Defining Perfect Location Privacy Using Anonymization

Zarrin Montazeri
Electrical and Computer
Engineering Department
University of Massachusetts
Amherst, Massachusetts

Email: seyedehzarin@umass.edu

Amir Houmansadr
College of Information and
Computer Sciences
University of Massachusetts
Amherst, Massachusetts
Email: amir@cs.umass.edu

Hossein Pishro-Nik
Electrical and Computer
Engineering Department
University of Massachusetts
Amherst, Massachusetts

Email: pishro@engin.umass.edu

***Abstract*—The popularity of mobile devices and location-based services (LBS) has created great concerns regarding the location privacy of users of such devices and services. Anonymization is a common technique that is often being used to protect the location privacy of LBS users. In this paper, we provide a general information theoretic definition for location privacy. In particular, we define perfect location privacy. We show that under certain conditions, perfect privacy is achieved if the pseudonyms of users is changed after $o(N^{\frac{2}{r-1}})$ observations by the adversary, where N is the number of users and r is the number of sub-regions or locations.**

I. Introduction

Mobile devices capable of communicating over the Internet with high-precision localization capability have become pervasive in the past several years. These communicating mobile devices provide a wide range of services based on the *geographic location* of the user. We refer to such services that use the geographic location of their users as *location-based services (LBS)*.

While LBSes provide so many services to their users thanks to their unrestricted access to the location information of the users, they also impose significant privacy threats to them. Some mechanisms have been proposed in order to protect location privacy of LBS users, [1]–[7], generally referred to as *location privacy protection mechanisms (LPPM)*.

Today's LPPMs can be classified into two main categories: *identity perturbation* LPPMs [5]–[7] modify the identity of mobile users in order to protect their location privacy (e.g., through anonymization techniques), in other words they aim at improving location privacy by concealing the mapping between users and location observations. *Location perturbation* LPPMs are the second category [1]–[4], [7], which add noise to mobile users' location coordinates. This can potentially improve location privacy by returning inaccurate location information to the LBS applications. Some LPPMs combine both mechanisms to address this problem, but this may degrade the *performance* of an LBS system. Unfortunately, despite previous studies on location privacy, the design of LPPM systems relies on ad-hoc algorithmic heuristics such as adding noise and shuffling identities.

In this paper, we propose a fundamental, analytical study of location privacy for location-based services. We assume the strongest model for the adversary, i.e., an adversary who has complete statistical knowledge of the users' movements. Then, we define location privacy based on the mutual information between the adversary's observation and actual location data. This allows us to define *perfect location privacy* wherein users have provably private locations. Then, we show that for r possible i.i.d locations, if the adversary obtains less than $o(N^{\frac{2}{r-1}})$ observations per user, then all users have *perfect location privacy* at all time. We show that perfect location privacy is indeed achievable if the LPPMs are designed appropriately.

II. Related Work

Existing work on the design of LPPM mechanisms can be classified into two main categories of *identity perturbation* LPPMs [5]–[7] and *location perturbation* LPPMs [1]–[4], [7]. Location perturbation LPPMs add noise to users' location coordinates while identity perturbation LPPMs modify the identities of mobile users.

A common approach used by identity perturbation LPPMs is to obfuscate user identities within a group of users, an approach known as k-anonymity [2], [8]. A second common approach to identity perturbation LPPMs is to exchange users' pseudonyms within specific areas called mix-zones [9], [10]. Freudiger et al. show that combining techniques from cryptography with mix-zones can result in higher levels of location privacy [5]. Also, Manshaei et al. use game theoretic approaches to improve the location privacy protection provided by mix-zones [11]. Location cryptography is another direction taken towards protecting location information [12].

Many proposed location perturbation LPPMs work by replacing each user's location information with a larger region, a technique known as cloaking [2], [3], [13], [14]. Another direction to location perturbation is including dummy locations in the set of possible locations of users [15], [16].

Several works [4], [17]–[20] use differential privacy to protect location privacy in location information datasets. This insures that the presence of no single user could significantly change the outcome of the aggregated location information.

For instance, Ho et al. [21] proposed a differentially private location pattern mining algorithm using quadtree spatial decomposition.

Dewri [22] combined k-anonymity and differential privacy to improve location privacy. Some location perturbation LPPMs are based on ideas from differential privacy [4], [23]–[26]. For instance, Andres et al. hide the exact location of each user in a region by adding Laplacian distributed noise to achieve a desired level of geo-indistinguishability [26].

Several works aimed at quantifying location privacy protection. Shokri et al. [7], [27] define the expected estimation error of the adversary as a metric to evaluate LPPM mechanisms. On the other hand, Ma et al. [6] uses uncertainty about users' location information to quantify user location privacy in vehicular networks.

