Matching Anonymized and Obfuscated Time Series to Users' Profiles

Nazanin Takbiri, *Student Member, IEEE*, Amir Houmansadr, *Member, IEEE*, Dennis L. Goeckel, *Fellow, IEEE*, and Hossein Pishro-Nik, *Member, IEEE*

***Abstract*— Many popular applications use traces of user data to offer various services to their users. However, even if user data are anonymized and obfuscated, a user's privacy can be compromised through the use of statistical matching techniques that match a user trace to prior user behavior. In this paper, we derive the theoretical bounds on the privacy of users in such a scenario. We build on our recent study in the area of location privacy, in which we introduced formal notions of location privacy for anonymization-based location privacy-protection mechanisms. Here, we derive the fundamental limits of user privacy when both anonymization and obfuscation-based protection mechanisms are applied to users' time series of data. We investigate the impact of such mechanisms on the tradeoff between privacy protection and user utility. We first study achievability results for the case where the time-series of users are governed by an independent and identically distributed (i.i.d.) process. The converse results are proved both for the i.i.d. case as well as the more general Markov chain model. We demonstrate that as the number of users in the network grows, the obfuscation-anonymization plane can be divided into two regions: in the first region, all users have perfect privacy; and, in the second region, no user has privacy.**

***Index Terms*— Anonymization, information theoretic privacy, obfuscation, privacy-protection mechanism (PPM), user-data driven (UDD) services.**

I. Introduction

A NUMBER of emerging systems and applications work by analyzing the data submitted by their users in order to serve them; we call such systems *User-Data Driven* (UDD) services. Examples of UDD services include smart cities, connected vehicles, smart homes, and connected healthcare devices, which have the promise of greatly improving users' lives. Unfortunately, the sheer volume of user data collected by these systems can compromise users' privacy [2]. Even the use of standard Privacy-Protection Mechanisms (PPMs), specifically anonymization of user identities and obfuscation of submitted data, does not guarantee users' privacy, as adversaries are able to use powerful statistical inference techniques to learn sensitive private information of the users [3]–[7].

To illustrate the threat of privacy leakage, consider three popular UDD services: (1) *Health care:* Wearable monitors that constantly track user health variables can be invaluable in assessing individual health trends and responding to emergencies. However, such monitors produce long time-series of user data uniquely matched to the health characteristics of each user; (2) *Smart homes:* Emerging smart-home technologies such as fine-grained power measurement systems can help users and utility providers to address one of the key challenges of the twenty-first century: energy conservation. But the measurements of power by such devices can be mapped to users and reveal their lifestyle habits; and, (3) *Connected vehicles:* The location data provided by connected vehicles promises to greatly improve everyday life by reducing congestion and traffic accidents. However, the matching of such location traces to prior behavior not only allows for user tracking, but also reveals a user's habits. In summary, despite their potential impact on society and their emerging popularity, these UDD services have one thing in common: their utility critically depends on their collection of user data, which puts users' privacy at significant risk.

There are two main approaches to augment privacy in UDD services: *identity perturbation (anonymization)* [8]–[14], and *data perturbation (obfuscation)* [15]–[17]. In anonymization techniques, privacy is obtained by concealing the mapping between users and data, and the mapping is changed periodically to thwart statistical inference attacks that try to de-anonymize the anonymized data traces by matching user data to known user profiles. Some approaches employ k-anonymity to keep each user's identity indistinguishable within a group of $k-1$ other users [18]–[24]. Other approaches employ users' pseudonyms within areas called mix-zones [25]–[27]. Obfuscation mechanisms aim at protecting privacy by perturbing user data, e.g., by adding noise to users' samples of data. For instance, cloaking replaces each user's sample of data with a larger region [28]–[33], while an alternative approach is to use dummy data in the set of possible data of the users [34]–[38]. In [39], a mechanism of obfuscation was introduced where the answer was changed randomly with some small probability. Here we consider the fundamental limits of a similar obfuscation technique for providing privacy in the long time series of emerging applications.

Manuscript received September 7, 2017; revised April 29, 2018; accepted July 17, 2018. Date of publication October 1, 2018; date of current version January 18, 2019. This work was supported by the National Science Foundation under Grants CCF–1421957 and CNS–1739462. This paper was presented in part at the 2017 IEEE International Symposium on Information Theory [1].

N. Takbiri, D. L. Goeckel, and H. Pishro-Nik are with the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003 USA (e-mail: ntakbiri@umass.edu; goeckel@ecs.umass.edu; pishro@engin.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).

Communicated by N. Kiyavash, Associate Editor for Statistical Learning.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2018.2873134

The anonymization and obfuscation mechanisms improve user privacy at the cost of user utility. The anonymization mechanism works by frequently changing the pseudonym mappings of users to reduce the length of time series that can be exploited by statistical analysis. However, this frequent change may also decrease the usability by concealing the temporal relation between a user's sample of data, which may be critical in the utility of some systems, e.g., a dining recommendation system that makes suggestions based on the dining history of its users. On the other hand, obfuscation mechanisms work by adding noise to users' collected data, e.g., location information. The added noise may degrade the utility of UDD applications. Thus, choosing the right level of the privacy-protection mechanism is an important question, and understanding what levels of anonymization and obfuscation can provide theoretical guarantees of privacy is of interest.

In this paper, we will consider the ability of an adversary to perform statistical analyses on time series and match the series to descriptions of user behavior. In related work, Unnikrishnan [7] provides a comprehensive analysis of the asymptotic (in the length of the time series) optimal matching of time series to source distributions. However, there are several key differences between that analysis and the work here. First, Unnikrishnan [7] looks at the optimal matching tests, but does not consider the privacy metrics in this paper. A significant component of our study is demonstrating that mutual information converges to zero so that we can conclude there is no privacy leakage (hence, "perfect privacy"). Second, the setting of Unnikrishnan [7] is different in two key aspects: (a) It does not consider the obfuscation, which is one of the two major protection mechanisms studied here; (b) Unnikrishnan [7] only focuses on sources that are independent and identically distributed (i.i.d.) while here, models based on Markov chains are also considered. Third, the setting of Unnikrishnan [7] assumes a fixed distribution on sources (i.e., classical inference), whereas we assume the existence of general (but possibly unknown) prior distributions for the sources (i.e., a Bayesian setting). Finally, we study the fundamental limits in terms of both the number of users and the number of observations, while Unnikrishnan [7] focuses on the case where the number of users is a fixed, finite value.

Numerous researchers have put forward ideas for quantifying privacy-protection. Shokri *et al.* [12], [40] define the expected estimation error of the adversary as a metric to evaluate PPMs. Ma *et al.* [11] use uncertainty about users' information to quantify user privacy in vehicular networks. To defeat localization attacks and achieve privacy at the same time, Shokri *et al.* [15] proposed a method which finds optimal PPM for a Location Based Service (LBS) given service quality constraints. In [41] and [42], privacy leakage of data sharing and interdependent privacy risks are quantified, respectively. A similar idea is proposed in [43] where the quantification model is based on the Bayes conditional risk. Kalantari *et al.* [44] derived the exact information theoretic privacy-utility tradeoff for finite blocklengths of data.

Previously, mutual information has been used as a privacy metric in a number of settings [45]–[51]. However, the framework and problem formulation for our setting (Internet of Things (IoT) privacy) are quite different from those encountered in previous works. More specifically, the IoT privacy problem we consider here is based on a large set of time-series data that belongs to different users with different statistical patterns that has gone through a privacy-preserving mechanism, and the adversary is aiming at de-anonymizing and de-obfuscating the data.

The discussed studies demonstrate the growing importance of privacy. What is missing from the current literature is a solid theoretical framework for privacy that is general enough to encompass various privacy-preserving methods in the literature. Such a framework will allow us to achieve provable privacy guarantees, obtain fundamental trade-offs between privacy and performance, and provide analytical tools to optimally achieve provable privacy. We derive the fundamental limits of user privacy in UDD services in the presence of both anonymization and obfuscation protection mechanisms. We build on our previous works on formalizing privacy in location-based services [52], [53], but we significantly expand those works here not just in application area but also user models and settings. In particular, our previous works introduced the notion of *perfect privacy* for location-based services, and we derived the rate at which an anonymization mechanism should change the pseudonyms in order to achieve the defined perfect privacy. In this work, we expand the notion of perfect privacy to UDD services in general and derive the conditions for it to hold when *both* anonymization and obfuscation-based protection mechanisms are employed.

In this paper, we consider two models for users' data: i.i.d. and Markov chains. After introducing the general framework in Section II, we consider an i.i.d. model extensively in Section III and the first half of Section IV. We obtain achievability and converse results for the i.i.d. model. The i.i.d. model would apply directly to data that is sampled at a low rate. In addition, understanding the i.i.d. case can also be considered the first step toward understanding the more complicated case where there is dependency, as was done for anonymization-only Location Privacy-Preserving Mechanisms (LPPMs) in [52], and will be done in Section IV-C. In particular, in Section IV-C, a general Markov chain model is used to model users' data pattern to capture the dependency of the user' data pattern over time. There, we obtain converse results for privacy for this model. In Section V, we provide some discussion about the achievability for the Markov chain case.

A. Summary of the Results

Given n, the total number of the users in a network, their degree of privacy depends on two parameters: (1) The number of observations $m = m(n)$ by the adversary per user for a fixed anonymization mapping (i.e., the number of observations before the pseudonyms are changed); and (2) the value of the noise added by the obfuscation technique (as defined in Section II, we quantify the obfuscation noise with a parameter a_n, where larger a_n means a higher level of obfuscation). Intuitively, smaller $m(n)$ and larger a_n result in stronger privacy, at the expense of lower utility for the users.

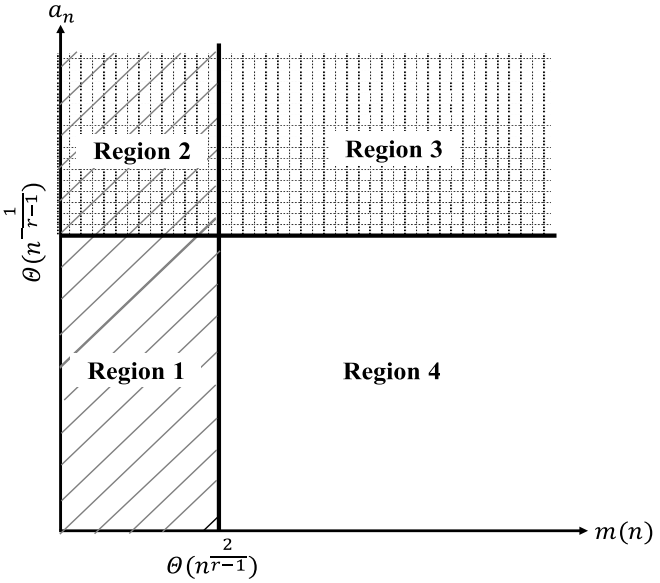

Fig. 1. Limits of privacy in the entire $m(n) - a_n$ plane: in regions 1, 2, and 3, users have perfect privacy, and in region 4 users have no privacy.

Our goal is to identify values of a_n and $m(n)$ that satisfy perfect privacy in the asymptote of a large number of users ($n \to \infty$). When the users' datasets are governed by an i.i.d. process, we show that the $m(n) - a_n$ plane can be divided into two areas. In the first area, all users have perfect privacy (as defined in Section II), and, in the second area, users have no privacy. Figure 1 shows the limits of privacy in the entire $m(n) - a_n$ plane. As the figure shows, in regions 1, 2, and 3, users have perfect privacy, while in region 4 users have no privacy.

For the case where the users' datasets are governed by irreducible and aperiodic Markov chains with r states and $|E|$ edges, we show that users will have no privacy if $m = cn^{\frac{2}{|E|-r}+\alpha}$ and $a_n = c'n^{-\left(\frac{1}{|E|-r}+\beta\right)}$, for any constants $c > 0$, $c' > 0$, $\alpha > 0$, and $\beta > \frac{\alpha}{4}$. We also provide some insights for the opposite direction (under which conditions users have perfect privacy) for the case of Markov chains.

II. Framework

In this paper, we adopt a similar framework to that employed in [52] and [53]. The general set up is provided here, and the refinement to the precise models for this paper will be presented in the following sections. We assume a system with n users with $X_u(k)$ denoting a sample of the data of user u at time k, which we would like to protect from an interested adversary. We consider a strong adversary that has complete statistical knowledge of the users' data patterns based on the previous observations or other resources. In order to secure data privacy of users, both obfuscation and anonymization techniques are used as shown in Figure 2. In Figure 2, $Z_u(k)$ shows the (reported) sample of the data of user u at time k after applying obfuscation, and $Y_u(k)$ shows the (reported) sample of the data of user u at time k after applying anonymization. The adversary observes only $Y_u(k)$,

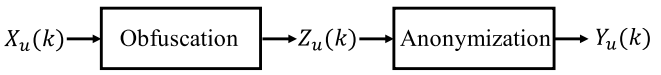

Fig. 2. Applying obfuscation and anonymization techniques to users' data samples.

$k = 1, 2, \cdots, m(n)$, where $m(n)$ is the number of observations of each user before the identities are permuted. The adversary then tries to estimate $X_u(k)$ by using those observations.

