Achieving Perfect Location Privacy in Wireless Devices Using Anonymization

Zarrin Montazeri *Student Member, IEEE,* Amir Houmansadr *Member, IEEE,* Hossein Pishro-Nik *Member, IEEE,*

***Abstract*—The popularity of mobile devices and location-based services (LBS) has raised significant concerns regarding the location privacy of their users. A popular approach to protect location privacy is anonymizing the users of LBS systems. In this paper, we introduce an information-theoretic notion for location privacy, which we call perfect location privacy. We then demonstrate how anonymization should be used by LBS systems to achieve the defined perfect location privacy.**

We study perfect location privacy under two models for user movements. First, we assume that a user's current location is independent from her past locations. Using this i.i.d. model, we show that if the pseudonym of the user is changed before $O(n^{\frac{2}{r-1}})$ observations are made by the adversary for that user, then the user has perfect location privacy. Here, n is the number of the users in the network and r is the number of all possible locations. Next, we model users' movements using Markov chains to better model real-world movement patterns. We show that perfect location privacy is achievable for a user if the user's pseudonym is changed before $O(n^{\frac{2}{|E|-r}})$ observations are collected by the adversary for that user, where $|E|$ is the number of edges in the user's Markov chain model.

***Index Terms*—Location Privacy, Mobile Networks, Information Theoretic Privacy, Anonymization, Location Privacy Protecting Mechanism (LPPM), Markov Chains.**

I. Introduction

MOBILE devices offer a wide range of services by collecting and processing the geographic locations of their users. Such services, broadly known as *location-based services (LBS)*, include navigation, ride-sharing, dining recommendation, and auto-collision warning applications. While such LBS applications offer a wide range of popular and important services to their users, they impose significant privacy threats because of their unrestricted access to the location information of these wireless devices over time. Privacy attacks can also be launched by other types of adversaries including third-party applications, nearby mobile users, and cellular service providers.

To protect the location privacy of LBS users, various mechanisms have been designed [3]–[5], which are known as *Location-Privacy Protection Mechanisms (LPPM)*. These mechanisms work by perturbing the privacy-sensitive information collected by wireless devices, such as users' identifiers or location coordinates, before revealing such information to an LBS application. LPPMs are mainly classified into two types; those that perturb users' identity information are known as *identity perturbation mechanisms*, and those that perturb the location information of the users are known as *location perturbation mechanisms*. Such LPPMs improve privacy usually at the cost of degrading the utility of the underlying LBS applications, e.g., perturbing the location information sent to a restaurant recommendation LBS improves privacy while degrading the accuracy of the recommendations. Therefore, LPPM mechanisms need to be designed to make the right tradeoff between privacy and utility.

In this paper, we develop an information-theoretic framework to assess location privacy for anonymization-based LPPMs. We formulate a user's location privacy based on the mutual information between the adversary's anonymized observations and the user's actual location information. We define the notion of *perfect location privacy* and show that it is possible to design anonymization-based LPPMs that achieve the defined perfect location privacy. We start our analysis for a simple scenario where users' mobility is governed by i.i.d. models, i.e., each user's current location is independent of her past locations. We show that if the pseudonym of the user is changed before $O(n^{\frac{2}{r-1}})$ observations are made by the adversary for that user, then the user has perfect location privacy. Here, n is the number of the users in the network and r is the number of all possible locations. We then extend the mobility model to the more realistic setting where user movements follow Markov chains. We assume the strongest model for the adversary by assuming that the adversary has complete statistical knowledge of the users' movements, i.e., she knows each user's Markov model. We show that perfect location privacy is achievable for a user if the user's pseudonym is changed before $O(n^{\frac{2}{|E|-r}})$ observations are collected by the adversary for that user, where $|E|$ is the number of edges in the user's Markov chain model. Our work in this paper significantly extends our previous study [1], [2] by offering new results, analysis, and proofs.

Note that our work derives at the theoretical bounds of location privacy for anonymization-based LPPMs under, e.g., when the number of users goes to infinity. Applying our results to practical settings need be done considering the specific threat models, e.g., the number of mobile entities, the capabilities of the adversary (the number and location of observation points, prior knowledge about mobile entities,

Z. Montazeri is with the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA, 01003 USA e-mail: (seyedehzarin@umass.edu).

A. Houmansadr is with the College of Information and Computer Sciences, University of Massachusetts, Amherst, MA, 01003 USA e-mail:(amir@cs.umass.edu)

H. Pishro-Nik is with the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA, 01003 USA e-mail:(pishro@engin.umass.edu)

This work was supported by National Science Foundation under grants CCF 0844725, CCF 1421957, and CNS 1525642. Parts of this work was presented in Annual Conference on Information Science and Systems (CISS 2016) [1], and in International Symposium on Information Theory and Its Applications (ISITA 2016) [2].

etc.), the extent of possible geographic locations, etc.

II. Related Work

Location privacy has been an active field of research over the past decade [3], [6]–[14]. Studies in this field can be classified into two main classes: 1) Designing effective LPPMs for specific LBS systems and platforms, 2) Deriving theoretical models for location privacy, e.g., deriving metrics to quantify the location privacy.

The designed LPPMs can be classified into two classes: 1) Location perturbation LPPMs, 2) Identity perturbation LPPMs. Location perturbation LPPMs aim at obfuscating the location information of the users over time and geographical domain with methods such as cloaking, [6], [9], and adding dummy locations, [10], [11]. On the other hand, identity perturbation LPPMs try to obfuscate the user's identity while using an LBS. Some common approaches to perturb the identity of the user is to either exchange users' identifiers, [15], or assign random pseudonyms to them, known as anonymization technique, [16], [17]. The former method usually uses some pre-defined regions, called *mixed-zones*, to exchange users' identifiers within those regions. As users cross such regions, they exchange their identifiers with other users in the same region using an encryption protocol to confuse the adversary, [18], [19].

Previous studies have shown that using anonymization alone is not enough to protect users' location privacy in real-world scenarios where users go to unique locations. Particularly, Zang et al. demonstrate that an adversary has a significant advantage in identifying users who visit unique locations [20]. Also, Golle and Partridge show that the possibility of user identification based on anonymized location traces is significantly increased when the individual's home and work locations are known [21]. Please note that this does not contradict the analysis and findings of our paper as we use a different setting. First, our analysis seeks to find the theoretical limits of privacy for situations where the number of users (N) goes to infinity, which is not the case in previous studies like [20], [21]. Increasing the number of user reduces an adversary's confidence in distinguishing different users. Second, in our analysis we assume "continuous" density functions for the movements of the users across different locations (e.g., the $f_P(p)$ function in Section IV-A). Therefore, user distributions do not contain Dirac delta functions representing their unique locations. Note that this is not an unrealistic assumption; in real-world scenarios with users having unique locations, we assume that the location information is pre-processed to satisfy this continuity requirement. Such pre-processing can be performed in two ways; first, by reducing the granularity of locations, e.g., in our analysis we divide a region of interest into a number of girds (i.e., into r coarse-grained locations). Second, an obfuscation mechanism can be applied to location traces to ensure they satisfy the continuity requirement. Further discussion on implementing such pre-processing is out of the scope of our work and we leave it to future work.

A related, but in parallel, approach to our study is differential privacy-based mechanisms. Differential privacy is mainly studied in the context of databases containing sensitive information, where the goal of differential privacy is to respond queries on the aggregation of the information in the database without revealing sensitive information about the individual entries in the database. Differential privacy has been extensively studied in the context of location privacy, i.e., to prevent data leakage from location information databases [7], [22]–[26]. The goal here is to insure that the presence of no single user could noticeably change the outcome of the aggregated location information. For instance, Ho et al. [27] proposed a differentially private location pattern mining algorithm using quadtree spatial decomposition. Some location perturbation LPPMs are based on ideas from differential privacy [23], [28]–[31]. For instance, Dewri [32] suggest to design obfuscation LPPMs by applying differential perturbations. Alternatively, 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 [31]. Note that our approach is entirely in parallel with this line of work. Our paper tries to achieve the theoretical limits on location privacy—independent of the LPPM mechanisms being used—while differential privacy based studies on location privacy try to *design* specific LPPM mechanisms under very specific application scenarios.

Several works aim at quantifying the location privacy of mobile users. A common approach is called K-anonymity, [4], [16]. In K-anonymity, each user's identity is kept indistinguishable within a group of $k-1$ other users. On the other hand, Shokri et al. [12], [13] define the expected estimation error of the adversary as a metric to evaluate LPPMs. Ma et al. [33] use the uncertainty of the users' location information to quantify the location privacy of the user in vehicular networks. Li et al. [34] define metrics to show the tradeoff between the privacy and utility of LPPMs.

Wang et al. tried to protect the privacy of the users for context sensing on smartphones, using Markov decision process (MDP) [35]. The adversary's approach and user's privacy preserving mechanism is changing during time. The goal is to obtain the optimal policy of the users.

Previously, the mutual information has been used as a privacy metric in different topics, [36]–[41]. However, in this paper we use the mutual information specifically for location privacy. For this reason, a new setting for this privacy problem will be provided and discussed. Specifically, we provide an information theoretic definition for location privacy using the mutual information. We show that wireless devices can achieve provable perfect location privacy by using the anonymization method in the suggested way.

In [42], the author studies asymptotically optimal matching of sequences to source distributions. However, there are two key differences between [42] and this paper. First, [42] looks only at the optimal matching tests, but does not consider any privacy metric (i.e., perfect privacy) as considered in this paper. The major part of our work is to show that the mutual information converges to zero so we can conclude there is no privacy leak (hence perfect privacy). Also, the setting of [42] is different as it assumes a fixed distribution on sources (i.e., classical inference) as we assume the existence of a general (but possibly unknown) prior distributions for the sources (i.e. a Bayesian setting).

III. Framework

A. Defining Perfect Location Privacy

In the proposed framework, we consider a region in which a large number of wireless devices are using an LBS. To support their location privacy, the anonymization technique is being used by the LBS. An outsider adversary $\mathcal{A}$ is interested in identifying users based on their locations and movements. We consider this adversary to be the strongest adversary that has complete statistical knowledge of the users' movements based on the previous observations or other resources. The adversary has a model that describes users' movements as a random process on the corresponding geographic area.

Let $X_u(t)$ be the location of user u at time t, and n be the number of users in our network. The location data of users can be represented in the form of the following stochastic processes:

$$\begin{array}{ccccccc} X_1(1) & X_1(2) & X_1(3) & \cdots & X_1(m) & X_1(m+1) & \cdots \\ X_2(1) & X_2(2) & X_2(3) & \cdots & X_2(m) & X_2(m+1) & \cdots \\ X_3(1) & X_3(2) & X_3(3) & \cdots & X_3(m) & X_3(m+1) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_n(1) & X_n(2) & X_n(3) & \cdots & X_n(m) & X_n(m+1) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{array}$$

The adversary's observations are anonymized versions of the $X_u(t)$'s produced by the anonymization technique. She is interested in knowing $X_u(t)$ for $u = 1, 2, ..., n$ based on her m anonymized observations for each of the n users, where m is a function of n, e.g., $m = m(n)$. Thus, at time m, the data shown in the box has been produced:

$$\begin{array}{|ccccc|cc} X_1(1) & X_1(2) & X_1(3) & \cdots & X_1(m) & X_1(m+1) & \cdots \\ X_2(1) & X_2(2) & X_2(3) & \cdots & X_2(m) & X_2(m+1) & \cdots \\ X_3(1) & X_3(2) & X_3(3) & \cdots & X_3(m) & X_3(m+1) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_n(1) & X_n(2) & X_n(3) & \cdots & X_n(m) & X_n(m+1) & \cdots \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{array}$$

The goal of this paper is to find the function $m = m(n)$ in a way that perfect privacy is guaranteed. Let $\mathbf{Y}^{(m)}$ be a collection of anonymized observations available to the adversary. That is, $\mathbf{Y}^{(m)}$ is the anonymized version of the data in the box. We define *perfect location privacy* as follows:

Definition 1. User u has perfect location privacy at time t, if and only if

$$\lim_{n\to\infty} I\left(X_u(t); \mathbf{Y}^{(m)}\right) = 0,$$

where $I(X;Y)$ shows the mutual information between X and Y.