III. Framework

A. Defining Location Privacy

To investigate the location privacy problem, we first need to provide a generic mathematical definition for location privacy. Consider a network consisting of N users, and suppose that an LPPM is used to protect the privacy of the users. Let $\mathcal{A}$ be an adversary who is interested in knowing the locations of the users as they move. To ensure privacy, we assume the strongest adversary in the sense that we assume the adversary has complete statistical knowledge of the users' movements. That is, through previous observations or other sources, the adversary has a complete model that describes the movement of users as a random process on the corresponding geographic area.

Now, starting at time zero, the users move through the area. In particular, let $X_i(t)$ be the location of user i at time t. Adversary $\mathcal{A}$ is interested in knowing $X_i(t)$ for $i = 1,2,...,N$. However, he can only observe the anonymized and obfuscated versions of $X_i(t)$'s produced by the LPPM. In particular, let $\mathbf{Y}$ be a collection of observations available to the adversary. We define *perfect location privacy* as follows:

Definition 1. User i has perfect location privacy at time t with respect to adversary $\mathcal{A}$, if and only if

$$\lim_{N\to\infty} I(X_i(t);\mathbf{Y}) = 0,$$

where $I(.)$ shows the mutual information.

The above definition requires that the observations of the adversary does not give her any useful information about the location of user i. It also assumes a large number of users ($N \to \infty$). This assumption is valid for almost all applications that we consider.

In this paper, to achieve location privacy, we use only anonymization techniques. That is, we perform a random permutation $\Pi^{(N)}$ on the set of N users, and then assign the pseudonym $\Pi^{(N)}(i)$ to user i.

$$\Pi^{(N)} : \{1,2,\cdots,N\} \to \{1,2,\cdots,N\}$$

Throughout the paper, we assume the permutation $\Pi^{(N)}$ is chosen uniformly at random among all $N!$ possible permutations. For simplicity of notations we sometimes drop the superscripts, e.g., $\Pi^{(N)} = \Pi$.

For $i = 1,2,\cdots,N$ let $\mathbf{X}_i^{(M)} = (X_i(1), X_i(2),\cdots,X_i(M))^T$ be a vector which shows the i^{th} user's locations at times $1,2,\cdots,M$. The adversary observes a permutation of users location vectors, $\mathbf{X}_i^{(M)}$'s, using the permutation function $\Pi^{(N)}$. In other words, the adversary observes

$$\begin{aligned}\mathbf{Y}^{(M)} &= \text{Perm}(\mathbf{X}_1^{(M)},\mathbf{X}_2^{(M)},\cdots,\mathbf{X}_N^{(M)};\Pi^{(N)})\\ &= (\mathbf{X}_{\Pi^{-1}(1)}^{(M)},\mathbf{X}_{\Pi^{-1}(2)}^{(M)},\cdots,\mathbf{X}_{\Pi^{-1}(N)}^{(M)})\\ &= (\mathbf{Y}_1^{(M)},\mathbf{Y}_2^{(M)},\cdots,\mathbf{Y}_N^{(M)})\end{aligned}$$

where,

$$\mathbf{Y}_{\Pi^{(N)}(i)}^{(M)} = \mathbf{X}_i^{(M)} = (X_i(1),X_i(2),\cdots,X_i(M))^T$$

We introduce two lemmas here that will be used to prove the main result through the paper.

Lemma 1. For $k = 1,2,...,$ let $\mathbf{Z}^{(k)} = (Z_1^{(k)},Z_2^{(k)},\cdots,Z_{n(k)}^{(k)})$ be a sequence of independent random vectors with size $n(k) \le k$, such that $n(k) = ak^b$ where $a > 0$ and $0 < b < 1$ are constants. Assume $Z_1^{(k)},Z_2^{(k)},\cdots,Z_n^{(k)}$ are independent discrete random variables with identical range, i.e., $P(Z_i^{(k)} = x) > 0$ if and only if $P(Z_j^{(k)} = x) > 0$. Further, suppose that their distributions $F_{Z_i^{(k)}}$ converge to the standard normal distribution. In particular, for each $\gamma > 0$, there exists $k_0 \in \mathbb{N}$ such that if $k > k_0$, then

$$\sup\left\{|F_{Z_i^{(k)}}(x) - \Phi(x)| : x \in \mathbb{R}, i \in \{1,2,...,n(k)\}\right\} \le \gamma$$

Let $\mathbf{Y}^{(k)}$ be a permuted version of the $\mathbf{Z}^{(k)}$ under the random permutation $\Pi^{(n)}$:

$$\mathbf{Y}^{(k)} = \text{Perm}(Z_1^{(k)},Z_2^{(k)},\cdots,Z_n^{(k)};\Pi^{(n)}),$$

For $\lambda > 0$ we define a set $A_\lambda^{(k)}$ as follows:

$$A_\lambda^{(k)} = \left\{\mathbf{y}^{(k)} : \left|1 - nP(\Pi^{(k)}(1) = j|\mathbf{Y}^{(k)} = \mathbf{y}^{(k)})\right| < \lambda\right\},$$

Then, for any $\lambda > 0$, we have

$$\lim_{k\to\infty} P(\mathbf{Y}^{(k)} \in A_\lambda^{(k)}) = 1$$

Proof. (Sketch) Here, the goal is to study the conditional probability $P(\Pi^{(n)}(1) = j|\mathbf{Y}^{(k)} = \mathbf{y}^{(k)})$. In particular, we want to study the power of an adversary in finding the permutated value of 1, i.e., $\Pi^{(k)}(1)$, based on the observed data $\mathbf{Y}^{(k)}$.