Let $\mathbf{X}_u$ be the $m(n) \times 1$ vector containing the sample of the data of user u, and $\mathbf{X}$ be the $m(n) \times n$ matrix with u^{th} column equal to $\mathbf{X}_u$;

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

A. Data Samples Model

We assume there are $r \geq 2$ possible values $(0, 1, \cdots, r-1)$ for each sample of the users' data. In the first part of the paper (perfect privacy analysis), we assume an i.i.d. model as motivated in Section I. In the second part of the paper (converse results: no privacy region), the users' datasets are governed by irreducible and aperiodic Markov chains. At any time, $X_u(k)$ is equal to a value in $\{0, 1, \cdots, r-1\}$ according to a user-specific probability distribution $(\mathbf{p}_u)$. $p_u(i)$ is the probability of user u having the data value i, so

$$\mathbf{p}_u = \begin{bmatrix} p_u(1) \\ p_u(2) \\ \vdots \\ p_u(r-1) \end{bmatrix}, \quad \text{for each } u \in \{1, 2, \cdots, n\}.$$

We also assume $\mathbf{p}_u$'s are drawn independently from some continuous density function, $f_{\mathbf{P}}(\mathbf{p}_u)$, which has support on a subset of the $(0, 1)^{r-1}$ hypercube.

All $\mathbf{p}_u$'s are know to the adversary, and he/she employs such to distinguish different users based on statistical matching of those user distributions to traces of user activity of length $m(n)$.

B. Obfuscation Model

The first step in obtaining privacy is to apply the obfuscation operation in order to perturb the users' data samples. In this paper, we assume that each user has only limited knowledge of the characteristics of the overall population and thus we employ a simple distributed method in which the samples of the data of each user are reported with error with a certain probability, where that probability itself is generated randomly for each user. In other words, the obfuscated data is obtained by passing the users' data through an r-ary symmetric channel with a random error probability. More precisely, let $\mathbf{Z}_u$ be the vector which contains the obfuscated versions of user u's data samples, and $\mathbf{Z}$ is the collection of $\mathbf{Z}_u$ for all users,

$$\mathbf{Z}_u = \begin{bmatrix} Z_u(1) \\ Z_u(2) \\ \vdots \\ Z_u(m) \end{bmatrix}, \quad \mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n].$$

To create a noisy version of data samples, for each user u, we independently generate a random variable R_u that is uniformly distributed between 0 and a_n, where $a_n \in (0, 1]$. The value of R_u gives the probability that a user's data sample is changed to a different data sample by obfuscation, and a_n is termed the "noise level" of the system. For the case of $r = 2$ where there are two states for users' data (state 0 and state 1), the obfuscated data is obtained by passing users' data through a Binary Symmetric Channel (BSC) with a small error probability [39]. Thus, we can write

$$Z_u(k) = \begin{cases} X_u(k), & \text{with probability } 1 - R_u. \\ 1 - X_u(k), & \text{with probability } R_u. \end{cases}$$

When $r > 2$, for $l \in \{0, 1, \cdots, r-1\}$:

$$P(Z_u(k) = l | X_u(k) = i) = \begin{cases} 1 - R_u, & \text{for } l = i. \\ \dfrac{R_u}{r-1}, & \text{for } l \neq i. \end{cases}$$

Note that the effect of the obfuscation is to alter the probability distribution function of each user across the r possibilities in a way that is unknown to the adversary, since it is independent of all past activity of the user, and hence the obfuscation inhibits user identification. For each user, R_u is generated once and is kept constant for the collection of samples of length $m(n)$, thus, providing a very low-weight obfuscation algorithm. We will discuss the extension to the case where R_u is regenerated independently over time in Section V. There, we will also provide a discussion about obfuscation using continuous noise distributions (e.g., Gaussian noise).

C. Anonymization Model

Anonymization is modeled by a random permutation Π on the set of n users. The user u is assigned the pseudonym $\Pi(u)$. $\mathbf{Y}$ is the anonymized version of $\mathbf{Z}$; thus,

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

where Perm(., Π) is permutation operation with permutation function Π. As a result, $\mathbf{Y}_u = \mathbf{Z}_{\Pi^{-1}(u)}$ and $\mathbf{Y}_{\Pi(u)} = \mathbf{Z}_u$.

D. Adversary Model

We protect against the strongest reasonable adversary. Through past observations or some other sources, the adversary is assumed to have complete statistical knowledge of the users' patterns; in other words, he/she knows the probability distribution for each user on the set of data samples $\{0, 1, \ldots, r-1\}$, shown by vector $\mathbf{p}_u = [p_u(1), p_u(2), \cdots, p_u(m)]^T$. As discussed in the model for the data samples, the parameters $\mathbf{p}_u$, $u = 1, 2, \cdots, n$ are drawn independently from a continuous density function, $f_{\mathbf{P}}(\mathbf{p}_u)$, which has support on a subset of a defined hypercube. The density $f_{\mathbf{P}}(\mathbf{p}_u)$ might be unknown to the adversary, as all that is assumed here is that such a density exists, and it will be evident from our results that knowing or not knowing $f_{\mathbf{P}}(\mathbf{p}_u)$ does not change the results asymptotically. Specifically, from the results of Section III, we conclude that user u has perfect privacy even if the adversary knows $f_{\mathbf{P}}(\mathbf{p}_u)$. In addition, in Section IV, it is shown that the adversary can recover the true data of user u at time k without using the specific density function of $f_{\mathbf{P}}(\mathbf{p}_u)$, and as result, users have no privacy even if the adversary does not know $f_{\mathbf{P}}(\mathbf{p}_u)$.

The adversary also knows the value of a_n as it is a design parameter. However, the adversary does not know the realization of the random permutation Π or the realizations of the random variables R_u, as these are independent of the past behavior of the users. It is critical to note that we assume the adversary does not have any other auxiliary information or side information about users' data.

In [52], perfect privacy is defined as follows:

Definition 1: User u has *perfect privacy* at time k, if and only if

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

where $I(X; Y)$ denotes the mutual information between random variables (vectors) X and Y.

In this paper, we also consider the situation in which there is no privacy.

Definition 2: For an algorithm for the adversary that tries to estimate the actual sample of data of user u at time k, define the error probability as

$$P_e(u, k) = P\left(\widetilde{X_u(k)} \neq X_u(k)\right),$$

where $\underbrace{X_u(k)}$ is the actual sample of the data of user u at time k, $\widetilde{X_u(k)}$ is the adversary's estimated sample of the data of user u at time k. Now, define $\mathcal{E}$ as the set of all possible adversary's estimators. Then, user u has *no privacy* at time k, if and only if for large enough n,

$$P_e^*(u, k) = \inf_{\mathcal{E}} P\left(\widetilde{X_u(k)} \neq X_u(k)\right) \to 0.$$

Hence, a user has no privacy if there exists an algorithm for the adversary to estimate $X_u(k)$ with diminishing error probability as n goes to infinity.

Discussion: Both of the privacy definitions given above (perfect privacy and no privacy) are asymptotic in the number of users ($n \to \infty$), which allows us to find clean analytical results for the fundamental limits. Moreover, in many IoT applications, such as ride sharing and dining recommendation applications, the number of users is large.

Notation: Note that the sample of data of user u at time k after applying obfuscation ($Z_u(k)$) and the sample of data of user u at time k after applying anonymization ($Y_u(k)$) depend on the number of users in the network (n), while the actual sample of data of user u at time k is independent of the number of users (n). Despite the dependency in the former cases, we omit this subscript (n) on $\left(Z_u^{(n)}(k), Y_u^{(n)}(k)\right)$ to avoid confusion and make the notation consistent.

Notation: Throughout the paper, $X_n \xrightarrow{d} X$ denotes convergence in distribution. Also, We use $P\left(X = x \middle| Y = y\right)$ for the conditional probability of $X = x$ given $Y = y$. When we write $P\left(X = x \middle| Y\right)$, we are referring to a random variable that is defined as a function of Y.

III. Perfect Privacy Analysis: I.I.D. Case

A. Two-State Model

We first consider the two-state case ($r = 2$) which captures the salient aspects of the problem. For the two-state case, the sample of the data of user u at any time is a Bernoulli random variable with parameter p_u, which is the probability of user u having data sample 1. Thus,

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

Per Section II, the parameters p_u, $u = 1, 2, \cdots, n$ are drawn independently from a continuous density function, $f_P(p_u)$, on the $(0, 1)$ interval. We assume there are $\delta_1, \delta_2 > 0$ such that:[1]

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

The adversary knows the values of p_u, $u = 1, 2, \cdots, n$ and uses this knowledge to identify users. We will use capital letters (i.e., P_u) when we are referring to the random variable, and use lower case (i.e., p_u) to refer to the realization of P_u.

In addition, since the user data ($X_u(k)$) are i.i.d. and have a Bernoulli distribution, the obfuscated data ($Z_u(k)$) are also i.i.d. with a Bernoulli distribution. Specifically,

$$Z_u(k) \sim \text{Bernoulli}\,(Q_u),$$

where

$$\begin{aligned} Q_u &= P_u(1 - R_u) + (1 - P_u)R_u \\ &= P_u + (1 - 2P_u)\,R_u, \end{aligned}$$

and recall that R_u is the probability that user u's data sample is altered at any time. For convenience, define a vector where element Q_u is the probability that an obfuscated data sample of user u is equal to one, and

$$\mathbf{Q} = [Q_1, Q_2, \cdots, Q_n].$$

Thus, a vector containing the permutation of those probabilities after anonymization is given by:

$$\begin{aligned} \mathbf{V} &= \text{Perm}\,(Q_1, Q_2, \cdots, Q_n; \Pi) \\ &= \left[Q_{\Pi^{-1}(1)}, Q_{\Pi^{-1}(2)}, \cdots, Q_{\Pi^{-1}(n)}\right] \\ &= [V_1, V_2, \cdots, V_n], \end{aligned}$$

where $V_u = Q_{\Pi^{-1}(u)}$ and $V_{\Pi(u)} = Q_u$. As a result, for $u = 1, 2, \cdots, n$, the distribution of the data symbols for the user with pseudonym u is given by:

$$Y_u(k) \sim \text{Bernoulli}\,(V_u) \sim \text{Bernoulli}\left(Q_{\Pi^{-1}(u)}\right).$$

The following theorem states that if a_n is significantly larger than $\frac{1}{n}$ in this two-state model, then all users have perfect privacy independent of the value of $m(n)$.

Theorem 1: For the above two-state model, if $\mathbf{Z}$ is the obfuscated version of $\mathbf{X}$, and $\mathbf{Y}$ is the anonymized version of $\mathbf{Z}$ as defined above, and

- $m = m(n)$ is arbitrary;
- $R_u \sim \text{Uniform}[0, a_n]$, where $a_n = c' n^{-(1-\beta)}$ for any $c' > 0$ and $0 < \beta < 1$;

then, user 1 has perfect privacy at time k as n goes to infinity.

[1]The condition $\delta_1 < f_P(p_u) < \delta_2$ is not actually necessary for the results and can be relaxed; however, we keep it here to avoid unnecessary technicalities.

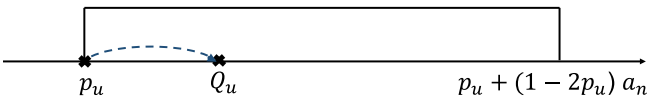

Fig. 3. Distribution of Q_u given $P_u = p_u$.

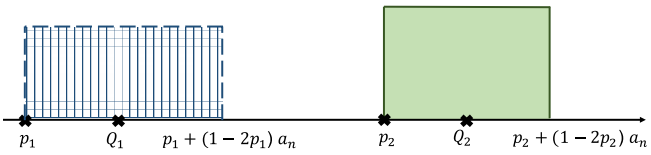

Fig. 4. Case 1: The support of the distributions is small relative to the difference between p_1 and p_2.

The proof of Theorem 1 will be provided for the case $0 \le p_1 < \frac{1}{2}$, as the proof for the case $\frac{1}{2} \le p_1 \le 1$ is analogous and is thus omitted.

Intuition behind the Proof of Theorem 1: Since $m(n)$ is arbitrary, the adversary is able to estimate very accurately (in the limit, perfectly) the distribution from which each data sequence $\mathbf{Y}_u$, $u = 1, 2, \cdots, n$ is drawn; that is, the adversary is able to accurately estimate the probability V_u, $u = 1, 2, \cdots, n$. Clearly, if there were no obfuscation for each user u, the adversary would then simply look for the j such that p_j is very close to V_u and set $\widehat{X_j(k)} = Y_u(k)$, resulting in no privacy for any user.

We want to make certain that the adversary obtains no information about $X_1(k)$, the sample of data of user 1 at time k. To do such, we will establish that there are a large number of users whom have a probability p_u that when obfuscated could have resulted in a probability consistent with p_1. Consider asking whether another probability p_2 is sufficiently close enough to be confused with p_1 after obfuscation; in particular, we will look for p_2 such that, even if the adversary is given the obfuscated probabilities $V_{\Pi(1)}$ and $V_{\Pi(2)}$, he/she cannot associate these probabilities with p_1 and p_2. This requires that the distributions Q_1 and Q_2 of the obfuscated data of user 1 and user 2 have significant overlap; we explore this next.

Recall that $Q_u = P_u + (1 - 2P_u)R_u$, and $R_u \sim \text{Uniform}[0, a_n]$. Thus, we know $Q_u \big| P_u = p_u$ has a uniform distribution with length $(1 - 2p_u)a_n$. Specifically,

$$Q_u \big| P_u = p_u \sim \text{Uniform}\,[p_u, p_u + (1 - 2p_u)a_n].$$

Figure 3 shows the distribution of Q_u given $P_u = p_u$.