The above definition implies that over time, the adversary's anonymized observations do not give any information about the user's location. The assumption of large n, $(n \to \infty)$, is valid for almost all applications that we consider since the numbers of users for such applications are usually very large.

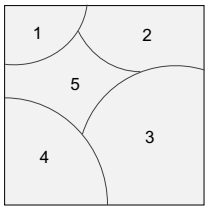

Fig. 1: An area is divided into five regions that users can occupy.

In order to achieve perfect location privacy, we only consider anonymization techniques to confuse the adversary. In particular, the anonymization can be modeled as follows:

We perform a random permutation $\Pi^{(n)}$, chosen uniformly at random among all $n!$ possible permutations on the set of n users, and then assign the pseudonym $\Pi^{(n)}(u)$ to user u

$$\Pi^{(n)} : \{1, 2, \cdots, n\} \mapsto \{1, 2, \cdots, n\}.$$

Throughout the paper, we may use $\Pi(u)$ instead of $\Pi^{(n)}(u)$ for simplicity of the notation.

For $u = 1, 2, \cdots, n$, let $\mathbf{X}_u^{(m)}$ be a vector which shows the u^{th} user's locations at times $t = 1, 2, \ldots, m$:

$$\mathbf{X}_u^{(m)} = (X_u(1), X_u(2), \cdots, X_u(m))^T$$

Using the permutation function $\Pi^{(n)}$, the adversary observes a permutation of users' location vectors, $\mathbf{X}_u^{(m)}$'s. In other words, the adversary observes

$$\begin{aligned} \mathbf{Y}^{(m)} &= \text{Perm}\left(\mathbf{X}_1^{(m)}, \mathbf{X}_2^{(m)}, \cdots, \mathbf{X}_n^{(m)}; \Pi^{(n)}\right) \\ &= \left(\mathbf{X}_{\Pi^{-1}(1)}^{(m)}, \mathbf{X}_{\Pi^{-1}(2)}^{(m)}, \cdots, \mathbf{X}_{\Pi^{-1}(n)}^{(m)}\right) \\ &= \left(\mathbf{Y}_1^{(m)}, \mathbf{Y}_2^{(m)}, \cdots, \mathbf{Y}_n^{(m)}\right) \\ \mathbf{Y}_u^{(m)} &= \mathbf{X}_{\Pi^{-1}(u)}^{(m)}, \ \mathbf{Y}_{\Pi(u)}^{(m)} = \mathbf{X}_u^{(m)} \end{aligned} \tag{1}$$

where Perm(.) shows the applied permutation function. Then,

$$\mathbf{Y}_{\Pi(u)}^{(m)} = \mathbf{X}_u^{(m)} = (X_u(1), X_u(2), \cdots, X_u(m))^T .$$

B. Example

Here we provide a simple example to further elaborate the problem setting. Assume that we have only three users, $n = 3$, and five locations, $r = 5$, that users can occupy (Figure 1). Also, let's assume that the adversary can collect $m(n) = 4$ observations per user. Each user creates a path as below:

user	path
user 1	$1 \to 2 \to 3 \to 4$
user 2	$2 \to 1 \to 3 \to 5$
user 3	$4 \to 5 \to 1 \to 3$

$$\mathbf{X}_1^{(4)} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, \quad \mathbf{X}_2^{(4)} = \begin{bmatrix} 2 \\ 1 \\ 3 \\ 5 \end{bmatrix}, \quad \mathbf{X}_3^{(4)} = \begin{bmatrix} 4 \\ 5 \\ 1 \\ 3 \end{bmatrix}, \ \mathbf{X}^{(4)} = \begin{bmatrix} 1 & 2 & 4 \\ 2 & 1 & 5 \\ 3 & 3 & 1 \\ 4 & 5 & 3 \end{bmatrix}$$

To anonymize the users, we will assign a pseudonym to each. The pseudonyms are determined by the function defined by a random permutation on the user set:

$$\Pi^{(3)} : \{1, 2, 3\} \mapsto \{1, 2, 3\}$$

For this example, suppose that the permutation function is given by $\Pi(1) = 3$, $\Pi(2) = 1$, and $\Pi(3) = 2$. The choice of the permutation is the only piece of information that is not available to the adversary. So here, the adversary observes anonymized users and their paths:

pseudonym	observation
user 1	$2 \to 1 \to 3 \to 5$
user 2	$4 \to 5 \to 1 \to 3$
user 3	$1 \to 2 \to 3 \to 4$

$$\mathbf{Y}^{(4)} = \begin{bmatrix} 2 & 4 & 1 \\ 1 & 5 & 2 \\ 3 & 1 & 3 \\ 5 & 3 & 4 \end{bmatrix}$$

and she wants to find which user (with the pseudonym user 3) actually made $1 \to 2 \to 3 \to 4$, and so on for the other users. Based on the number of observations that the adversary collects for each user, $m = m(n) = 4$, and also the statistical knowledge of the users' movements, she aims at breaking the anonymization function and de-anonymizing the users. The accuracy of this method depends on the number of observations that the adversary collects, and thus our main goal in this paper is to find the function $m(n)$ in a way that the adversary is unsuccessful and the users have perfect location privacy.

IV. I.I.D. Model

A. Perfect Location Privacy for a Simple Two-State Model

To get a better insight about the location privacy problem, here we consider a simple scenario where there are only two states to which users can go, states 0 and 1. At any time $k \in \{0, 1, 2, \cdots\}$, user u has probability $p_u \in (0, 1)$ to be at state 1, independent from her previous locations and other users' locations. Therefore,

$$X_u(k) \sim \text{Bern}(p_u).$$

To keep things general, we assume that p_u's are drawn independently from some continuous density function, $f_P(p)$, on the $(0, 1)$ interval. Specifically, there are $\delta_2 > \delta_1 > 0$ such that[1]

$$\begin{cases} \delta_1 < f_P(p) < \delta_2 & p \in (0, 1) \\ f_P(p) = 0 & p \notin (0, 1) \end{cases}$$

The values of p_u's are known to the adversary. Thus, the adversary can use this knowledge to potentially identify the users. Note that our results do not depend on the choice of $f_P(p)$ and we do not assume that we know the underlying distribution $f_P(p)$. All we assume here is the existence of such distribution. The following theorem gives a general condition to guarantee perfect location privacy:

[1]The condition $\delta_1 < f_P(p) < \delta_2$ for all $p \in (0, 1)$ is not actually necessary for the results and can be relaxed in several ways. For example, the range of P could be any sub-intervals of $(0, 1)$. In fact, as long as the condition $\delta_1 < f_P(p) < \delta_2$ is valid on an open interval around p_1, the statement of Theorem 1 remains valid.

Theorem 1. For two locations with the above assumptions on $f_P(p)$, and anonymized observation vector of the adversary, $\mathbf{Y}^{(m)}$, if all the following holds

1) $m = cn^{2-\alpha}$, for some positive constants c and $\alpha < 1$;
2) $p_1 \in (0, 1)$;
3) $(p_2, p_3, \cdots, p_n) \sim f_P$;
4) $P = (p_1, p_2, \cdots, p_n)$ be known to the adversary;

then, we have

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

i.e., user 1 has perfect location privacy.

Note that although the theorem is stated for user 1, the symmetry of the problem allows it to be restated for all users. Also note that the theorem is proven for any $0 < \alpha < 1$. Therefore, roughly speaking, the theorem states that if the adversary obtains less than $O(n^2)$ observations per user, then all users have location privacy.

B. The Intuition Behind The Proof

Here we provide the intuition behind the proof, and the rigorous proof for Theorem 1 is given in Appendix A. Let us look from the adversary's perspective. The adversary is observing anonymized locations of the first user and she wants to figure out the index of the user that she is observing, in other words she wants to obtain $X_1(k)$ from $\mathbf{Y}^{(m)}$. Note that the adversary knows the values of $p_1, p_2, \cdots, p_n$. To obtain $X_1(k)$, it suffices that the adversary obtains $\Pi(1)$. This is because $\mathbf{X}_1^{(m)} = \mathbf{Y}_{\Pi(1)}^{(m)}$, so

$$\begin{aligned} X_1(k) &= \text{the kth element of } \mathbf{X}_1^{(m)} \\ &= \text{the kth element of } \mathbf{Y}_{\Pi(1)}^{(m)}. \end{aligned}$$

Since $X_u(k)$ is a Bernoulli random variable with parameter p_u, to do so, the adversary can look at the averages

$$\overline{Y}_{\Pi(u)} = \frac{Y_{\Pi(u)}(1) + Y_{\Pi(u)}(2) + \ldots + Y_{\Pi(u)}(m)}{m}.$$

In fact, $\overline{Y}_{\Pi(u)}$'s provide 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 $u \neq 1$, $\overline{Y}_{\Pi(u)}$ is not too close to p_1.

Now, note that by the Central Limit Theorem (CLT)

$$\frac{\overline{Y}_{\Pi(u)} - p_u}{\sqrt{\frac{p_u(1-p_u)}{m}}} \to N(0, 1).$$

That is, loosely speaking, we can write

$$\overline{Y}_{\Pi(u)} \to N\left(p_u, \frac{p_u(1-p_u)}{m}\right).$$

Consider an interval $I \subseteq (0, 1)$ such that p_1 falls into that interval and length of I, $l^{(n)}$, is chosen to be

$$l^{(n)} = \frac{1}{n^{1-\eta}},$$

where $0 < \eta < \frac{\alpha}{2}$ (Remember α was defined by the equation $m = cn^{2-\alpha}$ in the statement of Theorem 1) . Note that $l^{(n)}$

goes to zero as n becomes large. Also, note that for any $u \in 1, 2, \cdots, n$, the probability that p_u is in I is larger than $\delta_1 l^{(n)}$. In other words, since there are n users, we can guarantee that a large number of p_u's fall in I since we have

$$nl^{(n)} \to \infty.$$

On the other hand, note that

$$\frac{\sqrt{\mathrm{Var}(\overline{Y}_{\Pi(u)})}}{\mathrm{length}(I)} = \frac{\sqrt{\frac{p_u(1-p_u)}{m}}}{\frac{1}{n^{1-\eta}}} = n^{\frac{\alpha}{2}-\eta} \to \infty.$$

Note that here, we will have a large number of normal random variables $\overline{Y}_{\Pi(u)}$ whose expected values are in the interval I (that has a vanishing length) with high probability and their standard deviations are much larger than the interval length (but equal to each other asymptotically, i.e. $\frac{p_1(1-p_1)}{m}$). Thus, distinguishing between them will become impossible for the adversary. In other words, the probability that the adversary will correctly identify $\Pi(1)$ goes to zero as n goes to infinity. That is, the adversary will most likely choose an incorrect value j for $\Pi(1)$. In this case, since the locations of different users are independent, the adversary will not obtain any useful information by looking at $X_j(k)$. Of course, the above argument is only intuitive. The rigorous proof has to make sure all the limiting conditions work out appropriately. This has been accomplished in Appendix A.

