Asymptotic Limits of Privacy in Bayesian Time Series Matching

Nazanin Takbiri
Electrical and
Computer Engineering
UMass-Amherst
ntakbiri@umass.edu

Dennis L. Goeckel
Electrical and
Computer Engineering
UMass-Amherst
goeckel@ecs.umass.edu

Amir Houmansadr
Information and
Computer Sciences
UMass-Amherst
amir@cs.umass.edu

Hossein Pishro-Nik
Electrical and
Computer Engineering
UMass-Amherst
pishro@ecs.umass.edu

Abstract—**Various modern and highly popular applications make use of user data traces in order to offer specific services, often for the purpose of improving the user's experience while using such applications. However, even when user data is privatized by employing privacy-preserving mechanisms (PPM), users' privacy may still be compromised by an external party who leverages statistical matching methods to match users' traces with their previous activities. In this paper, we obtain the theoretical bounds on user privacy for situations in which user traces are matchable to sequences of prior behavior, despite anonymization of data time series. We provide both achievability and converse results for the case where the data trace of each user consists of independent and identically distributed (i.i.d.) random samples drawn from a multinomial distribution, as well as the case that the users' data points are dependent over time and the data trace of each user is governed by a Markov chain model.**

Index Terms—**Anonymization, information theoretic privacy, Internet of Things (IoT), Markov chain model, statistical matching, Privacy-Preserving Mechanism (PPM).**

I. INTRODUCTION

THE Internet of Things (IoT) is an important emerging technology and is growing at a rapid pace: by 2020, over 50 billion devices will be connected together as part of the IoT network [1]. Environmental monitoring, infrastructure management, energy management, medical and healthcare systems, building and home automation, and transport systems are some examples which indicate that IoT devices will affect nearly every aspect of our daily lives. However, this ubiquity of impact also raises grave privacy concerns. In particular, each IoT user in each application is generating a sequence of data that can be modeled as a random process; for example, in location-based services, each user is generating location traces. These sequences of data in IoT systems often contain sensitive information about users, such as their locations, health information, and hobbies. As a result, such huge amount of data generated by IoT devices can critically damage users' privacy, thereby providing a significant obstacle to the adaption of IoT applications. Thus, IoT privacy has drawn the attention of the research community [2]–[4] to investigate effective privacy-preserving mechanisms (PPMs).

PPMs are used to increase the assurance that private data is not accessible to third parties. Two promising classes of

PPMs are *identity perturbation* and *data perturbation* [5]–[12]. The identity perturbation technique or anonymization is the process of hiding the true identity of the data owner [5]–[9]. This technique removes personal identifiers or converts personally identifiable information into aggregated data. The data perturbation or obfuscation is the process of hiding the users' data by adding noise [10]–[12]. However, perturbation techniques reduce utility to provide better privacy protection; thus, obtaining the optimum levels of anonymization and obfuscation is important.

In [7], [13], a comprehensive analysis of the asymptotic (in the length of the time series) optimal matching of time series to source distributions is presented in a non-Bayesian setting, where the number of users is a fixed, finite value. However, in [14]–[20], a Bayesian setting was adopted in which the adversary has accurate prior distributions for user behavior through past observations or other sources, and the asymptotic limits of user privacy were obtained.

In addition, Li et al. [21] provide an optimal hypothesis test in the case where the adversary has training sequences from the group of users rather than the exact probability distribution.

In this paper, we adopt the same setting as [21]; however, our work has significantly different flavor than that of [21]. First, [21] finds the optimal test in the non-asymptotic regime where there exist two users, while here, the asymptotic limits of user privacy for the case of a large number of users are obtained. Second, [21] obtains the necessary conditions for breaking privacy, while here, conditions for both perfect anonymity and no privacy are obtained. Third, [21] establishes the optimal test for the case with binary alphabets where each user's trace consists of independent and identically distributed (i.i.d.) samples drawn from a Bernoulli distribution, while here, we extend our results to the case where each user's trace is governed by i.i.d. random samples of a multinoulli distribution. We also extend our results for a more general Markov chain model.