Consider two cases: In the first case, the support of the distributions $Q_1 \big| P_1 = p_1$ and $Q_2 \big| P_2 = p_2$ are small relative to the difference between p_1 and p_2 (Figure 4); in this case, given the probabilities $V_{\Pi(1)}$ and $V_{\Pi(2)}$ of the anonymized data sequences, the adversary can associate those with p_1 and p_2 without error. In the second case, the support of the distributions $Q_1 \big| P_1 = p_1$ and $Q_2 \big| P_2 = p_2$ is large relative to the difference between p_1 and p_2 (Figure 5), so it is difficult for the adversary to associate the probabilities $V_{\Pi(1)}$ and $V_{\Pi(2)}$ of the anonymized data sequences with p_1 and p_2. In particular, if $V_{\Pi(1)}$ and $V_{\Pi(2)}$ fall into the overlap of the

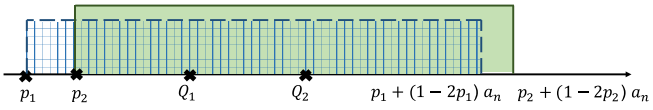

Fig. 5. Case 2: The support of the distributions is large relative to the difference between p_1 and p_2.

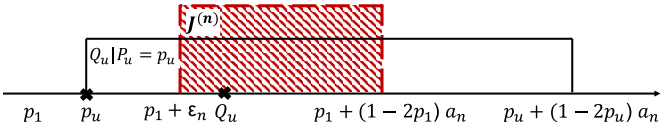

Fig. 6. Range of P_u and Q_u for elements of set $J^{(n)}$ and probability density function of $Q_u|P_u = p_u$.

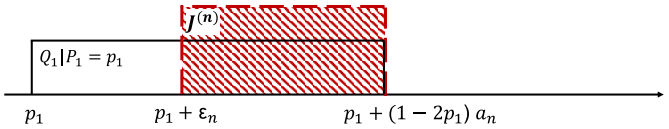

Fig. 7. Range of P_u and Q_u for elements of set $J^{(n)}$ and probability density function of $Q_1|P_1 = p_1$.

support of Q_1 and Q_2, we will show the adversary can only guess randomly how to de-anonymize the data. Thus, if the ratio of the support of the distributions to $|p_1 - p_2|$ goes to infinity, the adversary's posterior probability for each user converges to $\frac{1}{2}$, thus, implying no information leakage on the user identities. More generally, if we can guarantee that there will be a large set of users with p_u's very close to p_1 compared to the support of $Q_1|P_1 = p_1$, we will be able to obtain perfect privacy as demonstrated rigorously below.

Given this intuition, the formal proof proceeds as follows. Given p_1, we define a set $J^{(n)}$ of users whose parameter p_u of their data distributions is sufficiently close to p_1 (Figure 5; case 2), so that it is likely that Q_1 and Q_u cannot be readily associated with p_1 and p_u.

The purpose of Lemmas 1, 2, and 3 is to show that, from the adversary's perspective, the users in set $J^{(n)}$ are indistinguishable. More specifically, the goal is to show that the obfuscated data corresponding to each of these users could have been generated by any other users in $J^{(n)}$ in an equally likely manner. To show this, Lemma 1 employs the fact that, if the observed values of N uniformly distributed random variables (N is size of set $J^{(n)}$) are within the intersection of their ranges, it is impossible to infer any information about the matching between the observed values and the distributions. That is, all possible $N!$ matchings are equally likely. Lemmas 2 and 3 leverage Lemma 1 to show that even if the adversary is given a set that includes all of the pseudonyms of the users in set $J^{(n)}$ $\left(\text{i.e., } \Pi(J^{(n)}) \triangleq \left\{\Pi^{-1}(u) \in J^{(n)}\right\}\right)$ he/she still will not be able to infer any information about the matching of each specific user in set $J^{(n)}$ and his pseudonym. Then Lemma 5 uses the above fact to show that the mutual information between the data set of user 1 at time k and the observed data sets of the adversary converges to zero for large enough n.

Proof of Theorem 1:

Proof: Note, per Lemma 6 of Appendix A, it is sufficient to establish the results on a sequence of sets with high probability. That is, we can condition on high-probability events.

Now, define the critical set $J^{(n)}$ with size $N^{(n)} = |J^{(n)}|$ for $0 \le p_1 < \frac{1}{2}$ as follows:

$$J^{(n)} = \{u \in \{1, 2, \ldots, n\} : p_1 \le P_u \le p_1 + \epsilon_n; \\ p_1 + \epsilon_n \le Q_u \le p_1 + (1-2p_1)a_n\},$$

where $\epsilon_n = n^{-\left(1-\frac{\beta}{2}\right)}$, $a_n = c'n^{-(1-\beta)}$, and β is defined in the statement of Theorem 1.

Note for large enough n, if $0 \le p_1 < \frac{1}{2}$, we have $0 \le p_u < \frac{1}{2}$. As a result,

$$Q_u|P_u = p_u \sim \text{Uniform}\,(p_u, p_u + (1-2p_u)a_n).$$

We can prove that with high probability, $1 \in J^{(n)}$ for large enough n, as follows. First, Note that

$$Q_1|P_1 = p_1 \sim \text{Uniform}\,(p_1, p_1 + (1-2p_1)a_n).$$

Now, according to Figure 7,

$$P\left(1 \in J^{(n)}\right) = 1 - \frac{\epsilon_n}{(1-2p_1)\,a_n} \\ = 1 - \frac{1}{(1-2p_1)\,c'n^{\frac{\beta}{2}}},$$

thus, for any $c' > 0$ and large enough n,

$$P\left(1 \in J^{(n)}\right) \to 1.$$

Now in the second step, we define the probability $W_j^{(n)}$ for any $j \in \Pi(J^{(n)}) = \{\Pi(u) : u \in J^{(n)}\}$ as

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

$W_j^{(n)}$ is the conditional probability that $\Pi(1) = j$ after perfectly observing the values of the permuted version of obfuscated probabilities ($\mathbf{V}$) and set including all of the pseudonyms of the users in set $J^{(n)}$ $\left(\Pi(J^{(n)})\right)$. Since $\mathbf{V}$ and $\Pi(J^{(n)})$ are random, $W_j^{(n)}$ is a random variable. However, we will prove shortly that in fact $W_j^{(n)} = \frac{1}{N^{(n)}}$, for all $j \in \Pi(J^{(n)})$.

Note: Since we are looking from the adversary's point of view, the assumption is that all the values of P_u, $u \in \{1, 2, \cdots, n\}$ are known, so all of the probabilities are conditioned on the values of $P_1 = p_1, P_2 = p_2, \cdots, P_n = p_n$. Thus, to be accurate, we should write

$$W_j^{(n)} = P\left(\Pi(1) = j|\mathbf{V}, \Pi(J^{(n)}), P_1, P_2, \cdots, P_n\right).$$

Nevertheless, for simplicity of notation, we often omit the conditioning on $P_1, P_2, \cdots, P_n$.

First, we need a lemma from elementary probability.

Lemma 1: Let N be a positive integer, and let $a_1, a_2, \cdots, a_N$ and $b_1, b_2, \cdots, b_N$ be real numbers such that $a_u \le b_u$ for all u. Assume that $X_1, X_2, \cdots, X_N$ are independent random variables such that

$$X_u \sim \text{Uniform}[a_u, b_u].$$

Let also $\gamma_1, \gamma_2, \cdots, \gamma_N$ be distinct real numbers such that

$$\gamma_j \in \bigcap_{u=1}^{N} [a_u, b_u] \quad \text{for all } j \in \{1, 2, .., N\}.$$

Suppose that we know the event E has occurred, meaning that the observed values of X_u's are equal to the set of γ_j's (but with unknown ordering), i.e.,

$$E \equiv \{X_1, X_2, \cdots, X_N\} = \{\gamma_1, \gamma_2, \cdots, \gamma_N\},$$

then

$$P\left(X_1 = \gamma_j | E\right) = \frac{1}{N}.$$

Proof: Lemma 1 is proved in Appendix B. □

Using the above lemma, we can state our desired result for $W_j^{(n)}$.

Lemma 2: For all $j \in \Pi(J^{(n)})$, $W_j^{(n)} = \frac{1}{N^{(n)}}$.

Proof: We argue that the setting of this lemma is essentially equivalent to the assumptions in Lemma 1. First, remember that

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

Note that $Q_u = P_u + (1-2P_u)R_u$, and since R_u is uniformly distributed, Q_u conditioned on P_u is also uniformly distributed in the appropriate intervals. Moreover, since $V_u = Q_{\Pi^{-1}(u)}$, we conclude V_u is also uniformly distributed. So, looking at the definition of $W_j^{(n)}$, we can say the following: given the values of the uniformly distributed random variables Q_u, we would like to know which one of the values in $\mathbf{V}$ is the actual value of $Q_1 = V_{\Pi(1)}$, i.e., is $\Pi(1) = j$? This is equivalent to the setting of Lemma 1 as described further below.

Note that since $1 \in J^{(n)}$, $\Pi(1) \in \Pi(J^{(n)})$. Therefore, when searching for the value of $\Pi(1)$, it is sufficient to look inside set $\Pi(J^{(n)})$. Therefore, instead of looking among all the values of V_j, it is sufficient to look at V_j for $j \in \Pi(J^{(n)})$. Let us show these values by $\mathbf{V}_\Pi = \{v_1, v_2, \cdots, v_{N^{(n)}}\}$, so,

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

Thus, we have the following scenario: $Q_u, u \in J^{(n)}$ are independent random variables, and

$$Q_u \big| P_u = p_u \sim \text{Uniform}[p_u, p_u + (1-2p_u)a_n].$$

Also, $v_1, v_2, \cdots, v_{N^{(n)}}$ are the observed values of Q_u with unknown ordering (unknown mapping Π). We also know from the definition of set $J^{(n)}$ that

$$\begin{aligned} P_u &\le p_1 + \epsilon_n \le Q_u, \\ Q_u &\le p_1(1-2a_n) + a_n \le P_u(1-2a_n) + a_n, \end{aligned}$$

so, we can conclude

$$v_j \in \bigcap_{u=1}^{N^{(n)}} [p_u, p_u + (1-2p_u)a_n] \quad \text{for all } j \in \{1, 2, .., N^{(n)}\}.$$

We know the event E has occurred, meaning that the observed values of Q_u's are equal to set of v_j's (but with unknown ordering), i.e.,

$$E \equiv \{Q_u, u \in J^{(n)}\} = \{v_1, v_2, \cdots, v_{N^{(n)}}\}.$$

Then, according to Lemma 1,

$$P\left(Q_1 = v_j | E, P_1, P_2, \cdots, P_n\right) = \frac{1}{N^{(n)}}.$$

Note that there is a subtle difference between this lemma and Lemma 1. Here $N^{(n)}$ is a random variable while N is a fixed number in Lemma 1. Nevertheless, since the assertion holds for every fixed N, it also holds for the case where N is a random variable. Now, note that

$$\begin{aligned} &P(Q_1 = v_j | E, P_1, P_2, \cdots, P_n) \\ &\quad = P\left(\Pi(1) = j \middle| E, P_1, P_2, \cdots, P_n\right) \\ &\quad = P\left(\Pi(1) = j \middle| \mathbf{V}_\Pi, \Pi(J^{(n)}), P_1, P_2, \cdots, P_n\right) \\ &\quad = W_j^{(n)}. \end{aligned}$$

Thus, we can conclude

$$W_j^{(n)} = \frac{1}{N^{(n)}}.$$

□

In the third step, we define $\widetilde{W_j^{(n)}}$ for any $j \in \Pi(J^{(n)})$ as

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

$\widetilde{W_j^{(n)}}$ is the conditional probability that $\Pi(1) = j$ after observing the values of the anonymized version of the obfuscated samples of the users' data ($\mathbf{Y}$) and the aggregate set including all the pseudonyms of the users in set $J^{(n)}$ (i.e., $\Pi(J^{(n)}) \stackrel{\Delta}{=} \left\{\Pi^{-1}(j) \in J^{(n)}\right\}$). Since $\mathbf{Y}$ and $\Pi(J^{(n)})$ are random, $\widetilde{W_j^{(n)}}$ is a random variable. Now, in the following lemma, we will prove $\widetilde{W_j^{(n)}} = \frac{1}{N^{(n)}}$, for all $j \in \Pi(J^{(n)})$ by using Lemma 3.

Note in the following lemma, we want to show that even if the adversary is given a set including all of the pseudonyms of the users in set $J^{(n)}$, he/she cannot match each specific user in set $J^{(n)}$ and his pseudonym.

Lemma 3: For all $j \in \Pi(J^{(n)})$, $\widetilde{W_j^{(n)}} = \frac{1}{N^{(n)}}$.

Proof: First, note that

$$\begin{aligned} \widetilde{W_j^{(n)}} = \sum_{\text{for all } \mathbf{v}} &P\left(\Pi(1) = j \middle| \mathbf{Y}, \Pi(J^{(n)}), \mathbf{V} = \mathbf{v}\right) \\ &\cdot P\left(\mathbf{V} = \mathbf{v} \middle| \mathbf{Y}, \Pi(J^{(n)})\right). \end{aligned}$$

Also, we note that given $\mathbf{V}$, $\Pi(J^{(n)})$, and $\mathbf{Y}$ are independent. Intuitively, this is because when observing $\mathbf{Y}$, any information regarding $\Pi(J^{(n)})$ is leaked through estimating $\mathbf{V}$. This can be rigorously proved similar to the proof of [52, Lemma 1]. We can state this fact as

$$P\left(Y_u(k) \middle| V_u = v_u, \Pi(J^{(n)})\right) = P\left(Y_u(k) \middle| V_u = v_u\right) = v_u.$$

The right and left hand side are given by Bernoulli(v_u) distributions.