Note that if we do not take to account the fact that that p_u's are drawn independently from some continuous density function, $f_P(p)$, on the (0,1) interval, then the Theorem 1 is not necessarily valid. Here we provide an example. Consider we have two states and assume that the adversary has m observations of n user locations, and assume that the location of user 1$\sim$ Bern($\frac{1}{2}$), and the location of all other users 2, ..., n are drawn from Bern(p_k), with $p_k = \frac{1}{4}$, $k \in [2, \cdots, n]$. Now, from Hoeffding's inequality, for $S_k \triangleq \sum_{k=1}^{m} \frac{X_k(k)}{m}$,

$$Pr\{|S_1 - \frac{1}{2}| \geq \frac{1}{8}\} \leq \exp(-c_1 m^2),$$

where c_1 is some absolute positive constant. On the other hand, for each of the other users

$$Pr\{|S_k - \frac{1}{4}| \geq \frac{1}{8}\} \leq \exp(-c_2 m^2),$$

and, from the union bound,

$$Pr\{\exists k : |S_k - p_k| \geq \frac{1}{8}\} \leq n \exp(-c_3 m^2)$$

where c_3 is an absolute positive constant. Thus, unless n grows as $O(\exp(m^2))$ (i.e. exponentially in m), we can perfectly detect the location of user 1 by simply considering the sequence with average closest to $\frac{1}{2}$.

C. Extension to r-States Model

Here, we extend our results to a scenario in which we have r locations, $0, 1, \cdots, r-1$, where $r \geq 2$ is a fixed integer (not a function of n). At any time $k \in \{0, 1, 2, \cdots\}$, user u has probability $p_u(i) \in (0, 1)$ to be at location i, independent from her previous locations and other users' locations. At any given time k, $p_u(i)$ shows the probability of user u being at location i and vector $\mathbf{p}_u$ contains $p_u(i)$'s for all the locations

$$p_u(i) = P(X_u(k) = i),$$

$$\mathbf{p}_u = (p_u(0), p_u(1), \cdots, p_u(r-1)).$$

We assume that $p_u(i)$'s for $i = 0, 1, 2, \cdots, r-2$ are drawn independently from some $r-1$ dimensional continuous density function $f_{\mathbf{P}}(\mathbf{p})$ on the $(0, 1)^{r-1}$. In particular, define the range of distribution as

$$R_{\mathbf{P}} = \{(x_1, x_2, \cdots, x_{r-1}) \in (0, 1)^{r-1} : x_i > 0, x_1 + x_2 + \cdots + x_{r-1} < 1\}.$$

Then, we assume there exists positive constants $\delta_1, \delta_2 > 0$ such that

$$\begin{cases} \delta_1 < f_{\mathbf{P}}(\mathbf{p}_u) < \delta_2 & \mathbf{p}_u \in R_{\mathbf{P}} \\ f_{\mathbf{P}}(\mathbf{p}_u) = 0 & \mathbf{p}_u \notin R_{\mathbf{P}} \end{cases}$$

For example, Figure 2 shows the range $R_{\mathbf{P}}$ for the case where there are three locations, $r = 3$.

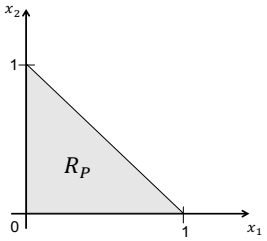

Fig. 2: R_P for case $r = 3$, $(d = 2)$.

Theorem 2. For r locations with the above definition and the adversary with an 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) $\mathbf{p}_1 \in (0, 1)^r$
3) $(\mathbf{p}_2, \mathbf{p}_3, \cdots, \mathbf{p}_n) \sim f_{\mathbf{P}}$, $0 < \delta_1 < f_{\mathbf{P}} < \delta_2$
4) $\mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n)$ is known to the adversary

then, we have

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

i.e., user 1 has perfect location privacy.

Discussion: Before discussing the proof of Theorem 2, it is worth noting that the number of locations r is assumed to be a fixed number. In particular, it is not a function of n. Intuitively, this implies that the number of users, n, is much larger that the number of locations. Also, the value of α is a constant as well. The statement of the Theorem is valid for any value of α, but the strongest result is obtained when α is chosen to be very small, and specifically much smaller than $\frac{2}{r-1}$. Thus, loosely speaking, the Theorem essentially states that if m is significantly smaller than $O\left(n^{\frac{2}{r-1}}\right)$, then users have perfect privacy.

Sketch of proof: Proof of the Theorem 2 is analogous to the proof of Theorem 1. Here, we provide the general intuition.

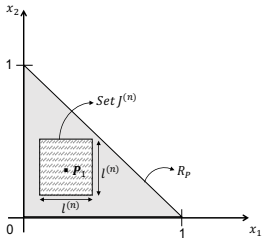

Fig. 3: $\mathbf{p}_1 = (p_1(0), p_1(1), \cdots, p_1(r-1))$ is in set $J^{(n)}$ in R_P.

We do not provide the entire rigorous proof as it is for the most part repetition of the arguments provided for Theorem 1 in Appendix A.

Let d be equal to $r-1$. As you can see in Figure 3, for three locations there exists a set $J^{(n)}$ such that $\mathbf{p}_1$ is in that set and we have

$$Vol(J^{(n)}) = (l^{(n)})^d.$$

We choose $l^{(n)} = \frac{1}{n^{\frac{1}{d}-\eta}}$, where $\eta < \frac{\alpha}{2}$. Thus, the average number of users with $\mathbf{p}$ vector in $J(n)$ is

$$n\frac{1}{\left(n^{\frac{1}{d}-\eta}\right)^d} = n^{d\eta} \to \infty \quad \text{as} \quad n \to \infty,$$

so we can guarantee that a large number of users are in the set $J^{(n)}$. This can be done exactly as in the proof of Theorem 1 using Chebyshev's Inequality.

Here, the number of times a user is at each location follows a multinomial distribution and in the long run, these numbers have a jointly Gaussian distribution asymptotically as n goes to infinity. The standard deviation of these variables are in the form of $\frac{const}{\sqrt{m}}$. Moreover, the standard deviation over length of this interval is also large

$$\frac{\frac{const.}{\sqrt{m}}}{l^{(n)}} = \frac{const.\ n^{\frac{1}{d}-\eta}}{\sqrt{m}} \sim \frac{n^{\frac{1}{d}-\eta}}{(n^{\frac{2}{d}-\alpha})^{\frac{1}{2}}} = n^{\frac{\alpha}{2}-\eta} \to \infty \quad \text{as} \quad n \to \infty.$$

Again, we have a large number of asymptotically jointly normal random variables that have a much larger standard deviation compared to the differences of their means. Thus, distinguishing between them becomes impossible.

This suggests that it is impossible for the adversary to correctly identify a user based on her observations even though she knows $\mathbf{P}$ and $\mathbf{Y}^{(m)}$. So, all the users have perfect location privacy. The proof can be made rigorous exactly the same way we did for the proof of Theorem 2, so we do not repeat the details here.

V. Markov Chain Model

Assume there are r possible locations to which users can go, where r is a fixed integer. We use a Markov chain with r states to model movements of each user. We define E, the set of edges in this Markov chain, such that (i, j) is in E if there exists an edge from i to j with probability $p(i, j) > 0$.

We assume that this Markov chain gives the movement pattern of each user and what differentiates between users is their transition probabilities. That is, for fixed locations i and j, two different users could have two different transition probabilities. For simplicity, let's assume that all users start at location (state) 1, i.e., $X_u(1) = 1$ for all $u = 1, 2, \cdots$. This condition is not necessary and can be easily relaxed; however, we assume it here for the clarity of exposition. We now state and prove the theorem that gives the condition for perfect location privacy for a user in the above setting.

Theorem 3. For an irreducible, aperiodic Markov chain with r states and $|E|$ edges, if $m = cn^{\frac{2}{|E|-r}-\alpha}$, where $c > 0$ and $\alpha > 0$ are constants, then

$$\lim_{n\to\infty} I(X_1(k); \mathbf{Y}^{(m)}) = 0, \quad \forall k \in \mathbb{N}, \tag{2}$$

i.e., user 1 has perfect location privacy.

Proof. Let $M_u(i, j)$ be the number of observed transitions from state i to state j for user u. We first show that $M_{\Pi(u)}(i, j)$'s provide a sufficient statistic for the adversary when the adversary's goal is to obtain the permutation $\Pi^{(n)}$. To make this statement precise, let's define $\mathbf{M}_u^{(m)}$ as the matrix containing $M_u(i, j)$'s for user u:

$$\mathbf{M}_u^{(m)} = \begin{bmatrix} M_u(1,1) & M_u(1,2) & \cdots & M_u(1,r) \\ M_u(2,1) & M_u(2,2) & \cdots & M_u(2,r) \\ \cdots & \cdots & \cdots & \cdots \\ M_u(r,1) & M_u(r,2) & \cdots & M_u(r,r) \end{bmatrix}$$

Also, let $\mathbf{M}^{(m)}$ be the ordered collection of $\mathbf{M}_u^{(m)}$'s. Specifically,

$$\mathbf{M}^{(m)} = \left(\mathbf{M}_1^{(m)}, \mathbf{M}_2^{(m)}, \cdots, \mathbf{M}_n^{(m)}\right)$$

The adversary can obtain $\text{Perm}(\mathbf{M}^{(m)}, \Pi^{(n)})$, a permuted version of $\mathbf{M}^{(m)}$. In particular, we can write

$$\begin{aligned}\text{Perm}(\mathbf{M}^{(m)}, \Pi^{(n)}) &= \text{Perm}\left(\mathbf{M}_1^{(m)}, \mathbf{M}_2^{(m)}, \cdots, \mathbf{M}_n^{(m)}; \Pi^{(n)}\right) \\ &= \left(\mathbf{M}_{\Pi^{-1}(1)}^{(m)}, \mathbf{M}_{\Pi^{-1}(2)}^{(m)}, \cdots, \mathbf{M}_{\Pi^{-1}(n)}^{(m)}\right).\end{aligned}$$

We now state a lemma that confirms $\text{Perm}(\mathbf{M}^{(m)}, \Pi^{(n)})$ is a sufficient statistic for the adversary, when the adversary's goal is to recover $\Pi^{(n)}$. Remember that $\mathbf{Y}^{(m)}$ is the collection of anonymized observations of users' locations available to the adversary.

Lemma 1. Given $\text{Perm}(\mathbf{M}^{(m)}, \Pi^{(n)})$, the random matrix $\mathbf{Y}^{(m)}$ and the random permutation $\Pi^{(n)}$ are conditionally independent. That is

$$\begin{aligned}P\left(\Pi^{(n)} = \pi \;\middle|\; \mathbf{Y}^{(m)} = \mathbf{y}, \text{Perm}(\mathbf{M}^{(m)}, \Pi^{(n)}) = \mathbf{m}\right) = \\ P\left(\Pi^{(n)} = \pi \;\middle|\; \text{Perm}(\mathbf{M}^{(m)}, \Pi^{(n)}) = \mathbf{s}\right)\end{aligned} \tag{3}$$

Lemma 1 is proved in the Appendix B.

Next note that since the Markov chain is irreducible and aperiodic, when we are determining $p(i, j)$'s, there are d degrees of freedom, where d is equal to $|E|-r$. This is because for each state i, we must have

$$\sum_{j=1}^{r} p(i, j) = 1.$$

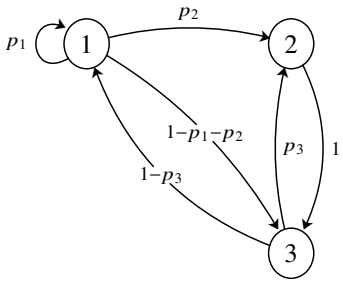

Fig. 4: Three states Markov chain example

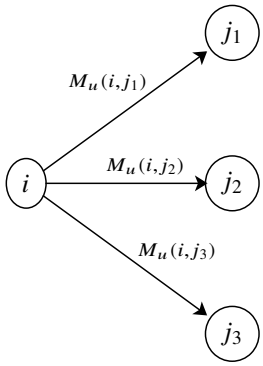

Fig. 5: If the total number of visits to State i is $M_i(u) = m_i(u)$, then the number of visits from state i to states j_k ($M_u(i,j_k)$) follow a multinomial distribution with probabilities $p_u(i,j_k)$.