To get the idea behind this lemma, let's assume that Z_i's have exactly normal distribution (instead of considering Z_i's distributions converging to $N(0,1)$). So, all Z_i's are i.i.d random variables.

$$Z_i^{(k)} \sim N(0,1)$$

If we observe $\mathbf{Y}^{(k)}$

$$\mathbf{Y}^{(k)} = \text{Perm}(Z_1^{(k)},Z_2^{(k)},\cdots,Z_n^{(k)};\Pi^{(n)}),$$

then, by considering that $Z_i^{(k)}$'s are i.i.d and using symmetry, probability of finding the right permutation function based on this observation is

$$P(\Pi^{(n)}(1) = j|\mathbf{Y}^{(k)} = y^{(k)}) = \frac{1}{n}.$$

In the lemma, as Z_i's distributions converge to $N(0,1)$, we are not able to say this probability is exactly $\frac{1}{n}$, but it is close to it. This can be shown using the continuity of the probability distribution functions. In particular, we can show

$$\frac{1-\lambda}{n} < P(\Pi(1)^{(n)} = j|\mathbf{Y}^{(n)} = \mathbf{y}^{(n)}) < \frac{1+\lambda}{n}$$

with high probability. Thus, with high probability

$$\left|1 - nP(\Pi(1)^{(n)} = j|\mathbf{Y}^{(k)} = \mathbf{y}^{(k)})\right| < \lambda$$

for any $\lambda > 0$, and that proves the lemma.

□

Lemma 2. Let $Y_i(k) = Y_i^M(k)$ be the adversary's observations as defined above. Let us define $\overline{Y}$ as,

$$\overline{Y}_i = \frac{1}{M}\sum_{k=1}^{M} Y_i(k).$$

Then, given $\overline{Y}_1,\overline{Y}_2,\cdots,\overline{Y}_N$, we have that $\Pi^{(N)}$ and $\mathbf{Y}_N$ are independent.

Proof. This is the immediate result of the fact that

$$(\overline{Y}_1,\overline{Y}_2,\cdots,\overline{Y}_N) = \overline{Y}^{(N)}$$

is a *sufficient statistic* for p_i's. In particular,

$$f(\mathbf{Y}^{(N)}|\overline{Y}^{(N)},\Pi) = f(\mathbf{Y}^{(N)}|\overline{Y}^{(N)})$$

which means that $\mathbf{Y}^{(N)}$ and Π are independent given the average values $\overline{Y}^{(N)}$. □

B. Location Privacy for a Simple Two-State Model

To get a better insight about the location privacy problem, here we consider a simple scenario. Consider a scenario where there are two locations, locations 0 and 1. At any time $k \in \{0,1,2,\cdots\}$, user i has probability $p_i \in (0,1)$ to be at location 1, independently from previous locations and independently from other users' locations. Therefore, $X_i(k) \sim Bernoulli(p_i)$.

To keep things general, we assume that p_i's are drawn independently from some continuous density $f_P(p)$ on the $(0,1)$ interval. Specifically, $f_P(p) = 0$ for all $p \notin (0,1)$ and there are $\delta_2 > \delta_1 > 0$ such that $\delta_1 < f_P(p) < \delta_2$ for all $p \in (0,1)$. The values of p_i's are known to the adversary.

Theorem 1. For two locations with above definition and observation vector $\mathbf{Y}^{(M)}$ if all the following holds,

1) $M = cN^{2-\alpha}$, which $c,\alpha > 0$ and are constant
2) $p_1 \in (0,1)$
3) $(p_2,p_3,\cdots,p_N) \sim f_p$, $0 < \delta_1 < f_p < \delta_2$
4) $P = (p_1,p_2,\cdots,p_N)$ be known to the adversary

then, we have

$$\forall k \in \mathbb{N}, \quad \lim_{N\to\infty} I(X_1(k);\mathbf{Y}^{(M)}) = 0$$

Before providing a formal proof for Theorem 1, let us provide the intuition behind it. Let us look from the adversary's perspective. The adversary would like to obtain $X_1(k)$. The adversary, knows the value of p_1. To obtain $X_1(k)$, it suffices that the adversary obtains $\Pi(1)$. Since $X_i(k) \sim Bernoulli(p_i)$, to do so, the adversary can look at the averages

$$\overline{Y}_{\Pi(i)} = \frac{Y_{\Pi(i)}(1) + Y_{\Pi(i)}(2) + ... + Y_{\Pi(i)}(M)}{M}.$$

In fact, we show in Lemma 2 that $\overline{Y}_{\Pi(i)}$'s provide a sufficient statistics for this problem. Now, intuitively, the adversary is successful in recovering $\Pi(1)$ if two conditions hold

1) $\overline{Y}_{\Pi(1)} \approx p_1$.
2) For all $i \neq 1$, $\overline{Y}_{\Pi(i)}$ is not too close to p_1.