As a result,

$$\widetilde{W_j^{(n)}} = \sum_{\text{for all } \mathbf{v}} P\left(\Pi(1) = j \middle| \Pi(J^{(n)}), \mathbf{V} = \mathbf{v}\right) \times P\left(\mathbf{V} = \mathbf{v} \middle| \mathbf{Y}, \Pi(J^{(n)})\right).$$

Note $W_j^{(n)} = P\left(\Pi(1) = j \middle| \Pi(J^{(n)}), \mathbf{V}\right)$, so

$$\begin{aligned}\widetilde{W_j^{(n)}} &= \sum_{\text{for all } \mathbf{v}} W_j^{(n)} P\left(\mathbf{V} = \mathbf{v} \middle| \mathbf{Y}, \Pi(J^{(n)})\right) \\ &= \frac{1}{N^{(n)}} \sum_{\text{for all } \mathbf{v}} P\left(\mathbf{V} = \mathbf{v} \middle| \mathbf{Y}, \Pi(J^{(n)})\right) \\ &= \frac{1}{N^{(n)}}.\end{aligned}$$

□

To show that no information is leaked, we need to show that the size of set $J^{(n)}$ goes to infinity. This is established in Lemma 4.

Lemma 4: If $N^{(n)} = |J^{(n)}|$, then $N^{(n)} \to \infty$ with high probability as $n \to \infty$. More specifically, there exists $\lambda > 0$ such that

$$P\left(N^{(n)} > \frac{\lambda}{2} n^{\frac{\beta}{2}}\right) \to 1.$$

Proof: Lemma 4 is proved in Appendix C. □

In the final step, we define $\widehat{W_j^{(n)}}$ for any $j \in \Pi(J^{(n)})$ as

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

$\widehat{W_j^{(n)}}$ is the conditional probability that $X_1(k) = 1$ after observing the values of the anonymized version of the obfuscated samples of the users' data ($\mathbf{Y}$) and the aggregate set including all of the pseudonyms of the users in set $J^{(n)}$ ($\Pi(J^{(n)})$). $\widehat{W_j^{(n)}}$ is a random variable because $\mathbf{Y}$ and $\Pi(J^{(n)})$ are random. Now, in the following lemma, we will prove $\widehat{W_j^{(n)}}$ converges in distribution to p_1.

Note that this is the probability from the adversary's point of view. That is, given that the adversary has observed $\mathbf{Y}$ as well as the extra information $\Pi(J^{(n)})$, what can he/she infer about $X_1(k)$?

Lemma 5: For all $j \in \Pi(J^{(n)})$, $\widehat{W_j^{(n)}} \xrightarrow{d} p_1$.

Proof: We know

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

and according to the definition $\widetilde{W_j^{(n)}} = P\left(\Pi(1) = j \middle| \mathbf{Y}, \Pi(J^{(n)})\right)$, we have

$$\begin{aligned}\widehat{W_j^{(n)}} &= \sum_{j \in \Pi(J^{(n)})} P\left(X_1(k) = 1 \middle| \Pi(1) = j, \mathbf{Y}, \Pi(J^{(n)})\right) \widetilde{W_j^{(n)}} \\ &= \frac{1}{N^{(n)}} \sum_{j \in \Pi(J^{(n)})} P\left(X_1(k) = 1 \middle| \Pi(1) = j, \mathbf{Y}, \Pi(J^{(n)})\right).\end{aligned}$$

We now claim that

$$P\left(X_1(k) = 1 \middle| \Pi(1) = j, \mathbf{Y}, \Pi(J^{(n)})\right) = p_1 + o(1).$$

The reasoning goes as follows. Given $\Pi(1) = j$ and knowing $\mathbf{Y}$, we know that

$$Y_{\Pi(1)}(k) = Z_1(k) = \begin{cases} X_1(k), & \text{with probability } 1 - R_1. \\ 1 - X_1(k), & \text{with probability } R_1. \end{cases}$$

Thus, given $Y_j(k) = 1$, Bayes' rule yields:

$$\begin{aligned}&P\left(X_1(k) = 1 \middle| \Pi(1) = j, \mathbf{Y}, \Pi(J^{(n)})\right) \\ &= (1 - R_1) \frac{P(X_1(k) = 1)}{P(Y_{\Pi(1)}(k) = 1)} \\ &= (1 - R_1) \frac{p_1}{p_1(1 - R_1) + (1 - p_1)R_1} \\ &= 1 - o(1),\end{aligned}$$

and similarly, given $Y_j(k) = 0$,

$$\begin{aligned}&P\left(X_1(k) = 1 \middle| \Pi(1) = j, \mathbf{Y}, \Pi(J^{(n)})\right) \\ &= R_1 \frac{P(X_1(k) = 1)}{P(Y_{\Pi(1)}(k) = 0)} \\ &= R_1 \frac{p_1}{p_1\ R_1 + (1 - p_1)(1 - R_1)} \\ &= o(1).\end{aligned}$$

Note that by the independence assumption, the above probabilities do not depend on the other values of $Y_u(k)$ (as we are conditioning on $\Pi(1) = j$). Thus, we can write

$$\begin{aligned}\widehat{W_j^{(n)}} &= \frac{1}{N^{(n)}} \sum_{j \in \Pi(J^{(n)})} P\left(X_1(k) = 1 \middle| \Pi(1) = j, \mathbf{Y}, \Pi(J^{(n)})\right) \\ &= \frac{1}{N^{(n)}} \sum_{j \in \Pi(J^{(n)}), Y_j(k) = 1} (1 - o(1)) \\ &\quad + \frac{1}{N^{(n)}} \sum_{j \in \Pi(J^{(n)}), Y_j(k) = 0} o(1).\end{aligned}$$

First, note that since $\left|\left\{j \in \Pi(J^{(n)}), Y_j(k) = 0\right\}\right| \leq N^{(n)}$, the second term above converges to zero, thus,

$$\widehat{W_j^{(n)}} \to \frac{\left|\left\{j \in \Pi(J^{(n)}), Y_{\Pi(1)}(k) = 1\right\}\right|}{N^{(n)}}.$$

Since for all $j \in \Pi(J^{(n)})$, $Y_j(k) \sim$ Bernoulli $(p_1 + o(1))$, by a simple application of Chebyshev's inequality, we can conclude $\widehat{W_j^{(n)}} \to p_1$. Appendix D provides the detail. □

As a result,

$$X_1(k) | \mathbf{Y}, \Pi(J^{(n)}) \to \text{Bernoulli}(p_1),$$

thus,

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

Since conditioning reduces entropy,

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

and as a result,

$$\lim_{n\to\infty} H\left(X_1(k)\right) - H\left(X_1(k)\middle|\mathbf{Y}\right) \leq 0,$$

and

$$\lim_{n\to\infty} I\left(X_1(k); \mathbf{Y}\right) \leq 0.$$

By knowing that $I\left(X_1(k); \mathbf{Y}\right)$ cannot take any negative value, we can conclude that

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

□

B. Extension to r-State

Now, assume users' data samples can have r possibilities $(0, 1, \cdots, r-1)$, and $p_u(i)$ shows the probability of user u having data sample i. We define the vector $\mathbf{p}_u$ and the matrix $\mathbf{p}$ as

$$\mathbf{p}_u = \begin{bmatrix} p_u(1) \\ p_u(2) \\ \vdots \\ p_u(r-1) \end{bmatrix}, \quad \mathbf{p} = \left[\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n\right].$$

We assume $p_u(i)$'s are drawn independently from some continuous density function, $f_{\mathbf{P}}(\mathbf{p}_u)$, which has support on a subset of the $(0, 1)^{r-1}$ hypercube (Note that the $p_u(i)$'s sum to one, so one of them can be considered as the dependent value and the dimension is $r-1$). In particular, define the range of the distribution as

$$\mathcal{R}_{\mathbf{p}} = \Bigg\{(x_1, x_2, \cdots, x_{r-1}) \in (0,1)^{r-1} : x_i > 0, x_1 + x_2 + \cdots + x_{r-1} < 1,\ i = 1, 2, \cdots, r-1\Bigg\}.$$

Figure 8 shows the range $\mathcal{R}_{\mathbf{p}}$ for the case where $r = 3$. Then, we assume there are $\delta_1, \delta_2 > 0$ such that:

$$\begin{cases} \delta_1 < f_{\mathbf{P}}(\mathbf{p}_u) < \delta_2, & \mathbf{p}_u \in \mathcal{R}_{\mathbf{p}}. \\ f_{\mathbf{P}}(\mathbf{p}_u) = 0, & \mathbf{p}_u \notin \mathcal{R}_{\mathbf{p}}. \end{cases}$$

The obfuscation is similar to the two-state case. Specifically, for $l \in \{0, 1, \cdots, r-1\}$, we can write

$$P(Z_u(k) = l | X_u(k) = i) = \begin{cases} 1 - R_u, & \text{for } l = i. \\ \dfrac{R_u}{r-1}, & \text{for } l \neq i. \end{cases}$$

Theorem 2: For the above r-state model, if $\mathbf{Z}$ is the obfuscated version of $\mathbf{X}$, and $\mathbf{Y}$ is the anonymized version of $\mathbf{Z}$ as defined previously, and

- $m = m(n)$ is arbitrary;
- $R_u \sim$ Uniform$[0, a_n]$, where $a_n = c' n^{-\left(\frac{1}{r-1} - \beta\right)}$ for any $c' > 0$ and $0 < \beta < \frac{1}{r-1}$;

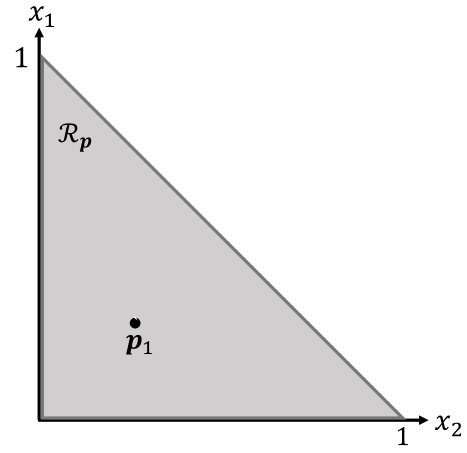

Fig. 8. $\mathcal{R}_{\mathbf{p}}$ for case $r = 3$.

then, user 1 has perfect privacy at time k as n goes to infinity.

The proof of Theorem 2 is similar to the proof of Theorem 1. The major difference is that instead of the random variables P_u, Q_u, V_u, we need to consider the random vectors $\mathbf{P}_u, \mathbf{Q}_u, \mathbf{V}_u$. Similarly, for user u, we define the vector $\mathbf{Q}_u$ as

$$\mathbf{Q}_u = \begin{bmatrix} Q_u(1) \\ Q_u(2) \\ \vdots \\ Q_u(r-1) \end{bmatrix}.$$

In the r-state case,

$$\begin{aligned} Q_u(i) &= P_u(i)\Big(1 - R_u(i)\Big) + \Big(1 - P_u(i)\Big)\frac{R_u}{r-1} \\ &= P_u(i) + \Big(1 - r P_u(i)\Big)\frac{R_u}{r-1}. \end{aligned}$$

We also need to define the critical set $J^{(n)}$. First, for $i = 0, 1, \cdots, r-1$, define set $J_i^{(n)}$ as follows. If $0 \leq p_1(i) < \frac{1}{r}$, then,

$$J_i^{(n)} = \Bigg\{\{u \in \{1, 2, \ldots, n\} : p_1(i) \leq P_u(i) \leq p_1(i) + \epsilon_n ; p_1(i) + \epsilon_n \leq Q_u(i) \leq p_1(i) + (1 - r p_1(i))\frac{a_n}{r-1}\Bigg\},$$

where $\epsilon_n = n^{-\left(\frac{1}{r-1} - \frac{\beta}{2}\right)}$, $a_n = c' n^{-\left(\frac{1}{r-1} - \beta\right)}$, and β is defined in the statement of Theorem 2.

We then define the critical set $J^{(n)}$ as:

$$J^{(n)} = \bigcap_{l=0}^{r-1} J_i^{(n)}.$$

We can then repeat the same arguments in the proof of Theorem 1 to complete the proof.

IV. Converse Results: No Privacy Region

In this section, we prove that if the number of observations by the adversary is larger than its critical value and the noise level is less than its critical value, then the adversary can find an algorithm to successfully estimate users' data samples

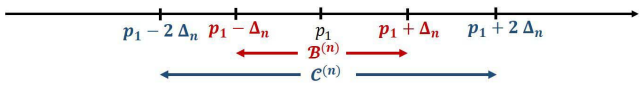

Fig. 9. p_1, sets $\mathcal{B}^{(n)}$ and $\mathcal{C}^{(n)}$ for case $r = 2$.

with arbitrarily small error probability. Combined with the results of the previous section, this implies that asymptotically (as $n \to \infty$), privacy can be achieved *if and only if* at least one of the two techniques (obfuscation or anonymization) are used above their thresholds. This statement needs a clarification as follows: Looking at the results of Montazeri *et al.* [52], we notice that anonymization alone can provide perfect privacy if $m(n)$ is below its threshold. On the other hand, the threshold for obfuscation requires some anonymization: In particular, the identities of the users must be permuted once to prevent the adversary from readily identifying the users.

A. Two-State Model

Again, we start with the i.i.d. two-state model. The data sample of user u at any time is a Bernoulli random variable with parameter p_u.

As before, we assume that p_u's are drawn independently from some continuous density function, $f_P(p_u)$, on the $(0, 1)$ interval. Specifically, there are $\delta_1, \delta_2 > 0$ such that:

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

Theorem 3: For the above two-state mode, if $\mathbf{Z}$ is the obfuscated version of $\mathbf{X}$, and $\mathbf{Y}$ is the anonymized version of $\mathbf{Z}$ as defined, and

- $m = cn^{2+\alpha}$ for any $c > 0$ and $\alpha > 0$;
- $R_u \sim$ Uniform$[0, a_n]$, where $a_n = c' n^{-(1+\beta)}$ for any $c' > 0$ and $\beta > \frac{\alpha}{4}$;

then,user 1 has no privacy at time k as n goes to infinity.