Thus, the Markov chain of the user u is completely determined by d values of $p(i,j)$'s which we show as

$$\mathbf{P}_u = (p_u(1), p_u(2), \cdots, p_u(d))$$

and $\mathbf{P}_u$'s are known to the adversary for all users. Note that the choice of $\mathbf{P}_u$ is not unique; nevertheless, as long as we fix a specific $\mathbf{P}_u$, we can proceed with the proof. We define E_d as the set of d edges whose $p(i,j)$'s belong to $\mathbf{P}_u$. Let $R_{\mathbf{p}} \subset \mathbb{R}^d$ be the range of acceptable values for $\mathbf{P}_u$. For example, in Figure 4 we have $|E| = 6$ and $r = 3$, so we have three independent transition probabilities. If we choose p_1, p_2, and p_3 according to the figure, we obtain the following region

$$R_{\mathbf{p}} = \{(p_1,p_2,p_3) \in \mathbb{R}^3 : \\ 0 \le p_i \le 1 \text{ for } i = 1,2,3 \text{ and } p_1 + p_2 \le 1\}.$$

The statistical properties of each user are completely known to the adversary since she knows the Markov chain of each user. The adversary wants to be able to distinguish between users by having m observations per user and also knowing $\mathbf{P}_u$'s for all users.

In this model, we assume that $\mathbf{P}_u$ for each user u is drawn independently from a d-dimensional continuous density function, $f_{\mathbf{P}}(\mathbf{p})$. As before, we assume there exist positive constants $\delta_1, \delta_2 > 0$, such that

$$\begin{cases} \delta_1 < f_{\mathbf{P}}(\mathbf{p}) < \delta_2 & \mathbf{p} \in R_{\mathbf{p}} \\ f_{\mathbf{P}}(\mathbf{p}) = 0 & \mathbf{p} \notin R_{\mathbf{p}} \end{cases}$$

It is worth noting that, by the above setting, the event $p(i,j) = 0$ has zero probability of occurring, so all users will have irreducible Markov chains if the initial graph is irreducible. If there are additional users with different graph structures, we can just ignore them as they do not impact our analysis.

We now claim that the adversary's position in this problem is mathematically equivalent to the i.i.d. model where the number of locations r is equal to $d+1$ where $d = |E| - r$. Before providing the step by step argument, we note that main idea behind the similarities between the above problem and the i.i.d. case lie in the fact both cases have similar forms of probabilities. In particular,

$$\prod_{i=1}^{r} p_i^{m_i} \quad \text{for the i.i.d. case,}$$

$$\prod_{(i,j)\in E} p_{ij}^{m_{ij}} \quad \text{for the Markov chain.}$$

(Here, p_i shows the probability of being at state i and m_i is the number of visits to state i for the i.i.d. case, while p_{ij} shows the probability of going from state i to state j and m_{ij} is the number of transitions from state i to state j for the Markov chain model.)

Let us now, discuss the equivalence in detail. First, note that since the Markov chain is irreducible and aperiodic, it has a unique stationary distribution which is equal to the limiting distribution. Next, define $\mathbf{Q}_u$ to be the vector consisting of all the transition probabilities of user u. In particular, based on the above argument, we can represent $\mathbf{Q}_u$ in the following way:

$$\mathbf{Q}_u = [\mathbf{P}_u, \mathbf{P}_u\mathbf{B}],$$

where $\mathbf{B}$ is a non-random d by $|E| - d$ matrix. Now, note that $\mathbf{P}_u\mathbf{B}$ is a non-random function of $\mathbf{P}_u$. In particular, if $M_u(i,j)$ shows the observed number transitions from state i to state j for user u, then we only need to know $M_u(i,j)$ for the edges in E_d, as the rest will be determined by the linear transform defined by $\mathbf{B}$. This implies that the decision problem for the adversary is reduced to the decision problem on transition probabilities in $\mathbf{P}_u$ and the adversary only needs to look at the $M_u(i,j)$'s for the edges in E_d. To finish the proof, we now argue that this problem has the same structure as the i.i.d. model where the number of locations r is equal to $d+1$ where $d = |E| - r$. To do so we follow the following steps:

1) If the total number of observations per user is $m = m(n)$, then define $M_i(u)$ to be the total number of visits by user u to state (location) i, for $i = 0,1,\cdots,r-1$.
2) Since the Markov chain is irreducible and aperiodic, and $m(n) \to \infty$, all $\frac{M_i(u)}{m(n)}$ converge to their stationary values.
3) Therefore, there exists positive constants that c_3 and c_4 such that $c_3 m(n) < M_i(u) < c_4 m(n)$ with arbitrarily high probability. The specific values of c_3 and c_4 do not impact the argument. In particular, let $m_i(u)$ be the observed value of $M_i(u)$, so $m_i(u) = c_5 m(n)$ for some positive c_5, where $c_3 < c_5 < c_4$.
4) Now we note that conditioned on $M_i(u) = m_i(u)$, the number transitions from state i to states j for user u

$(M_u(i,j))$ follow a multinomial distribution with probabilities $p_u(i,j)$. This is simply due to the Markov chain property (Figure 5).

5) Now let's choose one specific state i, by above the $M_u(i,j)$'s corresponding to this state for each user follow multinomial distribution with probabilities $p_u(i,j)$. We notice that this has the exact setting of the i.i.d. case. So if there are d_i edges connected to i which correspond to the probabilities in $\mathbf{P}_u$, we can argue as in the i.i.d. case (Theorem 2) that

$$m(n)^{d_i} = cn^{2-\alpha_i}, \text{ where } \alpha_i > 0$$

is a sufficient condition for perfect privacy if the adversary has access only to the information of edges connected to node i (that also correspond to one of the probabilities in $\mathbf{P}_u$).

6) But so far we have considered only one node. The adversary can look at $M_u(i,j)$'s for all $i = 0,1,\cdots,r-1$. The same line of reasoning can be repeated to obtain that

$$m(n)^{d_1+d_2+\cdots+d_r} = cn^{2-\alpha} \text{ where } \alpha > 0$$

is a sufficient condition for perfect privacy. Now note that $d_1 + d_2 + \cdots + d_r = d = |E| - r$ and this completes the proof.

□

Discussion: One limitation of the above formulations is that all users must have the same Markov chain graphs with only different transition probabilities. In reality, there might be users that have different Markov chain graphs. For example, we might have users that never visit a specific region. Nevertheless, we can address this issue in the following way. If we are considering the location privacy of user 1, we only consider users that have the same Markov chain graph (but with different transition probabilities). The other users are easily distinguishable from user 1 anyway. Now, n would be the total number of this new set of users and again we can apply Theorem 3. If n is not large enough in this case, then we need to use location obfuscation techniques in addition to anonymization to achieve perfect privacy.

VI. Simulation and Numerical Evaluation

In this section, we provide two sets of computer simulations and check their consistency with the theoretical results: First, we generate locations of users based on a Markov chain model. Next, we use real-world location traces of mobile users. In both cases, we compute the error probability of the adversary based on nearest-neighbor matching. We now provide the details of each set of simulations.

A. Simulated Locations

Here, we provide some simulation results that verify the result in Theorem 3. We consider a network with n users and r locations. Possible path of each user can be modeled as a Markov chain with r states and $|E|$ number of edges. After obtaining m observations per user, the adversary estimates transition probabilities $p_u(i,j)$ as $\frac{M_u(i,j)}{m_i(u)}$ and by using nearest neighbor decoding in $\mathbb{R}^d$, she matches a user to the observed paths.

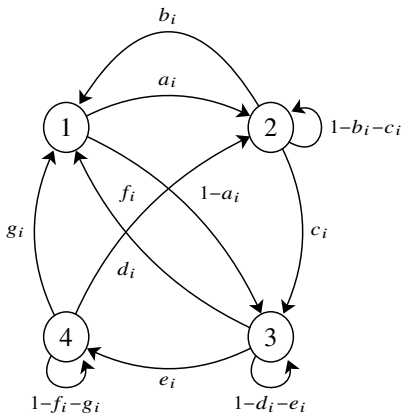

Fig. 6: The Markov chain MC which models of users' path.

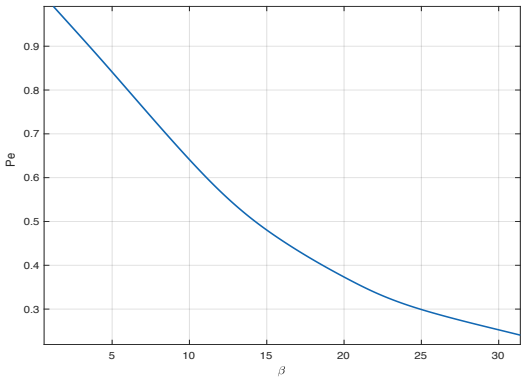

Fig. 7: $P_e(\beta)$ vs. β for Markov chain MC with $n = 500$.

We see that if the adversary's number of observations, m, is more than $O(n^{\frac{2}{|E|-r}})$, then the adversary's error probability (when adversary fails to map user with pseudonym $\Pi(u)$ to user u) goes to zero. On the other hand, if the number of observations is much smaller, then the error probability goes to one suggesting that users can achieve perfect location privacy.

In our simulations we consider $r = 4$ and $m = \beta n^{\frac{2}{|E|-r}}$. We model each user's path as a Markov chain MC shown in Figure 6. Since in this model $|E| = 11$ we can write $m = \beta n^{\frac{2}{7}}$.

In order to have n unique users, we generate each user's transition probabilities $\mathbf{P}_u$ at random based on a uniform distribution on R_p and we consider them known to the adversary. For $n = 500$, simulation results in Figure 7 show that as β grows, the adversary's error probability goes to zero, which shows that the adversary maps users with less error probability. On the other hand, as β becomes smaller, the error probability approaches 1. This is consistent with our main result that users have perfect privacy if the adversary obtains less than $O(n^{\frac{2}{|E|-r}})$ observations per user.

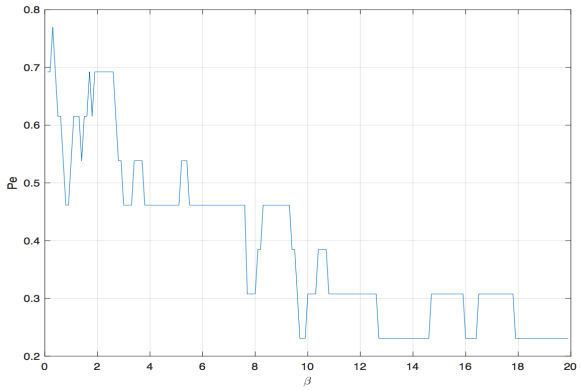

Fig. 8: $P_e(\beta)$ vs. β for 19 users using real-world traces.