Now, note that by the central limit theorem (CLT),

$$\frac{\overline{Y}_{\Pi(i)} - p_i}{\sqrt{\frac{p_i(1-p_i)}{M}}} \to N(0,1).$$

That is, loosely speaking, we can write

$$\overline{Y}_{\Pi(i)} \to N\left(p_i, \frac{p_i(1-p_i)}{M}\right).$$

Consider and interval $I \in (0,1)$ such that $p_1 \in I$ and the length of I, length(I), is equal to $L_N = \frac{c}{N}$ where $c > 0$ is an arbitrary constant. Note that for any $i \in 1,2,\cdots,N$ the probability that $p_i \in I$ is larger than $\delta L_N = \frac{c\delta}{N}$. In other words, by choosing c large enough, we can guarantee that a large number of p_i's be in I. On the other hand, note that we have

$$\begin{aligned}\frac{\sqrt{\text{Var}(\overline{Y}_{\Pi(i)})}}{\text{length}(I)} &= \frac{\sqrt{\frac{p_i(1-p_i)}{M}}}{\frac{c}{N}} \\ &= \frac{N}{\sqrt{o(N^2)}} \to \infty.\end{aligned}$$

Note that here, we will have a large number of normal random variables $\overline{Y}_{\Pi(i)}$ whose expected values are in interval I with high probability and their standard deviation is much larger than the interval length. Thus, distinguishing between them will become impossible for the adversary. In other words, the probability that the adversary will correctly identify $\Pi(l)$ goes to zero as N goes to infinity. That is, the adversary will most likely choose an incorrect value j for $\Pi(l)$. In this case, since the locations of different users are independent, the adversary will not obtain any useful information by looking at $X_j(k)$.

Proof of Theorem 1. We define $\overline{X}_i$

$$\overline{X}_i = \frac{1}{M}\sum_{k=1}^{M} X_i(k).$$

Since $X_i(k) \sim Bernoulli(p_i)$

$$\mathrm{E}\overline{X}_i = p_i, \quad \mathrm{Var}(\overline{X}_i) = \frac{p_i(1-p_i)}{M}.$$

As $M \to \infty$, by applying Central Limit Theorem,

$$\frac{\overline{X}_i - p_i}{\sqrt{\frac{p_i(1-p_i)}{M}}} = \sqrt{M}\frac{\overline{X}_i - p_i}{\sqrt{p_i(1-p_i)}} \xrightarrow{d} N(0,1)$$

and since $\overline{Y}_{\Pi^{(N)}(i)} = \overline{X}_i$, then we can conclude that

$$\sqrt{M}\frac{\overline{Y}_{\Pi(i)} - p_i}{\sqrt{p_i(1-p_i)}} \xrightarrow{d} N(0,1).$$

Next, we define a random set $J^{(N)}$ as a set which includes indices i as follows

$$J^{(N)} = \{i : p_1 - \epsilon < p_i < p_1 + \epsilon\}$$

where

$$\epsilon = \frac{1}{N^{1-\frac{\alpha}{3}}}$$

Remember that α is given by $M = cN^{2-\alpha}$. Also note that $1 \in J^{(N)}$.

Let us first find the distribution of $|J^{(N)}|$ which is the number of elements in $J^{(N)}$. Note that for N large enough,

$$\Pr\Big(p_1 - \epsilon < p_i < p_1 + \epsilon\Big) = \int_{p_1-\epsilon}^{p_1+\epsilon} f_P(p)dp.$$

Since $\delta_1 < f_P(p) < \delta_2$, we conclude that

$$2\epsilon\delta_1 < \Pr\Big(p_1 - \epsilon < p_i < p_1 + \epsilon\Big) < 2\epsilon\delta_2,$$

So we can write

$$\Pr\Big(p_1 - \epsilon < p_i < p_1 + \epsilon\Big) = 2\epsilon\delta.$$

for some $\delta > 0$. We conclude that $|J^{(N)}| \sim Bin(N, 2\epsilon\delta)$ when N is large enough. In particular, for the expected value and variance of $|J^{(N)}|$ we get

$$E[|J^{(N)}|] = 2\epsilon\delta N = 2 \times \frac{1}{N^{1-\frac{\alpha}{3}}}\delta N = 2\delta N^{\frac{\alpha}{3}}$$
$$Var(|J^{(N)}|) = 2N\epsilon\delta(1-2\epsilon\delta)$$

and as $N \to \infty$, $Var(|J^{(N)}|) \sim 2\delta N^{\frac{\alpha}{3}}$.