Since this is a converse result, we give an explicit detector at the adversary and show that it can be used by the adversary to recover the true data of user 1.

Proof: The adversary first inverts the anonymization mapping Π to obtain $Z_1(k)$, and then estimates the value of $X_1(k)$ from that. To invert the anonymization, the adversary calculates the empirical probability that each string is in state 1 and then assigns the string with the empirical probability closest to p_1 to user 1.

Formally, for $u = 1, 2, \cdots, n$, the adversary computes $\overline{Y_u}$, the empirical probability of user u being in state 1, as follows:

$$\overline{Y_u} = \frac{Y_u(1) + Y_u(2) + \cdots + Y_u(m)}{m},$$

thus,

$$\overline{Y_{\Pi(u)}} = \frac{Z_u(1) + Z_u(2) + \cdots + Z_u(m)}{m}.$$

As shown in Figure 9, define

$$\mathcal{B}^{(n)} = \{x \in (0,1);\, p_1 - \Delta_n \le x \le p_1 + \Delta_n\},$$

where $\Delta_n = n^{-\left(1+\frac{\alpha}{4}\right)}$ and α is defined in the statement of Theorem 3. We claim that for $m = cn^{2+\alpha}$, $a_n = c' n^{-(1+\beta)}$, and large enough n,

1) $P\left(\overline{Y_{\Pi(1)}} \in \mathcal{B}^{(n)}\right) \to 1$.
2) $P\left(\bigcup_{u=2}^{n} \left(\overline{Y_{\Pi(u)}} \in \mathcal{B}^{(n)}\right)\right) \to 0$.

As a result, the adversary can identify $\Pi(1)$ by examining $\overline{Y_u}$'s and assigning the one in $\mathcal{B}^{(n)}$ to user 1. Note that $\overline{Y_{\Pi(u)}} \in \mathcal{B}^{(n)}$ is a set (event) in the underlying probability space and can be written as $\left\{\omega \in \Omega : \overline{Y_{\Pi(u)}}(\omega) \in \mathcal{B}^{(n)}\right\}$.

First, we show that as n goes to infinity,

$$P\left(\overline{Y_{\Pi(1)}} \in \mathcal{B}^{(n)}\right) \to 1.$$

We can write

$$\begin{aligned} &P\left(\overline{Y_{\Pi(1)}} \in \mathcal{B}^{(n)}\right) \\ &= P\left(\frac{\sum_{k=1}^{m} Z_1(k)}{m} \in \mathcal{B}^{(n)}\right) \\ &= P\left(p_1 - \Delta_n \le \frac{\sum_{k=1}^{m} Z_1(k)}{m} \le p_1 + \Delta_n\right) \\ &= P\left(p_1 - \Delta_n - Q_1 \le \frac{\sum_{k=1}^{m} Z_1(k)}{m} - Q_1 \le p_1 + \Delta_n - Q_1\right). \end{aligned}$$

Note that for any $u \in \{1, 2, \cdots, n\}$, we have

$$\begin{aligned} |p_u - Q_u| &= |1 - 2p_u| R_u \\ &\le R_u \le a_n, \end{aligned}$$

so, we can conclude

$$\begin{aligned} &P\left(\overline{Y_{\Pi(1)}} \in \mathcal{B}^{(n)}\right) \\ &= P\left(\frac{\sum_{k=1}^{m} Z_1(k)}{m} \in \mathcal{B}^{(n)}\right) \\ &\ge P\left(-\Delta_n + a_n \le \frac{\sum_{k=1}^{m} Z_1(k)}{m} - Q_1 \le -a_n + \Delta_n\right) \\ &= P\left(\left|\sum_{k=1}^{m} Z_1(k) - mQ_1\right| \le m(\Delta_n - a_n)\right). \end{aligned}$$

From the Chernoff bound, for any $c, c', \alpha > 0$ and $\beta > \frac{\alpha}{4}$,

$$\begin{aligned} &P\left(\left|\sum_{k=1}^{m} Z_1(k) - mQ_1\right| \le m(\Delta_n - a_n)\right) \\ &\ge 1 - 2e^{-\frac{m(\Delta_n - a_n)^2}{3Q_1}} \\ &\ge 1 - 2e^{-\left(\frac{1}{3Q_1}\right)(cn^{2+\alpha})\left(\frac{1}{n^{1+\frac{\alpha}{4}}} - \frac{c'}{n^{1+\beta}}\right)^2} \\ &\ge 1 - 2e^{-\frac{1}{3}(cn^{2+\alpha})\left(\frac{1}{n^{1+\frac{\alpha}{4}}} - \frac{c'}{n^{1+\beta}}\right)^2} \to 1. \end{aligned}$$

As a result, as n becomes large,

$$P\left(\overline{Y_{\Pi(1)}} \in \mathcal{B}^{(n)}\right) \to 1.$$

Now, we need to show that as n goes to infinity,

$$P\left(\bigcup_{u=2}^{n}\left(\overline{Y_{\Pi(u)}}\in\mathcal{B}^{(n)}\right)\right)\to 0.$$

First, we define

$$\mathcal{C}^{(n)}=\{x\in(0,1);\, p_1-2\Delta_n\le x\le p_1+2\Delta_n\},$$

and claim as n goes to infinity,

$$P\left(\bigcup_{u=2}^{n}\left(P_u\in\mathcal{C}^{(n)}\right)\right)\to 0.$$

Note

$$4\Delta_n\delta_1<P\left(P_u\in\mathcal{C}^{(n)}\right)<4\Delta_n\delta_2,$$

and according to the union bound, for large enough n,

$$\begin{aligned}P\left(\bigcup_{u=2}^{n}\left(P_u\in\mathcal{C}^{(n)}\right)\right)&\le\sum_{u=2}^{n}P\left(P_u\in\mathcal{C}^{(n)}\right)\\&\le 4n\Delta_n\delta_2\\&=4n\frac{1}{n^{1+\frac{\alpha}{4}}}\delta_2\\&=4n^{-\frac{\alpha}{4}}\delta_2\to 0.\end{aligned}$$

As a result, we can conclude that all p_u's are outside of $\mathcal{C}^{(n)}$ for $u\in\{2,3,\cdots,n\}$ with high probability.

Now, we claim that given all p_u's are outside of $\mathcal{C}^{(n)}$, $P\left(\overline{Y_{\Pi(u)}}\in\mathcal{B}^{(n)}\right)$ is small. Remember that for any $u\in\{1,2,\cdots,n\}$, we have

$$|p_u-Q_u|\le a_n.$$

Now, noting the definitions of sets $\mathcal{B}^{(n)}$ and $\mathcal{C}^{(n)}$, we can write for $u\in\{2,3,\cdots,n\}$,

$$\begin{aligned}&P\left(\overline{Y_{\Pi(u)}}\in\mathcal{B}^{(n)}\right)\\&\le P\left(\left|\overline{Y_{\Pi(u)}}-Q_u\right|\ge(\Delta_n-a_n)\right)\\&=P\left(\left|\sum_{k=1}^{m}Z_u(k)-mQ_u\right|>m(\Delta_n-a_n)\right).\end{aligned}$$

According to the Chernoff bound, for any $c,c',\alpha>0$ and $\beta>\frac{\alpha}{4}$,

$$\begin{aligned}&P\left(\left|\sum_{k=1}^{m}Z_u(k)-mQ_u\right|>m(\Delta_n-a_n)\right)\\&\le 2e^{-\frac{m(\Delta_n-a_n)^2}{3Q_1}}\\&\le 2e^{-\left(\frac{1}{3Q_1}\right)(cn^{2+\alpha})\left(\frac{1}{n^{1+\frac{\alpha}{4}}}-\frac{c'}{n^{1+\beta}}\right)^2}\\&\le 2e^{-\frac{1}{3}(cn^{2+\alpha})\left(\frac{1}{n^{1+\frac{\alpha}{4}}}-\frac{c'}{n^{1+\beta}}\right)^2}.\end{aligned}$$

Now, by using a union bound, for any $\beta>\frac{\alpha}{4}$, we have

$$\begin{aligned}P\left(\bigcup_{u=2}^{n}\left(\overline{Y_{\Pi(u)}}\in\mathcal{B}^{(n)}\right)\right)&\le\sum_{u=2}^{n}P\left(\overline{Y_{\Pi(u)}}\in\mathcal{B}^{(n)}\right)\\&\le n\left(2e^{-\frac{1}{3}(cn^{2+\alpha})\left(\frac{1}{n^{1+\frac{\alpha}{4}}}-\frac{c'}{n^{1+\beta}}\right)^2}\right),\end{aligned}$$

and thus, as n goes to infinity,

$$P\left(\bigcup_{u=2}^{n}\left(\overline{Y_{\Pi(u)}}\in\mathcal{B}^{(n)}\right)\right)\to 0.$$

So, the adversary can successfully recover $Z_1(k)$. Since $Z_1(k)=X_1(k)$ with probability $1-R_1=1-o(1)$, the adversary can recover $X_1(k)$ with vanishing error probability for large enough n. □

B. Extension to r-State

Now, assume users' data samples can have r possibilities $(0,1,\cdots,r-1)$, and $p_u(i)$ shows the probability of user u having data sample i. We define the vector $\mathbf{p}_u$ and the matrix $\mathbf{p}$ as

$$\mathbf{p}_u=\begin{bmatrix}p_u(1)\\p_u(2)\\\vdots\\p_u(r-1)\end{bmatrix},\quad\mathbf{p}=\left[\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_n\right].$$

We also assume $\mathbf{p}_u$'s are drawn independently from some continuous density function, $f_P(\mathbf{p}_u)$, which has support on a subset of the $(0,1)^{r-1}$ hypercube. In particular, define the range of distribution as

$$\begin{aligned}&\mathcal{R}_{\mathbf{p}}\\&=\Big\{(x_1,x_2,\cdots,x_{r-1})\in(0,1)^{r-1}\\&\quad:x_i>0,x_1+x_2+\cdots+x_{r-1}<1,\ i=1,2,\cdots,r-1\Big\}.\end{aligned}$$

Then, we assume there are $\delta_1,\delta_2>0$ such that:

$$\begin{cases}\delta_1<f_{\mathbf{P}}(\mathbf{p}_u)<\delta_2, & \mathbf{p}_u\in\mathcal{R}_{\mathbf{p}}.\\ f_{\mathbf{P}}(\mathbf{p}_u)=0, & \mathbf{p}_u\notin\mathcal{R}_{\mathbf{p}}.\end{cases}$$

Theorem 4: For the above r-state mode, if $\mathbf{Z}$ is the obfuscated version of $\mathbf{X}$, and $\mathbf{Y}$ is the anonymized version of $\mathbf{Z}$ as defined, and

- $m=cn^{\frac{2}{r-1}+\alpha}$ for any $c>0$ and $0<\alpha<1$;
- $R_u\sim\text{Uniform}[0,a_n]$, where $a_n=c'n^{-\left(\frac{1}{r-1}+\beta\right)}$ for any $c'>0$ and $\beta>\frac{\alpha}{4}$;

then, user 1 has no privacy at time k as n goes to infinity.

The proof of Theorem 4 is similar to the proof of Theorem 3, so we just provide the general idea. We similarly define the empirical probability that the user with pseudonym u has data sample i $\left(\overline{Y_u}(i)\right)$ as follows:

$$\overline{Y_u}(i)=\frac{|\{k\in\{1,2,\cdots,m\}:Y_u(k)=i\}|}{m},$$

thus,

$$\overline{Y_{\Pi(u)}}(i)=\frac{|\{k\in\{1,2,\cdots,m\}:Z_u(k)=i\}|}{m}.$$

The difference is that now for each $u\in\{1,2,\cdots,n\}$, $\overline{\mathbf{Y}_u}$ is a vector of size $r-1$. In other words,

$$\overline{\mathbf{Y}_u}=\begin{bmatrix}\overline{Y_u}(1)\\\overline{Y_u}(2)\\\vdots\\\overline{Y_u}(r-1)\end{bmatrix}.$$

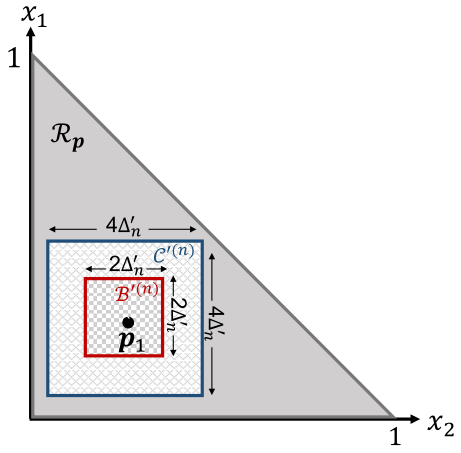

Fig. 10. $\mathbf{p}_1$, sets $\mathcal{B}'^{(n)}$ and $\mathcal{C}'^{(n)}$ in $\mathcal{R}_\mathbf{p}$ for case $r = 3$.

Define sets $\mathcal{B}'^{(n)}$ and $\mathcal{C}'^{(n)}$ as

$$\mathcal{B}'^{(n)} = \Big\{(x_1, x_2, \cdots, x_{r-1}) \in \mathcal{R}_\mathbf{p} : p_1(i) - \Delta'_n \le x_i \le p_1(i) + \Delta'_n,\ i = 1, 2, \cdots, r-1\Big\},$$

$$\mathcal{C}'^{(n)} = \Big\{(x_1, x_2, \cdots, x_{r-1}) \in \mathcal{R}_\mathbf{p} : p_1(i) - 2\Delta'_n \le x_i \le p_1(i) + 2\Delta'_n,\ i = 1, 2, \cdots, r-1\Big\},$$

where $\Delta'_n = n^{-\left(\frac{1}{r-1}+\frac{\alpha}{4}\right)}$. Figure 10 shows $\mathbf{p}_1$ and sets $\mathcal{B}'^{(n)}$ and $\mathcal{C}'^{(n)}$ for the case $r = 3$.