B. Evaluation Using Real-World Mobility Locations

We used real-world location traces of mobile users[2] in our evaluations. The location area that we considered is in the latitude range of $(39.1, 40.1)$ and the longitude range of $(116.2, 116.4)$, which gives us a 17km by 17km area that can well represent a typical city area. We divided this area into four regions with 6 edges connecting every two regions. We used 70 percent of the reported locations as our learning set, i.e., to obtain the users' Markov chains. We used the rest of the locations for the testing phase. We selected 19 users with sufficient number of reported locations in each hour to test our theorem. Our setting is comprised of $n = 19$ total users, $r = 4$ locations, and each Markov Chain has $E = 6$ edges. As we stated in Theorem 3, if the number of observations by the adversary, m, is more than $O(n^{\frac{2}{|E|-r}})$, then she can de-anonymize users with low error probability. Here we consider m to be $\beta n^{\frac{2}{|E|-r}} = \beta * 19$.

We simulated the adversary as follows: First, during the learning phase, the adversary builds Markov chains for each user using the learning data set. We simulate the detection phase by picking m observed locations for each user and constructing an estimated making Markov chain for each user. The goal of our adversary is to find the user who is most probable to create each trace. We therefore compare the observed Markov models with the Markov models of all the users obtained in the learning stage and pick a user whose Markov model has minimum distance with the observed Markov chain as the user who has created the trace. The distance between two chains is computed as the Euclidean distance between the vector representing their transition probabilities. Figure 8 shows the probability of error, i.e., incorrect matching by the adversary, for various β (larger values of β indicate more observations by the adversary). As the figure shows, increasing β reduces the error probability of the adversary, which is consistent with our main theorem. Even though our theorem was proved for a large number of users, the figure shows that its main finding holds even for our small-size setting.

[2]The traces are obtained from GeoLife GPS Trajectories. We use a GPS trajectory dataset collected by the Microsoft Research Asia Geolife project on 182 users in a period of over three years (from April 2007 to August 2012). The data set is available here: https://www.microsoft.com/en-us/download/details.aspx?id=52367

VII. Conclusion

We presented an information theoretic definition for perfect location privacy using the mutual information between the users' actual locations and the anonymized observations that the adversary collects. First, we modeled users' movements to be independent from their previous locations. In this model, we have n users and r locations. We prove that if the number of anonymized observations that the adversary collects, m, is less than $O(n^{\frac{2}{r-1}})$ then users will have perfect location privacy. So, if the anonymization method changes pseudonyms of users before $O(n^{\frac{2}{r-1}})$ observations is made by the adversary for each user, then the adversary cannot distinguish between users and they can achieve perfect location privacy. Then, we modeled users' movements using Markov chains so that their current locations affect their next moves. We proved that for such a user, perfect location privacy is achievable if the pseudonym of the user is changed before $O(n^{\frac{2}{|E|-r}})$ number of observations is made by the adversary.

Appendix A
Proof of Theorem 1 (Perfect Location Privacy for Two-State Model)

Here, we provide a formal proof for Theorem 1. In the proposed setting, we assume we have an infinite number of potential users indexed by integers, and at any step we consider a network consisting of n users, i.e., users $1, 2, \cdots, n$. We would like to show perfect location privacy when n goes to infinity. Remember that $X_u(t)$ shows the location of user u at time t.

In a two-state model, let us assume we have state 0 and state 1. There is a sequence $p_1, p_2, p_3, \cdots$ for the users. In particular, for user u we have $p_u = P(X_u(k) = 1)$ for times $k = 1, 2, \cdots$. Thus, the locations of each user u are determined by a Bern(p_u) process.

When we set $n \in \mathbb{N}$ as the number of users, we assume m to be the number of adversary's observations per user,

$$m = m(n) = cn^{2-\alpha} \quad \text{where} \quad 0 < \alpha < 1.$$

So, we have $n \to \infty$ if and only if $m \to \infty$.

As defined previously, $\mathbf{X}_u^{(m)}$ contains m number of user u's locations and $\mathbf{X}^{(m)}$ is the collection of $\mathbf{X}_u^{(m)}$'s for all users,

$$\mathbf{X}_u^{(m)} = \begin{bmatrix} X_u(1) \\ X_u(2) \\ \vdots \\ X_u(m) \end{bmatrix}, \quad \mathbf{X}^{(m)} = \left(\mathbf{X}_1^{(m)}, \mathbf{X}_2^{(m)}, \cdots, \mathbf{X}_n^{(m)}\right).$$

The permutation function applied to anonymize users is $\Pi^{(n)}$ (or simply Π). For any set $A \subset \{1, 2, \cdots, n\}$, we define

$$\Pi(A) = \{\Pi(u) : u \in A\}.$$

The adversary who knows all the p_u's, observes n anonymized users for m number of times each and collects their locations in $\mathbf{Y}^{(m)}$

$$\begin{aligned} \mathbf{Y}^{(m)} &= \text{Perm}\left(\mathbf{X}_1^{(m)}, \mathbf{X}_2^{(m)}, \cdots, \mathbf{X}_n^{(m)}; \Pi\right) \\ &= \left(\mathbf{Y}_1^{(m)}, \mathbf{Y}_2^{(m)}, \cdots, \mathbf{Y}_n^{(m)}\right) \end{aligned}$$

where $\mathbf{Y}_u^{(m)} = \mathbf{X}_{\Pi^{-1}(u)}^{(m)}, \mathbf{Y}_{\Pi(u)}^{(m)} = \mathbf{X}_u^{(m)}$.

Based on the assumptions of Theorem 1, if the following holds

1) $m = cn^{2-\alpha}$, which $c > 0, 0 < \alpha < 1$ 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 want to show

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

i.e., user 1 has perfect location privacy and the same applies for all other users.

A. Proof procedure

Steps of the proof are as follows:

1) We show that there exists a sequence of sets $J^{(n)} \subseteq \{1, 2, \cdots, n\}$ with the following properties:
   - $1 \in J^{(n)}$
   - If $N^{(n)} = |J^{(n)}|$ then, $N^{(n)} \to \infty$ as $n \to \infty$
   - Let $\{j_n\}_{n=1}^{\infty}$ be any sequence such that $j_n \in \Pi(J^{(n)})$ then
   $$P\left(\Pi(1) = j_n | \mathbf{Y}^{(m)}, \Pi(J^{(n)})\right) \to 0$$
2) We show that
   $$X_1(k) | \mathbf{Y}^{(m)}, \Pi(J^{(n)}) \xrightarrow{d} \text{Bern}(p_1).$$
3) Using 2, we conclude
   $$H\left(X_1(k) | \mathbf{Y}^{(m)}, \Pi(J^{(n)})\right) \to H\left(X_1(k)\right)$$
   and in conclusion,
   $$I\left(X_1(k); \mathbf{Y}^{(m)}\right) \to 0.$$

B. Detail of the proof

We define $S_u^{(m)}$ for $u = 1, 2, \cdots, n$ to be the number of times that user u was at state 1,

$$S_u^{(m)} = X_u(1) + X_u(2) + \cdots + X_u(m).$$

Based on the assumptions, we have $S_u^{(m)} \sim \text{Binomial}(m, p_u)$. One benefit of $S_u^{(m)}$'s is that they provide a sufficient statistic for the adversary when the adversary's goal is to obtain the permutation $\Pi^{(n)}$. To make this statement precise, let's define $\mathbf{S}^{(m)}$ as the vector containing $S_u^{(m)}$, for $u = 1, 2, \cdots, n$:

$$\mathbf{S}^{(m)} = \left(S_1^{(m)}, S_2^{(m)}, \cdots, S_n^{(m)}\right)$$

Note that for $u = 1, 2, \cdots, n$

$$\begin{aligned} S_u^{(m)} &= X_u(1) + X_u(2) + \cdots + X_u(m) \\ &= Y_{\Pi(u)}(1) + Y_{\Pi(u)}(2) + \cdots + Y_{\Pi(u)}(m). \end{aligned}$$

Thus, the adversary can obtain $\text{Perm}(\mathbf{S}^{(m)}, \Pi^{(n)})$, a permuted version of $\mathbf{S}^{(m)}$, by adding the elements in each column of $\mathbf{Y}^{(m)}$. In particular, we can write

$$\begin{aligned} \text{Perm}(\mathbf{S}^{(m)}, \Pi^{(n)}) &= \text{Perm}\left(S_1^{(m)}, S_2^{(m)}, \cdots, S_n^{(m)}; \Pi^{(n)}\right) \\ &= \left(S_{\Pi^{-1}(1)}^{(m)}, S_{\Pi^{-1}(2)}^{(m)}, \cdots, S_{\Pi^{-1}(n)}^{(m)}\right). \end{aligned}$$

We now state and prove a lemma that confirms $\text{Perm}(\mathbf{S}^{(m)}, \Pi^{(n)})$ is a sufficient statistic for the adversary when the adversary's goal is to recover $\Pi^{(n)}$. The usefulness of this lemma will be clear since we can use the law of total probability to break the adversary's decision problem into two steps of (1) obtaining the posterior probability distribution for $\Pi^{(n)}$ and (2) estimating the locations $X_u(k)$ given the choice of $\Pi^{(n)}$.

Lemma 2. Given $\text{Perm}(\mathbf{S}^{(m)}, \Pi^{(n)})$, the random matrix $\mathbf{Y}^{(m)}$ and the random permutation $\Pi^{(n)}$ are conditionally independent. That is

$$\begin{aligned} P\left(\Pi^{(n)} = \pi \;\middle|\; \mathbf{Y}^{(m)} = \mathbf{y}, \text{Perm}(\mathbf{S}^{(m)}, \Pi^{(n)}) = \mathbf{s}\right) = \\ P\left(\Pi^{(n)} = \pi \;\middle|\; \text{Perm}(\mathbf{S}^{(m)}, \Pi^{(n)}) = \mathbf{s}\right) \end{aligned} \tag{4}$$

Proof. Remember

$$\begin{aligned} \mathbf{Y}^{(m)} &= \text{Perm}\left(\mathbf{X}_1^{(m)}, \mathbf{X}_2^{(m)}, \cdots, \mathbf{X}_n^{(m)}; \Pi^{(n)}\right) \\ &= \left(\mathbf{X}_{\Pi^{-1}(1)}^{(m)}, \mathbf{X}_{\Pi^{-1}(2)}^{(m)}, \cdots, \mathbf{X}_{\Pi^{-1}(n)}^{(m)}\right). \end{aligned}$$

Note that $\mathbf{Y}^{(m)}$ (and therefore $\mathbf{y}$) is an m by n matrix, so we can write

$$\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n),$$

where for $u = 1, 2, \cdots, n$, we have

$$\mathbf{y}_u = \begin{bmatrix} y_u(1) \\ y_u(2) \\ \vdots \\ y_u(m) \end{bmatrix}.$$

Also, $\mathbf{s}$ is a 1 by n vector, so we can write

$$\mathbf{s} = (s_1, s_2, \cdots, s_n).$$

It suffices to show that

$$\begin{aligned} P\left(\mathbf{Y}^{(m)} = \mathbf{y} \;\middle|\; \Pi^{(n)} = \pi, \text{Perm}(\mathbf{S}^{(m)}, \Pi^{(n)}) = \mathbf{s}\right) = \\ P\left(\mathbf{Y}^{(m)} = \mathbf{y} \;\middle|\; \text{Perm}(\mathbf{S}^{(m)}, \Pi^{(n)}) = \mathbf{s}\right) \end{aligned} \tag{5}$$

More specifically, it suffices to show that the probability in the righthand side expression does not depend on $\Pi^{(n)} = \pi$. This is in fact a direct result of the fact that for a sequence of Bernoulli random variables with length m, given the total number of 1's is equal to s, then all sequences with s number of 1's are equally likely. In particular, we can write

$$\begin{aligned} P\left(\mathbf{Y}^{(m)} = \mathbf{y} \;\middle|\; \Pi^{(n)} = \pi, \text{Perm}(\mathbf{S}^{(m)}, \Pi^{(n)}) = \mathbf{s}\right) = \\ P\Bigg(\mathbf{Y}_u^{(m)} = \mathbf{y}_u,\ u \in [n] \;\Bigg|\; \sum_{k=1}^{m} \mathbf{Y}_u^{(m)}(k) = s_u,\ u \in [n], \\ \mathbf{Y}_u^{(m)}(k) \sim \text{Bern}(p_{\pi^{-1}(u)}),\ u \in [n]\Bigg) \end{aligned}$$

where $[n] = \{1, 2, \cdots, n\}$. Thus, by the independence assumption we can write

$$P\left(\mathbf{Y}^{(m)} = \mathbf{y} \;\middle|\; \Pi^{(n)} = \pi, \text{Perm}(\mathbf{S}^{(m)}, \Pi^{(n)}) = \mathbf{s}\right) = \prod_{u=1}^{n} P\left(\mathbf{Y}_u^{(m)} = \mathbf{y}_u \;\middle|\; \sum_{k=1}^{m} \mathbf{Y}_u^{(m)}(k) = s_u, \mathbf{Y}_u^{(m)}(k) \sim \text{Bern}(p_{\pi^{-1}(u)})\right) = \prod_{u=1}^{n} \frac{1}{\binom{m}{s_u}}.$$

Since this probability does not depend on π, we conclude that given $\text{Perm}(\mathbf{S}^{(m)}, \Pi^{(n)})$, the random matrix $\mathbf{Y}^{(m)}$ and the random permutation $\Pi^{(n)}$ are conditionally independent.