Using Chebyshev's inequality,

$$P\left\{\Big||J^{(N)}| - E[|J^{(N)}|]\Big| > \delta N^{\frac{\alpha}{3}}\right\} < \frac{Var(|J^{(N)}|)}{\delta^2 N^{2\frac{\alpha}{3}}}$$
$$\frac{Var(|J^{(N)}|)}{\delta^2 N^{2\frac{\alpha}{3}}} = \frac{2\delta N^{\frac{\alpha}{3}}}{\delta^2 N^{2\frac{\alpha}{3}}} \to 0, \ as \ N \to \infty$$

Thus, $|J^{(N)}| > \delta N^{\frac{\alpha}{3}}$ with high probability. In particular,

$$|J^{(N)}| \to \infty, \ as \ N \to \infty$$

Lemma 3. For all $i \in |J^{(N)}|$, the distribution of normalized random variable $\overline{X}_i$ converges to normal distribution,

$$\sqrt{M}\frac{\overline{X}_i - p_1}{\sqrt{p_1(1-p_1)}} \xrightarrow{d} N(0,1)$$

and since we have $\overline{Y}_{\Pi^{(N)}(i)} = \overline{X}_i$,

$$\sqrt{M}\frac{\overline{Y}_{\Pi(i)} - p_1}{\sqrt{p_1(1-p_1)}} \xrightarrow{d} N(0,1).$$

Proof. Note that, $|p_i - p_1| < \epsilon = \frac{1}{N^{1-\frac{\alpha}{3}}}$ and

$$p_i \to p_1, \ as \ N \to \infty.$$

By knowing that $\frac{\sqrt{p_i(1-p_i)}}{\sqrt{p_1(1-p_1)}} \to 1$,

$$\sqrt{M}\frac{\overline{X}_i - p_1}{\sqrt{p_1(1-p_1)}} = \sqrt{M}\frac{\overline{X}_i - p_i + p_i - p_1}{\sqrt{p_i(1-p_i)}}\frac{\sqrt{p_i(1-p_i)}}{\sqrt{p_1(1-p_1)}}$$
$$= \sqrt{M}\frac{\overline{X}_i - p_i}{\sqrt{p_i(1-p_i)}} + \sqrt{M}\frac{p_i - p_1}{\sqrt{p_i(1-p_i)}}$$

which we already know that $\frac{\overline{X}_i - p_i}{\sqrt{p_i(1-p_i)}} \xrightarrow{d} N(0,1)$, so we can write

$$\left|\sqrt{M}\frac{p_i - p_1}{\sqrt{p_i(1-p_i)}}\right| \leq \frac{\sqrt{M} \times \epsilon}{\sqrt{p_1(1-p_1)}}$$

and as $N \to \infty$

$$\frac{\sqrt{M} \times \epsilon}{\sqrt{p_1(1-p_1)}} = \frac{\sqrt{N^{1-\frac{\alpha}{2}}}}{\sqrt{p_1(1-p_1)}} \times \frac{1}{N^{1-\frac{\alpha}{3}}} \to 0.$$

Same thing holds for $\overline{Y}_i$ since $\overline{Y}_{\Pi^{(N)}(i)} = \overline{X}_i$.

□

Next we consider the case where $\Pi^{(N)}(J^{(N)})$ which is

$$\Pi^{(N)}(J^{(N)}) = \{\Pi^{(N)}(i) : i \in J^{(N)}\}$$

is known to the adversary (not the individual $\Pi(i)$'s, but the whole set $\Pi^{(N)}(J^{(N)})$). In this case, for the adversary to find $\Pi^{(N)}(i)$, she needs to just look into the set $J^{(N)}$.
We show that even if the adversary knows the set $\Pi^{(N)}(J^{(N)})$, her mutual information goes to zero.

For simplicity, we assume that

$$J^{(N)} = \{1, 2, \cdots, n\},$$

where $n = |J^{(N)}| > \delta N^{\frac{\alpha}{3}}$ and let

$$\mathbf{Y}^{(n)} = (\mathbf{Y}_{\Pi(1)}, \mathbf{Y}_{\Pi(2)}, \cdots, \mathbf{Y}_{\Pi(n)}).$$

Now for the adversary to find $\Pi(1)$, she can look into set $J^{(N)}$ with size n rather than all the N users.

To finish the proof of Theorem 1, it suffices to show that as $N \to \infty$,

$$H(X_1(k)|\mathbf{Y}^{(N)}) \to H(X_1(k)).$$

To continue, we first prove two lemmas.