We claim for $m = cn^{\frac{2}{r-1}+\alpha}$, $a_n = c'n^{-\left(\frac{1}{r-1}+\beta\right)}$, and large enough n,

1) $P\left(\overline{\mathbf{Y}_{\Pi(1)}} \in \mathcal{B}'^{(n)}\right) \to 1$.
2) $P\left(\bigcup_{u=2}^{n} \left(\overline{\mathbf{Y}_{\Pi(u)}} \in \mathcal{B}'^{(n)}\right)\right) \to 0$.

The proof follows that for the two-state case. Thus, the adversary can de-anonymize the data and then recover $X_1(k)$ with vanishing error probability in the r-state model.

C. Markov Chain Model

So far, we have assumed users' data samples can have r possibilities $(0, 1, \cdots, r-1)$ and users' pattern are i.i.d.. Here we model users' pattern using Markov chains to capture the dependency of the users' pattern over time. Again, we assume there are r possibilities (the number of states in the Markov chains). Let E be the set of edges. More specifically, $(i, l) \in E$ if there exists an edge from i to l with probability $p(i, l) > 0$. What distinguishes different users is their transition probabilities $p_u(i, l)$ (the probability that user u jumps from state i to state l). The adversary knows the transition probabilities of all users. The model for obfuscation and anonymization is exactly the same as before.

We show that the adversary will be able to estimate the data samples of the users with low error probability if $m(n)$ and a_n are in the appropriate range. The key idea is that the adversary can focus on a subset of the transition probabilities that are sufficient for recovering the entire transition probability matrix. By estimating those transition probabilities from the observed data and matching with the known transition probabilities of the users, the adversary will be able to first de-anonymize the data, and then estimate the actual samples of users' data. In particular, note that for each state i, we must have

$$\sum_{l=1}^{r} p_u(i, l) = 1, \quad \text{for each } u \in \{1, 2, \cdots, n\},$$

so, the Markov chain of user u is completely determined by a subset of size $d = |E| - r$ of transition probabilities. We define the vector $\mathbf{p}_u$ and the matrix $\mathbf{p}$ as

$$\mathbf{p}_u = \begin{bmatrix} p_u(1) \\ p_u(2) \\ \vdots \\ p_u(|E| - r) \end{bmatrix}, \quad \mathbf{p} = \big[\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n\big].$$

We also consider $\mathbf{p}_u$'s are drawn independently from some continuous density function, $f_P(\mathbf{p}_u)$, which has support on a subset of the $(0, 1)^{|E|-r}$ hypercube. Let $\mathcal{R}_\mathbf{p} \subset \mathbb{R}^d$ be the range of acceptable values for $\mathbf{p}_u$, so we have

$$\mathcal{R}_\mathbf{p} = \Big\{(x_1, x_2, \cdots, x_d) \in (0, 1)^d : x_i > 0, x_1 + x_2 + \cdots + x_d < 1,\ i = 1, 2, \cdots, d\Big\}.$$

As before, we assume there are $\delta_1, \delta_2 > 0$, such that:

$$\begin{cases} \delta_1 < f_\mathbf{P}(\mathbf{p}_u) < \delta_2, & \mathbf{p}_u \in \mathcal{R}_\mathbf{p}. \\ f_\mathbf{P}(\mathbf{p}_u) = 0, & \mathbf{p}_u \notin \mathcal{R}_\mathbf{p}. \end{cases}$$

Using the above observations, we can establish the following theorem.

Theorem 5: For an irreducible, aperiodic Markov chain with r states and $|E|$ edges as defined above, if $\mathbf{Z}$ is the obfuscated version of $\mathbf{X}$, and $\mathbf{Y}$ is the anonymized version of $\mathbf{Z}$, and

- $m = cn^{\frac{2}{|E|-r}+\alpha}$ for any $c > 0$ and $\alpha > 0$;
- $R_u \sim \text{Uniform}[0, a_n]$, where $a_n = c'n^{-\left(\frac{1}{|E|-r}+\beta\right)}$ for any $c' > 0$ and $\beta > \frac{\alpha}{4}$;

then, user 1 has no privacy at time k as n goes to infinity.

The proof has a lot of similarity to the i.i.d. case, so we provide a sketch, mainly focusing on the differences. We argue as follows. 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 i, for $i = 0, 1, \cdots, r-1$. 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. Now conditioned on $M_i(u) = m_i(u)$, the transitions from state i to state l for user u follow a multinomial distribution with probabilities $p_u(i, l)$.

Given the above, the setting is now very similar to the i.i.d. case. Each user is uniquely characterized by a vector $\mathbf{p}_u$ of

size $|E|-r$. We define sets $\mathcal{B}^{''(n)}$ and $\mathcal{C}^{''(n)}$ as

$$\mathcal{B}^{''(n)} = \Big\{(x_1, x_2, \cdots, x_d) \in \mathcal{R}_{\mathbf{p}} : p_1(i) - \Delta_n'' \le x_i \le p_1(i) + \Delta_n'',\ i = 1, 2, \cdots, d\Big\},$$

$$\mathcal{C}^{''(n)} = \Big\{(x_1, x_2, \cdots, x_d) \in \mathcal{R}_{\mathbf{p}} : p_1(i) - 2\Delta_n'' \le x_i \le p_1(i) + 2\Delta_n'',\ i = 1, 2, \cdots, d\Big\},$$

where $\Delta_n'' = n^{-\left(\frac{1}{|E|-r}+\frac{\alpha}{4}\right)}$, and $d = |E| - r$. Then, we can show that for the stated values of $m(n)$ and a_n, as n becomes large:

1) $P\left(\overline{\mathbf{Y}_{\Pi(1)}} \in \mathcal{B}^{''(n)}\right) \to 1$,
2) $P\left(\bigcup_{u=2}^{n}\left(\overline{\mathbf{Y}_{\Pi(u)}} \in B^{''(n)}\right)\right) \to 0$,

which means that the adversary can estimate the data of user 1 with vanishing error probability. The proof is very similar to the proof of the i.i.d. case; however, there are two differences that need to be addressed:

First, the probability of observing an erroneous observation is not exactly given by R_u. In fact, a transition is distorted if at least one of its nodes is distorted. So, if the actual transition is from state i to state l, then the probability of an erroneous observation is equal to

$$R_u' = R_u + R_u - R_u R_u = R_u(2 - R_u).$$

Nevertheless, here the order only matters, and the above expression is still in the order of $a_n = O\left(n^{-\left(\frac{1}{|E|-r}+\beta\right)}\right)$.

The second difference is more subtle. As opposed to the i.i.d. case, the error probabilities are not completely independent. In particular, if $X_u(k)$ is reported in error, then both the transition to that state and from that state are reported in error. This means that there is a dependency between errors of adjacent transitions. We can address this issue in the following way: The adversary makes his decision only based on a subset of the observations. More specifically, the adversary looks at only odd-numbered transitions: First, third, fifth, etc., and ignores the even-numbered transitions. In this way, the number of observations is effectively reduced from m to $\frac{m}{2}$ which again does not impact the order of the result (recall that the Markov chain is aperiodic). However, the adversary now has access to observations with independent errors.

V. Perfect Privacy Analysis: Markov Chain Model

So far, we have provided both achievability and converse results for the i.i.d. case. However, we have only provided the converse results for the Markov chain case. Here, we investigate achievability for Markov chain models. It turns out that for this case, the assumed obfuscation technique is not sufficient to achieve a reasonable level of privacy. Loosely speaking, we can state that if the adversary can make enough observations, then he can break the anonymity. The culprit is the fact that the sequence observed by the adversary is no longer modeled by a Markov chain; rather, it can be modeled by a hidden Markov chain. This allows the adversary to successfully estimate the obfuscation random variable R_u as well as the $p_u(i, l)$ values for each sequence, and hence successfully de-anonymize the sequences.

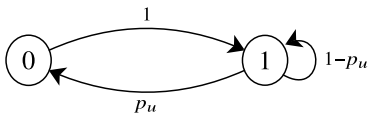

Fig. 11. A state transition diagram.

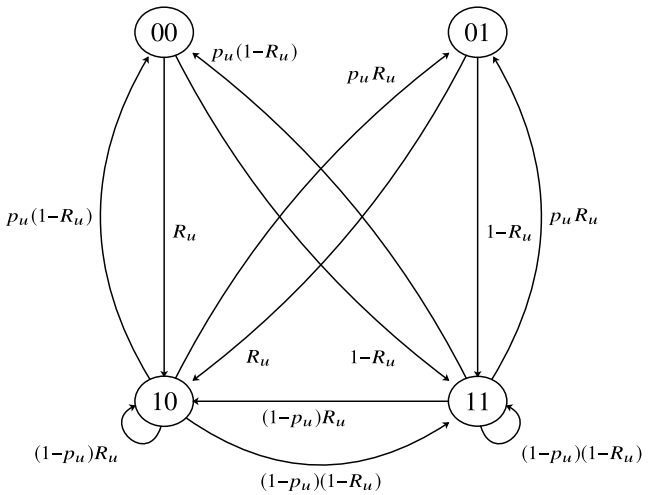

Fig. 12. The state transition diagram of the new Markov chain.

More specifically, as we will see below, there is a fundamental difference between the i.i.d. case and the Markov chain case. In the i.i.d. case, if the noise level is beyond a relatively small threshold, the adversary will be unable to de-anonymize the data and unable to recover the actual values of the data sets for users, *regardless of the (large) size of* $m = m(n)$. On the other hand, in the Markov chain case, if $m = m(n)$ is large enough, then the adversary can easily de-anonymize the data. To better illustrate this, let us consider a simple example.

Example 1: Consider the scenario where there are only two states and the users' data samples change between the two states according to the Markov chain shown in Figure 11. What distinguishes the users is their different values of p_u. Now, suppose we use the same obfuscation method as before. That is, to create a noisy version of the sequences of data samples, for each user u, we generate the random variable R_u that is the probability that the data sample of the user is changed to a different data sample by obfuscation. Specifically,

$$Z_u(k) = \begin{cases} X_u(k), & \text{with probability } 1 - R_u. \\ 1 - X_u(k), & \text{with probability } R_u. \end{cases}$$

To analyze this problem, we can construct the underlying Markov chain as follows. Each state in this Markov chain is identified by two values: the real state of the user, and the observed value by the adversary. In particular, we can write

$$(\text{Real value}, \text{Observed value}) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}.$$

Figure 12 shows the state transition diagram of this new Markov chain.

We know

$$\pi_{00} = \pi_0(1 - R_u) = \frac{p_u}{1 + p_u}(1 - R_u).$$
$$\pi_{01} = \pi_0 R_u = \frac{p_u}{1 + p_u} R_u.$$
$$\pi_{10} = \pi_1 R_u = \frac{1}{1 + p_u} R_u.$$
$$\pi_{11} = \pi_1(1 - R_u) = \frac{1}{1 + p_u}(1 - R_u).$$

The observed process by the adversary is not a Markov chain; nevertheless, we can define limiting probabilities. In particular, let θ_0 be the limiting probability of observing a zero. That is, we have

$$\frac{M_0}{m} \xrightarrow{d} \theta_0, \quad \text{as } n \to \infty,$$

where m is the total number of observations by the adversary, and M_0 is the number of 0's observed. Then,

$$\theta_0 = \pi_{00} + \pi_{10} = \frac{(1 - R_u)p_u + R_u}{1 + p_u}.$$

Also, let θ_1 be the limiting probability of observing a one, so

$$\theta_1 = \pi_{01} + \pi_{11} = \frac{p_u R_u + (1 - R_u)}{1 + p_u} = 1 - \theta_0.$$

Now the adversary's estimate of θ_0 is given by:

$$\widetilde{\theta_0} = \frac{(1 - R_u)p_u + R_u}{1 + p_u}. \tag{1}$$

Note that if the number of observations by the adversary can be arbitrarily large, the adversary can obtain an arbitrarily accurate estimate of θ_0. The adversary can obtain another equation easily, as follows. Let θ_{01} be the limiting value of the portion of transitions from state 0 to 1 in the chain observed by the adversary. We can write

$$\begin{aligned}\theta_{01} &= P\{(00 \to 01), (00 \to 11), (10 \to 01), (10 \to 11)\} \\ &= \pi_{00}(1 - R_u) + \pi_{10} p_u R_u + \pi_{10}(1 - p_u)(1 - R_u).\end{aligned}$$

As a result,

$$\widetilde{\theta_{01}} = \frac{p_u(1 - R_u)^2 + R_u\,(p_u R_u(1 - R_u)(1 - p_u))}{1 + p_u}. \tag{2}$$

Again, if the number of observations can be arbitrarily large, the adversary can obtain an arbitrarily accurate estimate of θ_{01}. By solving (1) and (2), the adversary can successfully recover R and p; thus, he/she can successfully determine the users' data values.