□

Next, we need to turn our attention to defining the critical set $J^{(n)}$. First, remember that

$$m = cn^{2-\alpha} \quad \text{where} \quad 0 < \alpha < 1.$$

We choose real numbers θ and ϕ such that $0 < \theta < \phi < \frac{\alpha}{2(2-\alpha)}$, and define

$$\epsilon_m \triangleq \frac{1}{m^{\frac{1}{2}+\phi}} \quad \beta_m \triangleq \frac{1}{m^{\frac{1}{2}-\theta}}.$$

We now define the set $J^{(n)}$ for any positive integer n as follows: Set $J^{(n)}$ consists of the indices of users such that the probability of them being at state 1 is within a range with ϵ_m difference around p_1,

$$J^{(n)} = \{i \in \{1, 2, \ldots, n\} : p_1 - \epsilon_m < p_i < p_1 + \epsilon_m\}.$$

Clearly for all n, $1 \in J^{(n)}$. The following lemma confirms that the number of elements in $J^{(n)}$ goes to infinity as $n \to \infty$.

Lemma 3. If $N^{(n)} \triangleq |J^{(n)}|$, then $N^{(n)} \to \infty$ as $n \to \infty$. More specifically, as $n \to \infty$,

$$\exists \lambda, c'' > 0 : \quad P(N^{(n)} > c''n^{\lambda}) \to 1 \qquad \text{as} \quad n \to \infty.$$

Proof. Remember that we assume p_u's are drawn independently from some continuous density function, $f_P(p)$, on the $(0, 1)$ interval which satisfies

$$\begin{cases} \delta_1 < f_P(p) < \delta_2 & p \in (0, 1) \\ f_P(p) = 0 & p \notin (0, 1) \end{cases}$$

So given $p_1 \in (0, 1)$, for n large enough (so that ϵ_m is small enough), we have

$$P(p_1 - \epsilon_m < p_i < p_1 + \epsilon_m) = \int_{p_1-\epsilon_m}^{p_1+\epsilon_m} f_P(p)dp,$$

so we can conclude that

$$2\epsilon_m\delta_1 < P(p_1 - \epsilon_m < p_i < p_1 + \epsilon_m) < 2\epsilon_m\delta_2.$$

We can find a δ such that $\delta_1 < \delta < \delta_2$ and

$$P(p_1 - \epsilon_m < p_i < p_1 + \epsilon_m) = 2\epsilon_m\delta.$$

Then, we can say that $N^{(n)} \sim \text{Binomial}(n, 2\epsilon_m\delta)$, where

$$\epsilon_m = \frac{1}{m^{\frac{1}{2}+\phi}} = \frac{1}{(cn^{2-\alpha})^{(\frac{1}{2}+\phi)}}.$$

The expected value of $N^{(n)}$ is $n2\epsilon_m\delta$, and by substituting ϵ_m we get

$$\text{E}[N^{(n)}] = n2\epsilon_m\delta = \frac{n2\delta}{(c'n^{2-\alpha})^{(\frac{1}{2}+\phi)}} = c''n^{(\frac{\alpha}{2}+\alpha\phi-2\phi)}.$$

Let us set $\lambda = \frac{\alpha}{2} + \alpha\phi - 2\phi$. Since $\phi < \frac{\alpha}{2(2-\alpha)}$, we have $\lambda > 0$. Therefore, we can write

$$\text{E}[N^{(n)}] = c''n^{\lambda},$$

$$\text{Var}(N^{(n)}) = n(2\epsilon_m\delta)(1 - 2\epsilon_m\delta) \to n^{\lambda}(1 + o(1)).$$

Using Chebyshev's inequality

$$P(|N^{(n)} - \text{E}[N^{(n)}]| > \frac{c''}{2}n^{\lambda}) < \frac{n^{\lambda}(1 + o(1))}{\frac{c''^2}{4}n^{2\lambda}} \to 0$$

$$P(N^{(n)} > \frac{c''}{2}n^{\lambda}) \to 1 \qquad \text{as} \quad n \to \infty.$$

□

The next step in the proof is to show that users that are identified by the set $J^{(n)}$ produce a very similar moving process as user 1. To make this statement precise, we provide the following definition. Define the set $A^{(m)}$ as the interval in $\mathbb{R}$ consisting of real numbers which are within the $m\beta_m$ distance from mp_1 (the expected number of times that user 1 is at state 1 during the m number of observations),

$$A^{(m)} = \{x \in R, m(p_1 - \beta_m) \le x \le m(p_1 + \beta_m)\}.$$

Lemma 4. We have

$$P\left(\bigcap_{j \in J^{(n)}} \left(S_j^{(m)} \in A^{(m)}\right)\right) \to 1 \qquad \text{as} \quad n \to \infty$$

Proof. Let $j \in J^{(n)}$ and $p_1 - \epsilon_m < p_j < p_1 + \epsilon_m$. Since $S_j^{(m)} \sim \text{Binomial}(m, p_j)$, by the Large Deviation Theory (Sanov's Theorem), we can write

$$P\left(S_j^{(m)} > m(p_1 + \beta_m)\right) < (m + 1)2^{-mD\left(\text{Bern}(p_1+\beta_m)\|\text{Bern}(p_j)\right)} \tag{6}$$

By using the fact that for all $p \in (0, 1)$ [3]

$$D\left(\text{Bern}(p + \epsilon)\|\text{Bern}(p)\right) = \frac{\epsilon^2}{2p(1-p)\ln 2} + O(\epsilon^3),$$

we can write

$$D\left(\text{Bern}(p_1 + \beta_m)\|\text{Bern}(p_j)\right) = \frac{(p_1 + \beta_m - p_j)^2}{2p_j(1-p_j)\ln 2} + O\left((p_1 + \beta_m - p_j)^3\right).$$

Note that $\left|p_1 - p_j\right| < \epsilon_m$, so for large m we can write

$$\left|p_1 + \beta_m - p_j\right| \ge \beta_m - \epsilon_m = \frac{1}{m^{\frac{1}{2}-\theta}} - \frac{1}{m^{\frac{1}{2}+\phi}} > \frac{\frac{1}{2}}{m^{\frac{1}{2}-\theta}}.$$

so we can write

$$D\left(\text{Bern}(p_1 + \beta_m)\|\text{Bern}(p_j)\right) =$$

[3] It is worth noting that in the standard large deviation literature, the case where $\epsilon > 0$ is a constant independent of n is usually considered, nevertheless, Equation 6 remains valid for all values of β_m>0.

$$\frac{1}{8p_j(1-p_j)m^{1-2\theta}\ln 2}+O((p_1+\beta_m-p_j)^3)$$

and for some constant $c'>0$

$$D\left(\text{Bern}(p_1+\beta_m)\|\text{Bern}(p_j)\right)>\frac{c'}{m^{1-2\theta}}\Rightarrow$$

$$mD\left(\text{Bern}(p_1+\beta_m)\|\text{Bern}(p_j)\right)>\frac{mc'}{m^{1-2\theta}}>c'm^{2\theta}\Rightarrow$$

$$P(S_j^{(m)}>m(p_1+\beta_m))<m2^{-c'm^{2\theta}}.$$

So in conclusion

$$P\left(\bigcup_{j\in J^{(n)}}S_j^{(m)}>m(p_1+\beta_m)\right)<|J^{(n)}|m2^{-c'm^{2\theta}}$$

$$|J^{(n)}|m2^{-c'm^{2\theta}}<nm2^{-c'm^{2\theta}}<m^2 2^{-c'm^{2\theta}}\to 0\quad\text{as}\quad m\to\infty.$$

Similarly we obtain

$$P\left(\bigcup_{j\in J^{(n)}}S_j^{(m)}<m(p_1-\beta_m)\right)\to 0\quad\text{as}\quad m\to\infty,$$

which completes the proof. This shows that for all users j for which p_j is within ϵ range around p_1, i.e. it is in set $J^{(n)}$, the average number of times that this user was at state 1 is within $m\beta_m$ from mp_1 with high probability. □

We are now in a position to show that distinguishing between the users in $J^{(n)}$ is not possible for an outside observer (i.e., the adversary) and this will pave the way in showing perfect location privacy.

Lemma 5. Let $\{a_m\}_{m=1}^{\infty}$, $\{b_m\}_{m=1}^{\infty}$ be such that a_m, b_m are in set $A^{(m)}$ and also $\{i_m\}_{m=1}^{\infty}$, $\{j_m\}_{m=1}^{\infty}$ be such that i_m, j_m are in set $J^{(n)}$. Then, we have

$$\frac{P\left(S_{i_m}^{(m)}=a_m,S_{j_m}^{(m)}=b_m\right)}{P\left(S_{i_m}^{(m)}=b_m,S_{j_m}^{(m)}=a_m\right)}\to 1\quad\text{as}\quad m\to\infty.$$