Lemma 4. Let $A_\lambda^{(N)}$ be as defined in Lemma 1, in particular, we have

$$\lim_{N\to\infty} P(\mathbf{Y}^{(N)} \in A_\lambda^{(N)}) = 1.$$

If for all $\mathbf{y}^{(N)} \in A_{\lambda}^{(N)}$, we have

$$H(X_1(k)|\mathbf{Y}^{(N)} = \mathbf{y}^{(N)}) \to H(X_1(k)) \quad \text{as} \quad N \to \infty.$$

Then, we have

$$\lim_{N\to\infty} H(X_1(k)|\mathbf{Y}^{(N)}) = H(X_1(k)).$$

Proof. We have

$$H(X_1(k)|\mathbf{Y}^{(N)}) = \sum_{\mathbf{y}^{(N)}} H(X_1(k)|\mathbf{Y}^{(N)} = \mathbf{y}^{(N)})P(\mathbf{Y}^{(N)} = \mathbf{y}^{(N)})$$

$$= \sum_{\mathbf{y}^{(N)}\in A_{\lambda}^{(N)}} H(X_1(k)|\mathbf{Y}^{(N)} = \mathbf{y}^{(N)})P(\mathbf{Y}^{(N)} = \mathbf{y}^{(N)})$$

$$+ \sum_{\mathbf{y}^{(N)}\notin A_{\lambda}^{(N)}} H(X_1(k)|\mathbf{Y}^{(N)} = \mathbf{y}^{(N)})P(\mathbf{Y}^{(N)} = \mathbf{y}^{(N)}).$$

Now note that the second sum converges to zero since $H(X_1(k)|\mathbf{Y}^{(N)} = \mathbf{y}^{(N)}) \leq H(X_1(k)) \leq 1$ and

$$\lim_{N\to\infty} P(\mathbf{Y}^{(N)} \notin A_{\lambda}^{(N)}) = 0.$$

On the other hand, the first sum converges to $H(X_1(k))$ by the assumptions of the lemma.

□

Lemma 4 allows us to consider only the observations $\mathbf{Y}^{(N)} = \mathbf{y}^{(N)}$ for $\mathbf{Y}^{(N)} \in A_{\lambda}^{(N)}$.

Lemma 5. Assume $P = (p_1, p_2, \cdots, p_N)$ is observed. We define q_N as the probability that the first user is in state (location) 1 at time k, i.e., $X_1(k) = 1$, given the observation vector $\mathbf{Y}^{(N)}$ and the set $\Pi(J^{(N)})$.

$$q_N = P(X_1(k) = 1|\mathbf{Y}^{(N)}, \Pi(J^{(N)}))$$

Since $\mathbf{Y}^{(N)}$ and $\Pi(J^{(N)})$ are random, q_N is a random variable. We have

$$q_N \xrightarrow{d} p_1.$$

Proof. This lemma is the result of the previous lemma. We have

$$q_N = P(X_1(k) = 1|\mathbf{Y}^{(N)} = \mathbf{y}^{(N)}, \Pi(J^{(N)})).$$

First, note that given set $\Pi(J^{(N)})$, we can ignore $\mathbf{Y}_i^{(N)}$ for $i \notin \Pi(J^{(N)})$, so we simply replace N with n to show this. By applying the Law of Total Probability we get

$$\sum_{j\in\Pi(J^{(N)})} P(X_1(k) = 1|\Pi(1) = j, \mathbf{Y}^{(n)} = \mathbf{y}^{(n)}, \Pi(J^{(N)}))$$

$$\times P(\Pi(1) = j|\mathbf{Y}^{(n)} = \mathbf{y}^{(n)}, \Pi(J^{(N)}))$$

$$= \sum_{j=1}^{n} 1_{[y_j^{(n)}(k)=1]} \times P(\Pi(1) = j|\mathbf{Y}^{(n)} = \mathbf{y}^{(n)}).$$

With the same reasoning as Lemma 4, it only suffices to consider $\mathbf{Y}^{(n)} = \mathbf{y}^{(n)}$ for $\mathbf{y}^{(n)} \in A_{\lambda}^{(N)}$. Also, by Lemma 2 if we define $\overline{Y}$ as,

$$\overline{Y}_i = \frac{1}{M}\sum_{k=1}^{M} Y_i(k).$$

then given $\overline{Y}_1, \overline{Y}_2, \cdots, \overline{Y}_n$, we have $\Pi^{(N)}$ and $\mathbf{Y}_n$ are independent. Thus,

$$P(\Pi(1) = j|\mathbf{Y}^{(n)} = \mathbf{y}^{(n)}) = P(\Pi(1) = j|\overline{\mathbf{Y}}^{(n)} = \overline{\mathbf{y}}^{(n)}).$$

But by Lemma 3,

$$\sqrt{M}\frac{\overline{Y}_i - p_1}{\sqrt{p_1(1-p_1)}} \xrightarrow{d} N(0,1).$$