VI. Discussion

A. Markov Chain Model

As opposed to the i.i.d. case, we see from Section V that if we do not limit $m = m(n)$, the assumed obfuscation method will not be sufficient to achieve perfect privacy. There are a few natural questions here. First, for a given noise level, what would be the maximum $m(n)$ that could guarantee perfect privacy in this model? The more interesting question is, how can we possibly modify the obfuscation technique to make it more suitable for the Markov chain model? A natural solution seems to be re-generating the obfuscation random variables R_u periodically. This will keep the adversary from easily estimating them by observing a long sequence of data at a small increase in complexity. In fact, this will make the obfuscation much more *robust* to modeling uncertainties and errors. It is worth noting, however, that this change would not affect the other results in the paper. That is, even if the obfuscation random variables are re-generated frequently, it is relatively easy to check that all the previous theorems in the paper remain valid. However, the increase in robustness to modeling errors will definitely be a significant advantage. Thus, the question is how often should the random variable R_u be re-generated to strike a good balance between complexity and privacy? These are all interesting questions for future research.

B. Obfuscating the Samples of Users' Data Using Continuous Noise

Here we argue that for the setting of this paper, continuous noise such as that drawn from a Gaussian distribution is not a good option to obfuscate the sample of users' data drawn from a finite alphabet when we want to achieve perfect privacy. For a better understanding, let us consider a simple example.

Example 2: Consider the scenario where the users' datasets are governed by an i.i.d. model and the number of possible values for each sample of the users' data (r) is equal to 2 (two-state model). Note that the data sequence for user u is a Bernoulli random variable with parameter p_u.

Assume that the actual sample of the data of user u at time k ($X_u(k)$) is obfuscated using noise drawn from a Gaussian distribution ($S_u(k)$), and $Z_u(k)$ is the obfuscated version of $X_u(k)$. That is, we can write

$$Z_u(k) = X_u(k) + S_u(k);$$

where $S_u(k) \sim N\left(\mu(R_u), \sigma^2(R_u)\right)$, is independent of $X_u(k)$, and R_u is the noise parameter which is chosen from some distribution. Here, $\mu(R_u)$ and $\sigma^2(R_u)$ are some known functions of R_u. We also apply anonymization to $Z_u(k)$, and, as before, $Y_u(k)$ is the reported sample of the data of user u at time k after applying anonymization. Per Section II, anonymization is modeled by a random permutation $\Pi(u)$ on the set of n users.

Now, the question is as follows: Is it possible to achieve perfect privacy independent of the number of adversary's observation (m) while using this continuous noise ($S_u(k)$) to obfuscate the sample of users' data?

Note that

$$E[Z_u(k)] = p_u + \mu(R_u), \tag{3}$$

and

$$Var\,(Z_u(k)) = p_u(1 - p_u) + \sigma^2(R_u). \tag{4}$$

In this case, when the adversary's number of observations is arbitrarily large, the adversary can obtain good estimates of $E[Z_u(k)]$ and $Var\,(Z_u(k))$ for each user with an arbitrarily small error probability. Then, by using (3) and (4), the adversary can recover p_u and R_u. As a result, the adversary can de-anonymize the data and then recover $X_u(k)$. The conclusion

here is that a continuous noise distribution can potentially give information to the adversary when used for obfuscation of finite alphabet data. A method to mitigate this issue is to regenerate the random variables R_u frequently (similar to our previous discussion for Markov chains). Understanding the optimal frequency of such a regeneration and detailed analysis in this case is an interesting future research direction.

VII. Conclusions

In this paper, we have considered both obfuscation and anonymization techniques to achieve privacy. The privacy level of the users depends on both $m(n)$ (number of observations per user by the adversary for a fixed anonymization mapping) and a_n (noise level). That is, larger $m(n)$ and smaller a_n indicate weaker privacy. We characterized the limits of privacy in the entire $m(n) - a_n$ plane for the i.i.d. case; that is, we obtained the exact values of the thresholds for $m(n)$ and a_n required for privacy to be maintained. We showed that if $m(n)$ is fewer than $O\left(n^{\frac{2}{r-1}}\right)$, or a_n is larger than $\Omega\left(n^{-\frac{1}{r-1}}\right)$, users have perfect privacy. On the other hand, if neither of these two conditions is satisfied, users have no privacy. For the case where the users' patterns are modeled by Markov chains, we obtained a no-privacy region in the $m(n) - a_n$ plane.

Future research in this area needs to characterize the exact privacy/no-privacy regions when user data sequences obey Markov models. It is also important to consider different ways to obfuscate users' data sets and study the utility-privacy trade-offs for different types of obfuscation techniques.

Appendix A
Lemma 6 and Its Proof

Here we state that we can condition on high-probability events.

Lemma 6: Let $p \in (0, 1)$, and $X \sim Bernoulli(p)$ be defined on a probability space $(\Omega, \mathcal{F}, P)$. Consider $B_1, B_2, \cdots$ be a sequence of events defined on the same probability space such that $P(B_n) \to 1$ as n goes to infinity. Also, let $\mathbf{Y}$ be a random vector (matrix) in the same probability space, then:

$$I(X; \mathbf{Y}) \to 0 \quad \text{iff } I(X; \mathbf{Y}|B_n) \to 0.$$

Proof: First, we prove that as n becomes large,

$$H(X|B_n) - H(X) \to 0. \tag{5}$$

Note that as n goes to infinity,

$$\begin{aligned} P(X = 1) &= P\left(X = 1\middle|B_n\right) P(B_n) + P\left(X = 1\middle|\overline{B_n}\right) P\left(\overline{B_n}\right) \\ &= P\left(X = 1\middle|B_n\right), \end{aligned}$$

thus, $\left(X\middle|B_n\right) \xrightarrow{d} X$, and as n goes to infinity,

$$H(X|B_n) - H(X) \to 0.$$

Similarly, as n becomes large,

$$P\left(X = 1\middle|\mathbf{Y} = \mathbf{y}\right) \to P\left(X = 1\middle|\mathbf{Y} = \mathbf{y}, B_n\right),$$

and

$$H(X|\mathbf{Y} = \mathbf{y}, B_n) - H(X|\mathbf{Y} = \mathbf{y}) \to 0. \tag{6}$$

Remembering that

$$I(X; \mathbf{Y}) = H(X) - H(X|\mathbf{Y}), \tag{7}$$

and using (5), (6), and (7), we can conclude that as n goes to infinity,

$$I(X; \mathbf{Y}|B_n) - I(X, \mathbf{Y}) \to 0.$$

As a result, for large enough n,

$$I(X; \mathbf{Y}) \to 0 \Longleftrightarrow I(X; \mathbf{Y}|B_n) \to 0.$$

□

Appendix B
Proof of Lemma 1

Here we provide a formal proof for Lemma 1 which we restate as follows.

Let N be a positive integer, and let $a_1, a_2, \cdots, a_N$ and $b_1, b_2, \cdots, b_N$ be real numbers such that $a_u \leq b_u$ for all u. Assume that $X_1, X_2, \cdots, X_N$ are N independent random variables such that

$$X_u \sim Uniform[a_u, b_u].$$

Let also $\gamma_1, \gamma_2, \cdots, \gamma_N$ be real numbers such that

$$\gamma_j \in \bigcap_{u=1}^{N} [a_u, b_u] \quad \text{for all } j \in \{1, 2, \cdots, N\}.$$

Suppose that we know the event E has occurred, meaning that the observed values of X_u's is equal to the set of γ_j's (but with unknown ordering), i.e.,

$$E \equiv \{X_1, X_2, \cdots, X_N\} = \{\gamma_1, \gamma_2, \cdots, \gamma_N\},$$

then

$$P\left(X_1 = \gamma_j|E\right) = \frac{1}{N}.$$

Proof: Define sets $\mathfrak{P}$ and $\mathfrak{P}_j$ as follows:

$\mathfrak{P}$ = The set of all permutations Π on $\{1, 2, \cdots, N\}$.

$\mathfrak{P}_j$ = The set of all permutations Π on $\{1, 2, \cdots, N\}$ such that $\Pi(1) = j$.

We have $|\mathfrak{P}| = N!$ and $|\mathfrak{P}| = (N-1)!$. Then

$$\begin{aligned} &P(X_1 = \alpha_j|E) \\ &= \frac{\sum_{\pi \in \mathfrak{P}_j} f_{X_1, X_2, \cdots, X_N}(\gamma_{\pi(1)}, \gamma_{\pi(2)}, \cdots, \gamma_{\pi(N)})}{\sum_{\pi \in \mathfrak{P}} f_{X_1, X_2, \cdots, X_N}(\gamma_{\pi(1)}, \gamma_{\pi(2)}, \cdots, \gamma_{\pi(N)})} \\ &= \frac{(N-1)! \prod_{u=1}^{N} \frac{1}{b_u - a_u}}{N! \prod_{u=1}^{N} \frac{1}{b_u - a_u}} \\ &= \frac{1}{N}. \end{aligned}$$

□

APPENDIX C
PROOF OF LEMMA 4

Here, we provide a formal proof for Lemma 4 which we restate as follows. The following lemma confirms that the number of elements in $J^{(n)}$ goes to infinity as n becomes large.

If $N^{(n)} \triangleq |J^{(n)}|$, then $N^{(n)} \to \infty$ with high probability as $n \to \infty$. More specifically, there exists $\lambda > 0$ such that

$$P\left(N^{(n)} > \frac{\lambda}{2} n^{\frac{\beta}{2}}\right) \to 1.$$

Proof: Define the events A, B as

$$A \equiv p_1 \le P_u \le p_1 + \epsilon_n$$
$$B \equiv p_1 + \epsilon_n \le Q_u \le p_1 + (1-2p_1)a_n.$$

Then, for $u \in \{1, 2, \ldots, n\}$ and $0 \le p_1 < \frac{1}{2}$:

$$\begin{aligned} P\left(u \in J^{(n)}\right) &= P\left(A \cap B\right) \\ &= P\left(A\right) P\left(B\middle|A\right). \end{aligned}$$

So, given $p_1 \in (0,1)$ and the assumption $0 < \delta_1 < f_p < \delta_2$, for n large enough, we have

$$P(A) = \int_{p_1}^{p_1+\epsilon_n} f_P(p)dp,$$

so, we can conclude that

$$\epsilon_n \delta_1 < P(A) < \epsilon_n \delta_2.$$

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

$$P(A) = \epsilon_n \delta. \tag{8}$$

We know

$$Q_u\big|P_u = p_u \sim Uniform\left[p_u, p_u + (1-2p_u)a_n\right],$$

so, according to Figure 6, for $p_1 \le p_u \le p_1 + \epsilon_n$,

$$\begin{aligned} P\left(B|P_u = p_u\right) &= \frac{p_1 + (1-2p_1)a_n - p_1 - \epsilon_n}{p_u + (1-2p_u)a_n - p_u} \\ &= \frac{(1-2p_1)a_n - \epsilon_n}{(1-2p_u)a_n} \\ &\ge \frac{(1-2p_1)a_n - \epsilon_n}{(1-2p_1)a_n} \\ &= 1 - \frac{\epsilon_n}{(1-2p_1)a_n}, \end{aligned}$$

which implies

$$P\left(B|A\right) \ge 1 - \frac{\epsilon_n}{(1-2p_1)a_n}. \tag{9}$$

Using (8) and (9), we can conclude

$$P\left(u \in J^{(n)}\right) \ge \epsilon_n \delta \left(1 - \frac{\epsilon_n}{(1-2p_1)a_n}\right).$$

Then, we can say that $N^{(n)}$ has a binomial distribution with expected value of $N^{(n)}$ greater than $n\epsilon_n\delta\left(1 - \frac{\epsilon_n}{(1-2p_1)a_n}\right)$, and by substituting ϵ_n and a_n, for any $c' > 0$, we get

$$E\left[N^{(n)}\right] \ge \delta\left(n^{\frac{\beta}{2}} - \frac{1}{c'(1-2p_1)}\right) \ge \lambda n^{\frac{\beta}{2}}.$$

Now by using Chernoff bound, we have

$$P\left(N^{(n)} \le (1-\theta)E\left[N^{(n)}\right]\right) \le e^{-\frac{\theta^2}{2}E\left[N^{(n)}\right]},$$

so, if we assume $\theta = \frac{1}{2}$, we can conclude for large enough n,

$$\begin{aligned} P\left(N^{(n)} \le \frac{\lambda}{2} n^{\frac{\beta}{2}}\right) &\le P\left(N^{(n)} \le \frac{E\left[N^{(n)}\right]}{2}\right) \\ &\le e^{-\frac{E\left[N^{(n)}\right]}{8}} \\ &\le e^{-\frac{\lambda n^{\frac{\beta}{2}}}{8}} \to 0. \end{aligned}$$

As a result, $N^{(n)} \to \infty$ with high probability for large enough n. □

APPENDIX D
COMPLETION OF PROOF OF LEMMA 5

Let $p_1 \in (0,1)$, and let $N^{(n)}$ be a random variable as above, i.e., $N^{(n)} \to \infty$ as $n \to \infty$. Consider the sequence of independent random variables $Y_u \sim Bernoulli(p_u)$ for $u = 1, 2, \cdots, N^{(n)}$ such that

1) For all n and all $u \in \left\{1, 2, \cdots, N^{(n)}\right\}$, $|p_u - p_1| \le \zeta_n$.
2) $\lim_{n\to\infty} \zeta_n = 0$.

Define

$$\overline{Y} = \frac{1}{N^{(n)}} \sum_{u=1}^{N^{(n)}} Y_u,$$

then $\overline{Y} \xrightarrow{d} p_1$.