Proof. Remember that

$$A^{(m)}=\{x\in R, m(p_1-\beta_m)\le x\le m(p_1+\beta_m)\}$$

where $\beta_m=\frac{1}{m^{\frac{1}{2}-\theta}}$ and $S_j^{(m)}\sim\text{Binomial}(m,p_j)$. Thus, $S_{i_m}^{(m)}\sim\text{Binomial}(m,p_{i_m})$ and $S_{j_m}^{(m)}\sim\text{Binomial}(m,p_{j_m})$,

$$P(S_{i_m}^{(m)}=a_m)=\binom{m}{a_m}p_{i_m}^{a_m}(1-p_{i_m})^{m-a_m},$$

$$P(S_{j_m}^{(m)}=b_m)=\binom{m}{b_m}p_{j_m}^{b_m}(1-p_{j_m})^{m-b_m}.$$

In conclusion,

$$\Delta_m=\frac{P\left(S_{i_m}^{(m)}=a_m,S_{j_m}^{(m)}=b_m\right)}{P\left(S_{i_m}^{(m)}=b_m,S_{j_m}^{(m)}=a_m\right)}=$$

$$(\frac{p_{i_m}}{p_{j_m}})^{a_m-b_m}(\frac{1-p_{j_m}}{1-p_{i_m}})^{a_m-b_m}$$

$$\ln\Delta_m=(a_m-b_m)\ln(\frac{p_{i_m}}{p_{j_m}})+(a_m-b_m)\ln(\frac{1-p_{j_m}}{1-p_{i_m}})$$

and since $\{i_m,j_m\}\in J^{(n)}$ we have

$$\left|p_{i_m}-p_{j_m}\right|\le 2\epsilon_m=\frac{2}{m^{\frac{1}{2}+\phi}}.$$

Also, since $\{a_m,b_m\}\in A^{(m)}$ we can say that

$$|a_m-b_m|\le 2m\beta_m.$$

Since $p_{i_m}\le p_{j_m}+2\epsilon_m$ and $1-p_{j_m}\le(1-p_{i_m})+2\epsilon_m$ and

$$\ln(1+\epsilon_m)=\epsilon_m+O(\epsilon_m^2)$$

we can write

$$\ln\Delta_m\le 2m\beta_m\epsilon_m+2m\beta_m\epsilon_m+2m\beta_m O(\epsilon_m^2)$$

and since $\phi>\theta$,

$$m\beta_m\epsilon_m=m\frac{1}{m^{\frac{1}{2}+\phi}}\frac{1}{m^{\frac{1}{2}-\theta}}=\frac{1}{m^{\phi-\theta}}\to 0,$$

$$\Rightarrow\ln\Delta_m\to 0$$

$$\Rightarrow\Delta_m\to 1.$$

Note that the convergence is uniform in the sense that we only used the fact that a_m, b_m are in set $A^{(m)}$ and not the specific choices of a_m and b_m. That is, following the above argument allows us to choose for any $\eta>0$, a value $n_\eta>0$, such that

$$\left|\frac{P\left(S_{i_m}^{(m)}=a_m,S_{j_m}^{(m)}=b_m\right)}{P\left(S_{i_m}^{(m)}=b_m,S_{j_m}^{(m)}=a_m\right)}-1\right|<\eta\quad\text{for all } n>n_\eta,$$

as long as the sequences $\{a_m\}_{m=1}^{\infty}$, $\{b_m\}_{m=1}^{\infty}$ be such that a_m, b_m are in set $A^{(m)}$.

This shows that for two users i and j, if the probability of them being at state 1 is in set $J^{(n)}$, $p_i, p_j\in J^{(n)}$, and also the observed number of times for these users to be at state 1 is in set $A^{(m)}$, then distinguishing between these two users is impossible. □

Lemma 6. For any $j\in\Pi(J^{(n)})$, we define $W_j^{(n)}$ as follows

$$W_j^{(n)}=P\left(\Pi(1)=j|\mathbf{Y}^{(m)},\Pi(J^{(n)})\right).$$

Then, for all $j^{(n)}\in\Pi(J^{(n)})$,

$$N^{(n)}W_j^{(n)}\xrightarrow{p}1.$$

More specifically, for all $\gamma_1,\gamma_2>0$, there exists n_o such that if $n>n_o$:

$$\forall j\in\Pi(J^{(n)}):\quad P\left(\left|N^{(n)}W_j^{(n)}-1\right|>\gamma_1\right)<\gamma_2.$$

Proof. This is the result of Lemma 5. First, remember that

$$\sum_{j\in\Pi(J^{(n)})}W_j^{(n)}=1,$$

and also note that

$$|\Pi(J^{(n)})|=|J^{(n)}|=N^{(n)}\to\infty\quad\text{as}\quad n\to\infty.$$

Here, we show that for any $\{j_n\}_{n=1}^{\infty}\in\Pi(J^{(n)})$,

$$\frac{W_{j_n}^{(n)}}{W_1^{(n)}}=\frac{P\left(\Pi(1)=j|D\right)}{P\left(\Pi(1)=1|D\right)}\xrightarrow{p}1$$

where $D = \left(\mathbf{Y}^{(m)}, \Pi(J^{(n)})\right)$.

Let a_i, for $i \in \Pi(J^{(n)})$, be the permuted observed values of $S_i^{(m)}$'s. Then note that

$$P(\Pi(1) = j|D) \propto \sum_{\substack{\text{permutation} \\ \text{such that } \Pi(1)=j}} \prod_{i \in \Pi(J)} P(S_i^{(m)} = a_i).$$

Then, in

$$\frac{W_{j_n}^{(n)}}{W_1^{(n)}} = \frac{P(\Pi(1) = j|D)}{P(\Pi(1) = 1|D)}$$

the numerator and denominator have the same terms. In particular, for each term

$$P(S_j^{(m)} = a_{j_n}) \times P(S_1^{(m)} = b_{j_n})$$

in $W_j^{(n)}$, there is a corresponding term

$$P(S_j^{(m)} = b_{j_n}) \times P(S_1^{(m)} = a_{j_n})$$

in $W_1^{(n)}$. Since by Lemma 5

$$\frac{P(S_j^{(m)} = a_{j_n}) \times P(S_1^{(m)} = b_{j_n})}{P(S_j^{(m)} = b_{j_n}) \times P(S_1^{(m)} = a_{j_n})}$$

converges uniformly to 1, we conclude

$$\frac{W_{j_n}^{(n)}}{W_1} \to 1.$$

We conclude that for any $\zeta > 0$, we can write (for large enough n)

$$(1 - \zeta) < \frac{W_{j_n}^{(n)}}{W_1} < (1 + \zeta),$$

$$\sum_{j \in \Pi(J^{(n)})} (1 - \zeta) W_1^{(n)} < \sum_{j \in \Pi(J^{(n)})} W_{j_n}^{(n)} < \sum_{j \in \Pi(J^{(n)})} (1 + \zeta) W_1^{(n)}$$

and since $\sum_{j \in \Pi(J^{(n)})} W_{j_n}^{(n)} = 1$, $|\Pi(J^{(n)})| = N^{(n)}$, we have

$$(1 - \zeta) N^{(n)} W_1^{(n)} < 1 < (1 + \zeta) N^{(n)} W_1^{(n)}$$

Therefore

$$\frac{1}{1 + \zeta} < N^{(n)} W_1^{(n)} < \frac{1}{1 - \zeta}$$

and, we conclude that $N^{(n)} W_1^{(n)} \to 1$ as $n \to \infty$. We can repeat the same argument for all users in set $j \in J^{(n)}$ and we can show for all $\zeta > 0$

$$\frac{1}{1 + \zeta} < N^{(n)} W_j^{(n)} < \frac{1}{1 - \zeta} \tag{7}$$

and $N^{(n)} W_j^{(n)} \to 1$ as $n \to \infty$. □

Now to finish the proof of Theorem 1,

$$P\left(X_1(k) = 1|\mathbf{Y}^{(m)}, \Pi(J^{(n)})\right) =$$

$$\sum_{j \in \Pi(J^{(n)})} P\left(X_1(k) = 1|\mathbf{Y}^{(m)}, \Pi(1) = j, \Pi(J^{(n)})\right) \times$$

$$P\left(\Pi(1) = j|\mathbf{Y}^{(m)}, \Pi(J^{(n)})\right)$$

$$= \sum_{j \in \Pi(J^{(n)})} 1_{[Y_j^{(m)}(k)=1]} W_j^{(n)} \triangleq Z_n.$$

But, since $Y_j^{(m)}(k) \sim \text{Bern}(p_j^{(n)})$ and $p_j^{(n)} \to p_1$ for all $j \in \Pi(J^{(n)})$, applying Chebyshev's inequality (law of large numbers) obtains:

$$\frac{1}{N^{(n)}} \sum_{j \in \Pi(J^{(n)})} 1_{[Y_j^{(m)}(k)=1]} \to p_1$$

Thus, we can write for any $\eta > 0$

$$\begin{aligned} Z_n &= \frac{1}{N^{(n)}} \sum_{j \in \Pi(J^{(n)})} (1_{[Y_j^{(m)}(k)=1]})(N^{(n)} W_j^{(n)}) \\ &< \frac{1}{N^{(n)}} \sum_{j \in \Pi(J^{(n)})} (1_{[Y_j^{(m)}(k)=1]}) \left(\frac{1}{1 - \zeta}\right) \quad \text{by (7)} \\ &\to \frac{p_1}{1 - \zeta}. \end{aligned}$$

We can also write

$$\begin{aligned} Z_n &= \frac{1}{N^{(n)}} \sum_{j \in \Pi(J^{(n)})} (1_{[Y_j^{(m)}(k)=1]})(N^{(n)} W_j^{(n)}) \\ &> \frac{1}{N^{(n)}} \sum_{j \in \Pi(J^{(n)})} (1_{[Y_j^{(m)}(k)=1]}) \left(\frac{1}{1 + \zeta}\right) \quad \text{by (7)} \\ &\to \frac{p_1}{1 + \zeta}. \end{aligned}$$

Since $\zeta > 0$ can be chosen arbitrarily, we conclude $Z_n \to p_1$.

In conclusion $X_1(k)|\mathbf{Y}^{(m)}, \Pi(J^{(n)}) \xrightarrow{d} \text{Bern}(p_1)$ which means that

$$H\left(X_1(k)|\mathbf{Y}^{(m)}, \Pi(J^{(n)})\right) \to H(X_1(k))$$

$$\Rightarrow H\left(X_1(k)|\mathbf{Y}^{(m)}\right) \geq H\left(X_1(k)|\mathbf{Y}^{(m)}, \Pi(J^{(n)})\right) \to H(X_1(k))$$

$$\Rightarrow I\left(X_1(k); \mathbf{Y}^{(m)})\right) \to 0$$

Appendix B
Proof of Lemma 1

Here, we provide a formal proof for Lemma 1 which we restate as follows. In the Markov chain setting of section V, we have the following: Given $\text{Perm}(\mathbf{M}^{(m)}, \Pi^{(n)})$, the random matrix $\mathbf{Y}^{(m)}$ and the random permutation $\Pi^{(n)}$ are conditionally independent. That is

$$P\left(\Pi^{(n)} = \pi \;\middle|\; \mathbf{Y}^{(m)} = \mathbf{y}, \text{Perm}(\mathbf{M}^{(m)}, \Pi^{(n)}) = \mathbf{m}\right) = P\left(\Pi^{(n)} = \pi \;\middle|\; \text{Perm}(\mathbf{M}^{(m)}, \Pi^{(n)}) = \mathbf{m}\right) \tag{8}$$

Proof. Remember

$$\begin{aligned} \mathbf{Y}^{(m)} &= \text{Perm}\left(\mathbf{X}_1^{(m)}, \mathbf{X}_2^{(m)}, \cdots, \mathbf{X}_n^{(m)}; \Pi^{(n)}\right) \\ &= \left(\mathbf{X}_{\Pi^{-1}(1)}^{(m)}, \mathbf{X}_{\Pi^{-1}(2)}^{(m)}, \cdots, \mathbf{X}_{\Pi^{-1}(n)}^{(m)}\right). \end{aligned}$$

Note that $\mathbf{Y}^{(m)}$ (and therefore $\mathbf{y}$) is an m by n matrix, so we can write

$$\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n),$$

where for $u = 1, 2, \cdots, n$, we have

$$\mathbf{y}_u = \begin{bmatrix} y_u(1) \\ y_u(2) \\ \vdots \\ y_u(m) \end{bmatrix}.$$

Also, $\mathbf{m}$ is a collection of n matrices so we can write

$$\mathbf{m} = (\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_n).$$

For an $r \times r$ matrix $\mathbf{m} = [m(i,j)]$, let's define $D(\mathbf{m})$ as the set of sequences $(x_1, x_2, \cdots, x_m) \in \{1, 2, \cdots, r\}^m$ that satisfy the following properties:

1) $x_0 = 1$;
2) The number of transitions from i to j in $(x_1, x_2, \cdots, x_m)$ is equal to m_{ij} for all i and j. That is, the number of indices k for which we have $x_k = i$ and $x_{k+1} = j$ is equal to $m(i,j)$.