This along with Lemma 1 tells us that with high probability

$$\frac{1-\lambda}{n} \leq P(\Pi(1) = j|\mathbf{Y}^{(n)} = \mathbf{y}^{(n)}) \leq \frac{1+\lambda}{n}.$$

We obtain

$$q_N \to \sum_j 1_{[y_j^{(N)}(k)=1]} \times P(\Pi(1) = j|\mathbf{Y}^{(n)} = \mathbf{y}^{(n)})$$

so we get

$$\frac{1-\lambda}{n}\sum_{j\in A_{\lambda}^{(n)}} 1_{[y_j^{(N)}(k)=1]} \leq q$$

$$\leq \frac{1+\lambda}{n}\sum_{j\in A_{\lambda}^{(n)}} 1_{[y_j^{(N)}(k)=1]}$$

Now, note that $Y_j(K)$ are $Bernoulli(p_{\Pi(j)})$ and since we are summing over n, by Law of Large Number and $p_{\Pi(j)} \to p_1$ then we have

$$\sum_j 1_{[Y_j^{(N)}(k)=1]} \to p_1, \quad as\ N \to \infty$$

and considering that we can write

$$(1-\lambda_1)p_1 \leq q_N \leq (1+\lambda_1)p_1$$

where λ_1 can be made arbitrarily small so that $q_N \xrightarrow{d} p_1$.

□

Thus,

$$(X_1(k) = 1|\mathbf{Y}^{(N)} = \mathbf{y}^{(N)}, \Pi(J^{(N)})) \to Bernoulli(p_1).$$

We already know that $X_1(k) \sim Bernoulli(p_1)$. It means that knowing $\mathbf{Y}^{(N)} = \mathbf{y}^{(N)}, \Pi(J^{(N)})$ does not change the distribution. In other words the entropy of $(X_1(k)|\mathbf{Y}^{(N)} = \mathbf{y}^{(N)}, \Pi(J^{(N)}))$ converges to $H(X_1(k))$.

$$H((X_1(k)|\mathbf{Y}^{(N)}, \Pi(J^{(N)}))) \to H(X_1(k))$$

and we know that conditioning does not increase entropy,

$$H((X_1(k)|\mathbf{Y}^{(N)}, \Pi(J^{(N)}))) \leq H(X_1(k))$$

$$H((X_1(k)|\mathbf{Y}^{(N)}, \Pi(J^{(N)}))) \leq H((X_1(k)|\mathbf{Y}^{(N)})$$

so,

$$H((X_1(k)|\mathbf{Y}^{(N)}) \leq H(X_1(k))$$

and since $H((X_1(k)|\mathbf{Y}^{(N)}, \Pi(J^{(N)}))) \to H(X_1(k))$

$$H(X_1(k)|\mathbf{Y}^{(N)}) \to H(X_1(k))$$

and finally we can write

$$I((X_1(k); \mathbf{Y}^{(N)}) \to 0, \ as\ N \to \infty$$

which completes the proof of Theorem 1.

C. Extension to r States (locations)

Here we extend the results to a scenario in which we have $r \geq 2$ locations or regions, locations $0, 1, \cdots, r-1$. At any time $k \in \{0, 1, 2, \cdots\}$, user i has probability $p_i^j \in (0, 1)$ to be at location j, independently from previous locations and independently from other users' locations.

We assume that p_i^j's (for $j = 0, 1, \cdots, r-2$) are drawn independently from some $r-1$ dimensional continuous density $f_P(p)$ on the $(0, 1)^{r-1}$. Specifically, $f_P(p) = 0$ for all $p \notin (0, 1)^{r-1}$ and there are $\delta_2 > \delta_1 > 0$ such that $\delta_1 < f_P(p) < \delta_2$ for all $p \in \{(p_0, p_1, \cdots, p_{r-2}) \in (0, 1)^{r-1} : p_0 + p_1 + \cdots + p_{r-2} \leq 1\}$. The values of p_i^j's are fixed and do not change as time goes on. We then can state the following theorem.

Theorem 2. For r locations with above definition and observation vector $\mathbf{Y}^{(M)}$ if all the following holds,

1) $M = cN^{\frac{2}{r-1}-\alpha}$, which $c, \alpha > 0$ and are constant
2) $p_i \in (0, 1)$
3) $(p_2, p_3, \cdots, p_N) \sim f_p$, $0 < \delta_1 < f_p < \delta_2$
4) $P = (p_1, p_2, \cdots, p_N)$ be known to the adversary

then, we have

$$\forall k \in \mathbb{N}, \quad \lim_{N \to \infty} I(X_1(k); \mathbf{Y}^{(M)}) = 0$$

Theorem 2 can be proved using similar ideas introduced in the proof of Theorem 1. We omit the proof due to space limitation.