Proof: Note

$$\begin{aligned} E[\overline{Y}] &= \frac{1}{N^{(n)}} \sum_{u=1}^{N^{(n)}} p_u \\ &\le \frac{1}{N^{(n)}} \sum_{u=1}^{N^{(n)}} (p_1 + \zeta_n) \\ &= \frac{1}{N^{(n)}} \cdot N^{(n)}(p_1 + \zeta_n) \\ &= p_1 + \zeta_n. \end{aligned}$$

Similarly we can prove $E\left[\overline{Y}\right] \ge p_1 - \zeta_n$. Since as n becomes large, $\zeta_n \to 0$ and $p_1 \in (0,1)$, we can conclude

$$\lim_{n\to\infty} E\left[\overline{Y}\right] = p_1. \tag{10}$$

Also,

$$\begin{aligned} Var\left(\overline{Y}\right) &= \frac{1}{\left(N^{(n)}\right)^2} \sum_{u=1}^{N^{(n)}} p_u\left(1-p_u\right) \\ &\le \frac{1}{(N^{(n)})^2} \sum_{u=1}^{N^{(n)}} (p_1+\zeta_n)\left(1-p_1+\zeta_n\right) \\ &= \frac{1}{(N^{(n)})^2} \cdot N^{(n)}\left(p_1+\zeta_n\right)\left(1-p_1+\zeta_n\right) \\ &= \frac{1}{N^{(n)}}\left(p_1+\zeta_n\right)\left(1-p_1+\zeta_n\right). \end{aligned}$$

Thus,

$$\lim_{n\to\infty} Var\left(\overline{Y}\right) = 0. \tag{11}$$

By using (10), (11), and Chebyshev's inequality, we can conclude

$$\overline{Y} \xrightarrow{d} p_1.$$

□

REFERENCES

[1] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Limits of location privacy under anonymization and obfuscation," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 764–768.

[2] *Internet of Things: Privacy and Security in a Connected World*, Federal Trade Commission Staff, Washington, DC, USA, 2015.

[3] P. Porambag, M. Ylianttila, C. Schmitt, P. Kumar, A. Gurtov, and A. V. Vasilakos, "The quest for privacy in the Internet of Things," *IEEE Cloud Comput.*, vol. 3, no. 2, pp. 36–45, Mar./Apr. 2016.

[4] A. Ukil, S. Bandyopadhyay, and A. Pal, "IoT-privacy: To be private or not to be private," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Toronto, ON, Canada, Apr./May 2014, pp. 123–124.

[5] S. Hosseinzadeh, S. Rauti, S. Hyrynsalmi, and V. Leppänen, "Security in the Internet of Things through obfuscation and diversification," in *Proc. Int. Conf. Comput., Commun. Secur. (ICCCS)*, Pamplemousses, Mauritius, Dec. 2015, pp. 1–5.

[6] N. Apthorpe, D. Reisman, and N. Feamster, "A smart home is no castle: Privacy vulnerabilities of encrypted iot traffic," in *Proc. Workshop Data Algorithmic Transparency*, New York, NY, USA, 2016.

[7] J. Unnikrishnan, "Asymptotically optimal matching of multiple sequences to source distributions and training sequences," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 452–468, Jan. 2015.

[8] G. P. Corser, H. Fu, and A. Banihani, "Evaluating location privacy in vehicular communications and applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 9, pp. 2658–2667, Sep. 2016.

[9] B. Hoh and M. Gruteser, "Protecting location privacy through path confusion," in *Proc. 1st Int. Conf. Secur. Privacy Emerg. Areas Commun. Netw. (SECURECOMM)*, Pamplemousses, Mauritius, Sep. 2005, pp. 194–205.

[10] J. Freudiger, M. Raya, M. Félegyházi, P. Papadimitratos, and J. P. Hubaux, "Mix-zones for location privacy in vehicular networks," in *Proc. ACM Workshop Wireless Netw. Intell. Transp. Syst. (WiN-ITS)*, Vancouver, BC, Canada, 2007.

[11] Z. Ma, F. Kargl, and M. Weber, "A location privacy metric for V2X communication systems," in *Proc. IEEE Sarnoff Symp.*, Princeton, NJ, USA, Mar./Apr. 2009, pp. 1–6.

[12] R. Shokri, G. Theodorakopoulos, G. Danezis, J.-P. Hubaux, and J. Y. Le Boudec, "Quantifying location privacy: The case of sporadic location exposure," in *Proc. Int. Symp. Privacy Enhancing Technol. Symp.* Waterloo, ON, Canada: Springer, 2011, pp. 57–76.

[13] F. M. Naini, J. Unnikrishnan, P. Thiran, and M. Vetterli, "Where you are is who you are: User identification by matching statistics," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 2, pp. 358–372, Feb. 2016.

[14] R. Soltani, D. Goeckel, D. Towsley, and A. Houmansadr, "Towards provably invisible network flow fingerprints," in *Proc. 51st Asilomar Conf. Signals, Syst., Comput.*, Oct. 2017, pp. 258–262.

[15] R. Shokri, G. Theodorakopoulos, C. Troncoso, J.-P. Hubaux, and J.-Y. Le Boudec, "Protecting location privacy: Optimal strategy against localization attacks," in *Proc. ACM Conf. Comput. Commun. Secur.*, Raleigh, NC, USA, 2012, pp. 617–627.

[16] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *Proc. 1st Int. Conf. Mobile Syst., Appl. Services*, San Francisco, CA, USA, 2003, pp. 31–42.

[17] N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Optimal geo-indistinguishable mechanisms for location privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Scottsdale, AZ, USA, 2014, pp. 251–262.

[18] Y. Zhang, W. Tong, and S. Zhong, "On designing satisfaction-ratio-aware truthful incentive mechanisms for k-anonymity location privacy," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 11, pp. 2528–2541, Nov. 2016.

[19] R. Dewri and R. Thurimella, "Exploiting service similarity for privacy in location-based search queries," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 2, pp. 374–383, Feb. 2014.

[20] B. Gedik and L. Liu, "Location privacy in mobile systems: A personalized anonymization model," in *Proc. 25th IEEE Int. Conf. Distrib. Comput. Syst.*, Columbus, OH, USA, Jun. 2005, pp. 620–629.

[21] G. Zhong and U. Hengartner, "A distributed k-anonymity protocol for location privacy," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, Galveston, TX, USA, Mar. 2009, pp. 1–10.

[22] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[23] P. Kalnis, G. Ghinita, K. Mouratidis, and D. Papadias, "Preventing location-based identity inference in anonymous spatial queries," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 12, pp. 1719–1733, Dec. 2007.

[24] X. Liu, K. Liu, L. Guo, X. Li, and Y. Fang, "A game-theoretic approach for achieving k-anonymity in location based services," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Turin, Italy, Apr. 2013, pp. 2985–2993.

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

[26] J. Freudiger, R. Shokri, and J.-P. Hubaux, "On the optimal placement of mix zones," in *Proc. 9th Int. Symp. Privacy Enhancing Technol.*, Seattle, WA, USA, 2009, pp. 216–234.

[27] B. Palanisamy and L. Liu, "MobiMix: Protecting location privacy with mix-zones over road networks," in *Proc. IEEE 27th Int. Conf. Data Eng. (ICDE)*, Hannover, Germany, Apr. 2011, pp. 494–505.

[28] R. Shokri, G. Theodorakopoulos, P. Papadimitratos, E. Kazemi, and J. Hubaux, "Hiding in the mobile crowd: Location privacy through collaboration," *IEEE Trans. Dependable Secure Comput.*, vol. 11, no. 3, pp. 266–279, Mar. 2013.

[29] M. A. Zurbarán, K. Ávila, P. Wightman, and M. Fernández, "Near-rand: Noise-based location obfuscation based on random neighboring points," *IEEE Latin Amer. Trans.*, vol. 13, no. 11, pp. 3661–3667, Nov. 2015.

[30] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Preserving privacy in gps traces via uncertainty-aware path cloaking," in *Proc. 14th ACM Conf. Comput. Commun. Secur.*, Alexandria, VA, USA, 2007, pp. 161–171.

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

[32] C.-Y. Chow, M. F. Mokbel, and X. Liu, "Spatial cloaking for anonymous location-based services in mobile peer-to-peer environments," *GeoInformatica*, vol. 15, no. 2, pp. 351–380, 2011.

[33] J.-H. Um, H.-D. Kim, and J.-W. Chang, "An advanced cloaking algorithm using Hilbert curves for anonymous location based service," in *Proc. IEEE 2nd Int. Conf. Social Comput. (SocialCom)*, Minneapolis, MN, USA, Aug. 2010, pp. 1093–1098.

[34] H. Kido, Y. Yanagisawa, and T. Satoh, "Protection of location privacy using dummies for location-based services," in *Proc. 21st Int. Conf. Data Eng. Workshops*, Tokyo, Japan, Apr. 2005, p. 1248.

[35] P. Shankar, V. Ganapathy, and L. Iftode, "Privately querying location-based services with SybilQuery," in *Proc. 11th Int. Conf. Ubiquitous Comput.*, Orlando, FL, USA, 2009, pp. 31–40.

[36] R. Chow and P. Golle, "Faking contextual data for fun, profit, and privacy," in *Proc. 8th ACM Workshop Privacy Electron. Society*, Chicago, IL, USA, 2009, pp. 105–108.

[37] H. Kido, Y. Yanagisawa, and T. Satoh, "An anonymous communication technique using dummies for location-based services," in *Proc. Int. Conf. Pervasive Services*, Santorini, Greece, Jul. 2005, pp. 88–97.

[38] H. Lu, C. S. Jensen, and M. L. Yiu, "PAD: Privacy-area aware, dummy-based location privacy in mobile services," in *Proc. 7th ACM Int. Workshop Data Eng. Wireless Mobile Access*, Vancouver, BC, Canada, 2008, pp. 16–23.

[39] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *J. Amer. Statist. Assoc.*, vol. 60, no. 309, pp. 63–69, 1965.

[40] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, "Quantifying location privacy," in *Proc. IEEE Symp. Secur. Privacy (SP)*, Berkeley, CA, USA, 2011, pp. 247–262.

[41] H. Li, H. Zhu, S. Du, X. Liang, and X. Shen, "Privacy leakage of location sharing in mobile social networks: Attacks and defense," *IEEE Trans. Dependable Secure Comput.*, vol. 15, no. 4, pp. 646–660, Jul./Aug. 2016.

[42] A.-M. Olteanu, K. Huguenin, R. Shokri, M. Humbert, and J.-P. Hubaux, "Quantifying interdependent privacy risks with location data," *IEEE Trans. Mobile Comput.*, vol. 16, no. 3, pp. 829–842, Mar. 2016.

[43] X. Zhang, X. Gui, F. Tian, S. Yu, and J. An, "Privacy quantification model based on the Bayes conditional risk in location-based services," *Tsinghua Sci. Technol.*, vol. 19, no. 5, pp. 452–462, Oct. 2014.
[44] K. Kalantari, L. Sankar, and O. Kosut, "On information-theoretic privacy with general distortion cost functions," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jul. 2017, pp. 2865–2869.
[45] S. Salamatian *et al.*, "How to hide the elephant- or the donkey- in the room: Practical privacy against statistical inference for large data," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Austin, TX, USA, Dec. 2013, pp. 269–272.
[46] I. Csiszár, "Almost independence and secrecy capacity," *Problemy Peredachi Informatsii*, vol. 32, no. 1, pp. 48–57, 1996.
[47] F. P. Calmon, A. Makhdoumi, and M. Médard, "Fundamental limits of perfect privacy," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Hong Kong, China, Jun. 2015, pp. 1796–1800.
[48] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 6, pp. 838–852, Jun. 2013.
[49] J. Liao, L. Sankar, F. P. Calmon, and V. Y. F. Tan, "Hypothesis testing under maximal leakage privacy constraints," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Aachen, Germany, Jun. 2017, pp. 779–783.
[50] H. Yamamoto, "A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers (Corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 6, pp. 918–923, Nov. 1983.
[51] J. Liao, L. Sankar, V. Y. F. Tan, and F. P. Calmon, "Hypothesis testing in the high privacy limit," in *Proc. 54th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Monticello, IL, USA, Sep. 2016, pp. 649–656.
[52] Z. Montazeri, A. Houmansadr, and H. Pishro-Nik, "Achieving perfect location privacy in wireless devices using anonymization," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 11, pp. 2683–2698, Nov. 2017.
[53] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Fundamental limits of location privacy using anonymization," in *Proc. 51st Annu. Conf. Inf. Sci. Syst. (CISS)*, Baltimore, MD, USA, Mar. 2017, pp. 1–6.

Nazanin Takbiri received her B.S. degree from University of Tehran, Tehran, Iran, in 2012 and M.S. degree from Boğaziçi University, Istanbul, Turkey, in 2016. She is currently working toward the Ph.D. degree in Electrical Engineering at the University of Massachusetts Amherst, Amherst, MA, USA. Her research interests include Privacy & Security issues with focus on IoT privacy.

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

Dennis L. Goeckel (F'11) received the B.S. from Purdue University in 1992 and the M.S. and Ph.D. degrees from the University of Michigan in 1993 and 1996, respectively. Since 1996, he has been with the Electrical and Computer Engineering Department, University of Massachusetts at Amherst, where he is currently a Professor. He was a Lilly Teaching Fellow from 2000 to 2001. He received the NSF CAREER Award in 1999 and the University of Massachusetts Distinguished Teaching Award in 2007. He has served on the Editorial Boards of a number of international journals in communications and networking, including the IEEE TRANSACTIONS ON NETWORKING, the IEEE TRANSACTIONS ON MOBILE COMPUTING, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and the IEEE TRANSACTIONS ON COMMUNICATIONS.

Hossein Pishro-Nik received the B.S. degree from Sharif University of Technology, and the M.Sc. and Ph.D. degrees from the Georgia Institute of Technology, all in electrical and computer engineering. He is currently an Associate Professor of electrical and computer engineering with the University of Massachusetts at Amherst, Amherst. His research interests include information theoretic privacy and 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.