We now show that the two sides of Equation 8 are equal. The right hand side probability can be written as

$$P\left(\Pi^{(n)} = \pi \;\middle|\; \text{Perm}(\mathbf{M}^{(m)}, \Pi^{(n)}) = \mathbf{m}\right) = \frac{P\left(\text{Perm}(\mathbf{M}^{(m)}, \Pi^{(n)}) = \mathbf{m} \;\middle|\; \Pi^{(n)} = \pi\right) P\left(\Pi^{(n)} = \pi\right)}{P\left(\text{Perm}(\mathbf{M}^{(m)}, \Pi^{(n)}) = \mathbf{m}\right)}$$
$$= \frac{P\left(\text{Perm}(\mathbf{M}^{(m)}, \pi) = \mathbf{m} \;\middle|\; \Pi^{(n)} = \pi\right)}{n! P\left(\text{Perm}(\mathbf{M}^{(m)}, \Pi^{(n)}) = \mathbf{m}\right)}$$
$$= \frac{P\left(\text{Perm}(\mathbf{M}^{(m)}, \pi) = \mathbf{m}\right)}{n! P\left(\text{Perm}(\mathbf{M}^{(m)}, \Pi^{(n)}) = \mathbf{m}\right)}.$$

Now note that

$$P\left(\text{Perm}(\mathbf{M}^{(m)}, \pi) = \mathbf{m}\right) = P\left(\bigcap_{j=1}^{n} \left(\mathbf{M}^{(m)}_{\pi^{-1}(j)} = \mathbf{m}_j\right)\right)$$
$$= P\left(\bigcap_{u=1}^{n} \left(\mathbf{M}^{(m)}_{u} = \mathbf{m}_{\pi(u)}\right)\right)$$
$$= \prod_{u=1}^{n} P\left(\mathbf{M}^{(m)}_{u} = \mathbf{m}_{\pi(u)}\right)$$
$$= \prod_{u=1}^{n} \sum_{\substack{(x_1, x_2, \cdots, x_m) \in \\ D(\mathbf{m}_{\pi(u)})}} P(X_u(1) = x_1, X_u(2) = x_2, \cdots, X_u(m) = x_m)$$
$$= \prod_{u=1}^{n} \sum_{(x_1, x_2, \cdots, x_m) \in D(\mathbf{m}_{\pi(u)})} \prod_{i,j} p_u(i,j)^{\mathbf{m}_{\pi(u)}(i,j)}$$
$$= \prod_{u=1}^{n} \left(|D(\mathbf{m}_{\pi(u)})| \prod_{i,j} p_u(i,j)^{\mathbf{m}_{\pi(u)}(i,j)} \right)$$
$$= \left(\prod_{k=1}^{n} |D(\mathbf{m}_k)| \right) \left(\prod_{u=1}^{n} \prod_{i,j} p_u(i,j)^{\mathbf{m}_{\pi(u)}(i,j)} \right)$$

Similarly, we obtain

$$P\Big(\text{Perm}(\mathbf{M}^{(m)}, \Pi^{(n)}) = \mathbf{m}\Big) = \sum_{\text{all permutations } \pi'} P\left(\text{Perm}(\mathbf{M}^{(m)}, \pi') = \mathbf{m} \;\middle|\; \Pi^{(n)} = \pi'\right) P\left(\Pi^{(n)} = \pi'\right)$$
$$= \frac{1}{n!} \sum_{\text{all permutations } \pi'} \left(\prod_{k=1}^{n} |D(\mathbf{m}_k)| \right) \left(\prod_{u=1}^{n} \prod_{i,j} p_u(i,j)^{\mathbf{m}_{\pi'(u)}(i,j)} \right)$$
$$= \frac{1}{n!} \left(\prod_{k=1}^{n} |D(\mathbf{m}_k)| \right) \sum_{\text{all permutations } \pi'} \left(\prod_{u=1}^{n} \prod_{i,j} p_u(i,j)^{\mathbf{m}_{\pi'(u)}(i,j)} \right).$$

Thus, we conclude that the right hand side of Equation 8 is equal to

$$\frac{\prod_{u=1}^{n} \prod_{i,j} p_u(i,j)^{\mathbf{m}_{\pi(u)}(i,j)}}{\sum_{\text{all permutations } \pi'} \left(\prod_{u=1}^{n} \prod_{i,j} p_u(i,j)^{\mathbf{m}_{\pi'(u)}(i,j)} \right)}.$$

Now let's look at the left hand side of Equation 8. We can write

$$P\left(\Pi^{(n)} = \pi \;\middle|\; \mathbf{Y}^{(m)} = \mathbf{y}, \text{Perm}(\mathbf{M}^{(m)}, \Pi^{(n)}) = \mathbf{m}\right) = P\left(\Pi^{(n)} = \pi \;\middle|\; \mathbf{Y}^{(m)} = \mathbf{y}\right).$$

This is because $\text{Perm}(\mathbf{M}^{(m)}, \Pi^{(n)})$ is a function of $\mathbf{Y}^{(m)}$. We have

$$P\left(\Pi^{(n)} = \pi \;\middle|\; \mathbf{Y}^{(m)} = \mathbf{y}\right) = \frac{P\left(\mathbf{Y}^{(m)} = \mathbf{y} \;\middle|\; \Pi^{(n)} = \pi\right) P\left(\Pi^{(n)} = \pi\right)}{P\left(\mathbf{Y}^{(m)} = \mathbf{y}\right)}$$

We have

$$P\left(\mathbf{Y}^{(m)} = \mathbf{y} \;\middle|\; \Pi^{(n)} = \pi\right) = \prod_{u=1}^{n} P\Big(X_u(1) = y_{\pi(u)}(1), X_u(2) = y_{\pi(u)}(2), \cdots, X_u(m) = y_{\pi(u)}(m)\Big)$$
$$= \prod_{u=1}^{n} \prod_{i,j} p_u(i,j)^{\mathbf{m}_{\pi(u)}(i,j)}.$$

Similarly, we obtain

$$P\left(\mathbf{Y}^{(m)} = \mathbf{y}\right) = \frac{1}{n!} \sum_{\text{all permutations } \pi'} \prod_{u=1}^{n} \prod_{i,j} p_u(i,j)^{\mathbf{m}_{\pi'(u)}(i,j)}$$

Thus, we conclude that the left hand side of Equation 8 is equal to

$$\frac{\prod_{u=1}^{n} \prod_{i,j} p_u(i,j)^{\mathbf{m}_{\pi(u)}(i,j)}}{\sum_{\text{all permutations } \pi'} \left(\prod_{u=1}^{n} \prod_{i,j} p_u(i,j)^{\mathbf{m}_{\pi'(u)}(i,j)} \right)},$$

which completes the proof.

□

ACKNOWLEDGMENT

We want to thank the reviewers for their insightful comments. Specifically, the example at the end of Section IV-B was mentioned by one of the reviewers which perfectly points out some specific details of Theorem 1.

REFERENCES

[1] Z. Montazeri, A. Houmansadr, and H. Pishro-Nik, "Defining perfect location privacy using anonymization," in *2016 Annual Conference on Information Science and Systems (CISS)*. IEEE, 2016, pp. 204–209.

[2] Z. Montazeri, A. Houmansadr, and H.Pishro-Nik, "Achieving perfect location privacy in markov models using anonymization," in *2016 International Symposium on Information Theory and its Applications (ISITA2016)*, Monterey, USA, Oct. 2016.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] W. Wang and Q. Zhang, "Privacy-preserving collaborative spectrum sensing with multiple service providers," *IEEE Transactions on Wireless Communications*, vol. 14, no. 2, pp. 1011–1019, 2015.

[8] B. Niu, Q. Li, X. Zhu, G. Cao, and H. Li, "Enhancing privacy through caching in location-based services," in *2015 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2015, pp. 1017–1025.

[9] 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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] W. Wang and Q. Zhang, "Toward long-term quality of protection in mobile networks: a context-aware perspective," *IEEE Wireless Communications*, vol. 22, no. 4, pp. 34–40, 2015.

[15] A. R. Beresford and F. Stajano, "Mix zones: User privacy in location-aware services," 2004.

[16] 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.

[17] G. Ghinita, P. Kalnis, and S. Skiadopoulos, "Prive: anonymous location-based queries in distributed mobile systems," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 371–380.

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

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

[20] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *Proceedings of the 17th annual international conference on Mobile computing and networking*. ACM, 2011, pp. 145–156.

[21] P. Golle and K. Partridge, "On the anonymity of home/work location pairs," in *International Conference on Pervasive Computing*. Springer, 2009, pp. 390–397.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

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

[31] 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.

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

[33] 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.

[34] T. Li and N. Li, "On the tradeoff between privacy and utility in data publishing," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 517–526.

[35] W. Wang and Q. Zhang, "Privacy preservation for context sensing on smartphone."

[36] S. Salamatian, A. Zhang, F. du Pin Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, P. Oliveira, and N. Taft, "How to hide the elephant-or the donkey-in the room: Practical privacy against statistical inference for large data." in *GlobalSIP*, 2013, pp. 269–272.

[37] I. Csiszár, "Almost independence and secrecy capacity," *Problemy Peredachi Informatsii*, vol. 32, no. 1, pp. 48–57, 1996.

[38] F. P. Calmon, A. Makhdoumi, and M. Médard, "Fundamental limits of perfect privacy," in *Information Theory (ISIT), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 1796–1800.

[39] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *Information Forensics and Security, IEEE Transactions on*, vol. 8, no. 6, pp. 838–852, 2013.

[40] H. Yamamoto, "A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers (corresp.)," *IEEE Transactions on Information Theory*, vol. 29, no. 6, pp. 918–923, 1983.

[41] D. Rebollo-Monedero, J. Forne, and J. Domingo-Ferrer, "From t-closeness-like privacy to postrandomization via information theory," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 11, pp. 1623–1636, 2010.

[42] J. Unnikrishnan, "Asymptotically optimal matching of multiple sequences to source distributions and training sequences," *IEEE Transactions on Information Theory*, vol. 61, no. 1, pp. 452–468, 2015.

Zarrin Montazeri received her B.S. degree in Information Technology from Sharif University of Technology, Tehran, Iran, in 2014. She then received her M.Sc. degree in Electrical and Computer Engineering from University of Massachusetts Amherst in 2017. During her research, she worked on Privacy & Security issues with focus on location privacy.

Amir Houmansadr is an Assistant Professor at the College of Information and Computer Sciences at the University of Massachusetts Amherst. He received his PhD from the University of Illinois at Urbana-Champaign in 2012. Amir's area of research is network security and privacy, which includes problems such as Internet censorship resistance, statistical traffic analysis, location privacy, cover communications, and privacy in next-generation network architectures. Amir has received several awards including the Best Practical Paper award at the IEEE Symposium on Security & Privacy (Oakland) in 2013, a Google Faculty Research Award in 2015, and an NSF CAREER Award in 2016.

Hossein Pishro-Nik is an Associate Professor of Electrical and Computer Engineering at the University of Massachusetts, Amherst. He received a B.S. degree from Sharif University of Technology, and M.Sc. and Ph.D. degrees from the Georgia Institute of Technology, all in Electrical and Computer Engineering. His research interests include Information Theoretic Privacy & Security, Error Control Coding, vehicular communications, and mathematical analysis of wireless networks. His awards include an NSF Faculty Early Career Development (CAREER) Award, an Outstanding Junior Faculty Award from UMass, and an Outstanding Graduate Research Award from the Georgia Institute of Technology.