IV. Conclusion

In this paper, we defined perfect location privacy based on the mutual information between the adversary's observation information and the actual user's location data. Then we simplified the problem into two-state locations with N number of users and M number of adversary's observations. We derived the relation between M and N. We showed that perfect location privacy is achievable if $M \leq cN^{2-\alpha}$. We then extended our model to r-state locations and obtained $M \leq cN^{\frac{2}{r-1}-\alpha}$ to have perfect location privacy.

References

[1] R. Shokri, G. Theodorakopoulos, C. Troncoso, J.-P. Hubaux, and J.-Y. Le Boudec, "Protecting location privacy: optimal strategy against localization attacks," in *Proceedings of the 2012 ACM conference on Computer and communications security*. ACM, 2012, pp. 617–627.

[2] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *Proceedings of the 1st international conference on Mobile systems, applications and services*. ACM, 2003, pp. 31–42.

[3] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Preserving privacy in gps traces via uncertainty-aware path cloaking," in *Proceedings of the 14th ACM conference on Computer and communications security*. ACM, 2007, pp. 161–171.

[4] N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Optimal geo-indistinguishable mechanisms for location privacy," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2014, pp. 251–262.

[5] J. Freudiger, M. Raya, M. Félegyházi, P. Papadimitratos, and J.-P. Hubaux, "Mix-zones for location privacy in vehicular networks," 2007.

[6] Z. Ma, F. Kargl, and M. Weber, "A location privacy metric for v2x communication systems," in *Sarnoff Symposium, 2009. SARNOFF'09. IEEE*. IEEE, 2009, pp. 1–6.

[7] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, "Quantifying location privacy," in *Security and Privacy (SP), 2011 IEEE Symposium on*. IEEE, 2011, pp. 247–262.

[8] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

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

[10] J. Freudiger, R. Shokri, and J.-P. Hubaux, "On the optimal placement of mix zones," in *Privacy enhancing technologies*. Springer, 2009, pp. 216–234.

[11] M. H. Manshaei, Q. Zhu, T. Alpcan, T. Bacşar, and J.-P. Hubaux, "Game theory meets network security and privacy," *ACM Computing Surveys (CSUR)*, vol. 45, no. 3, p. 25, 2013.

[12] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-L. Tan, "Private queries in location based services: anonymizers are not necessary," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 2008, pp. 121–132.

[13] M. Wernke, P. Skvortsov, F. Dürr, and K. Rothermel, "A classification of location privacy attacks and approaches," *Personal and Ubiquitous Computing*, vol. 18, no. 1, pp. 163–175, 2014.

[14] Y. Cai and G. Xu, "Cloaking with footprints to provide location privacy protection in location-based services," Jan. 1 2015, uS Patent App. 14/472,462. [Online]. Available: https://www.google.com/patents/US20150007341

[15] H. Kido, Y. Yanagisawa, and T. Satoh, "An anonymous communication technique using dummies for location-based services," in *Pervasive Services, 2005. ICPS'05. Proceedings. International Conference on*. IEEE, 2005, pp. 88–97.

[16] H. Lu, C. S. Jensen, and M. L. Yiu, "Pad: privacy-area aware, dummy-based location privacy in mobile services," in *Proceedings of the Seventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*. ACM, 2008, pp. 16–23.

[17] J. Lee and C. Clifton, "Differential identifiability," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1041–1049.

[18] K. Chatzikokolakis, C. Palamidessi, and M. Stronati, "Geo-indistinguishability: A principled approach to location privacy," in *Distributed Computing and Internet Technology*. Springer, 2015, pp. 49–72.

[19] H. H. Nguyen, J. Kim, and Y. Kim, "Differential privacy in practice," *Journal of Computing Science and Engineering*, vol. 7, no. 3, pp. 177–186, 2013.

[20] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory meets practice on the map," in *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*. IEEE, 2008, pp. 277–286.

[21] S.-S. Ho and S. Ruan, "Differential privacy for location pattern mining," in *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*. ACM, 2011, pp. 17–24.

[22] R. Dewri, "Local differential perturbations: Location privacy under approximate knowledge attackers," *Mobile Computing, IEEE Transactions on*, vol. 12, no. 12, pp. 2360–2372, 2013.

[23] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi, "Broadening the scope of differential privacy using metrics." in *Privacy Enhancing Technologies*. Springer, 2013, pp. 82–102.

[24] R. Shokri, "Optimal user-centric data obfuscation," *arXiv preprint arXiv:1402.3426*, 2014.

[25] K. Chatzikokolakis, C. Palamidessi, and M. Stronati, "Location privacy via geo-indistinguishability," *ACM SIGLOG News*, vol. 2, no. 3, pp. 46–69, 2015.

[26] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. ACM, 2013, pp. 901–914.

[27] R. Shokri, G. Theodorakopoulos, G. Danezis, J.-P. Hubaux, and J.-Y. Le Boudec, "Quantifying location privacy: the case of sporadic location exposure," in *Privacy Enhancing Technologies*. Springer, 2011, pp. 57–76.