9th International Symposium on Privacy Enhancing Technologies (PETS'09)*, 2009, pp. 216–234.

[9] T. Cormen, *Introduction to algorithms*. The MIT press, 2001.

[10] A. Arulselvan, C. Commander, L. Elefteriadou, and P. Pardalos, "Detecting critical nodes in sparse graphs," *Computers & Operations Research*, vol. 36, no. 7, pp. 2193–2200, 2009.

[11] L. Huang, H. Yamane, K. Matsuura, and K. Sezaki, "Silent cascade: Enhancing location privacy without communication qos degradation," in *Proc. SPC'06*, 2006, pp. 165–180.

[12] H. Shin, J. Vaidya, V. Atluri, and S. Choi, "Ensuring privacy and security for LBS through trajectory partitioning," in *Eleventh International Conference on Mobile Data Management*. IEEE, 2010, pp. 224–226.

[13] "IBM ILOG CPLEX Optimizer [Online]," Available: http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/.

[14] D. Kotz, T. Henderson, and I. Abyzov, "CRAWDAD data set dartmouth/campus (v. 2004-12-18)," Downloaded from http://www.crawdad.org/dartmouth/campus.

[15] L. Barkhuus and A. Dey, "Location-based services for mobile telephony: a study of users' privacy concerns," in *Proc. of the 9th IFIP TC13 International Conference on Human-Computer interaction (INTERACT'03)*, 2003.

[16] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *Proc. of MobiSys'03*, 2003, pp. 31–42.

[17] B. Gedik and L. Liu, "Location privacy in mobile systems: A personalized anonymization model," in *Proc. of ICDCS'05*, 2005, pp. 620–629.

[18] R. Cheng, Y. Zhang, E. Bertino, and S. Prabhakar, "Preserving user location privacy in mobile data management infrastructures," in *Proc. of the 6th Workshop on Privacy Enhancing Technologies*, 2006, pp. 393–412.

[19] T. Wang and L. Liu, "Privacy-aware mobile services over road networks," *Proc. VLDB Endow.*, vol. 2, no. 1, pp. 1042–1053, 2009.

[20] M. Dahl, S. Delaune, and G. Steel, "Formal analysis of privacy for vehicular mix-zones," in *Proc. of the 15th European conference on Research in computer security (ESORICS'10)*, 2010, pp. 55–70.