The remainder of this paper is organized as follows. Section II discusses the system model and the metrics used in the paper. Achievability and converse results for the two-state i.i.d. model are presented in Section III, and their extensions to the r-state i.i.d. model are presented in Section IV. In addition, achievability and converse results for a more general Markov

chain model are presented in Section V. Section VI provides some final conclusions and directions for future work.

II. FRAMEWORK

We assume a system with n users. Each user creates a length-m sequence of data, which is denoted by $\mathbf{X}_u$,

$$\mathbf{X}_u = \left[X_u(1), X_u(2), \cdots X_u(m) \right]^T, \quad \mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n],$$

where $X_u(k)$ is the actual data point of user u at time k. For each user, there also exists a length-l sequence of its past behavior which is denoted as $\mathbf{W}_u$,

$$\mathbf{W}_u = \left[W_u(1), W_u(2), \cdots W_u(l) \right]^T, \quad \mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_n].$$

where $W_u(k)$ is the observation of the prior behavior of user u at time k.

The adversary has access to the observations of the prior users' behavior and wants to use this knowledge to break users' privacy despite the usage of some PPMs. As shown in Figure 1, an anonymization technique is employed in order to perturb the users' identity before the data is provided to the IoT application. In this figure, $Y_u(k)$ is the reported data point of user u at time k after applying anonymization; hence, the adversary observes

$$\mathbf{Y}_u = \left[Y_u(1), Y_u(2), \cdots Y_u(m) \right]^T, \quad \mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n].$$

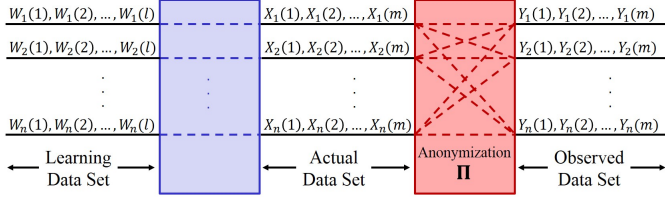where $\mathbf{Y}$ is the permuted version of $\mathbf{X}$.

Fig. 1: The goal of the adversary: match each sequence of $\mathbf{W}_u$ of user $u \in \{1, 2, \cdots, n\}$ to an observed sequence $\mathbf{Y}_u$ for $u \in \{1, 2, \cdots, n\}$.

A. Models and Metrics

Data Points Model: We assume there exist r possible values $\{0, 1, \cdots, r-1\}$ for each data point. As shown in Figure 1, there exist two traces for each user: one that is termed "training data" and one that is termed "actual data," which needs to be protected from a malicious adversary. Remember that these two traces are generated from the same unknown probability distribution. In other words, for $k \in \{1, 2, \cdots, m\}$ and $k' \in \{1, 2, \cdots, l\}$, both $X_u(k)$ and $W_u(k')$ are drawn from a user-specific probability distribution denoted as $\mathbf{p}_u$. While all $\mathbf{p}_u$'s are unknown to the adversary, each of them is drawn independently from a continuous density function $f_{\mathbf{P}}(\mathbf{x})$, where for all x in the support of $f_{\mathbf{P}}(\mathbf{x})$, we assume

$$0 < \delta_1 < f_{\mathbf{P}}(\mathbf{x}) < \delta_2 < \infty.$$

Anonymization Mechanism: As shown in Figure 1, the mapping between users and data sequences is randomly permuted in order to achieve privacy. This random permutation is chosen uniformly at random among all $n!$ possible permutations on the set of n users $\left(\mathbf{\Pi} : \{1, 2, \cdots, n\} \mapsto \{1, 2, \cdots, n\} \right)$; then, $\mathbf{Y}_u = \mathbf{X}_{\Pi^{-1}}$, $\mathbf{Y}_{\Pi(u)} = \mathbf{X}_u$.

Adversary Model: The adversary tries to match each sequence in the collection of training data traces $\{\mathbf{W}_u, u = 1, 2, \cdots, n\}$ with the sequence in the observation data traces $\{\mathbf{Y}_u, u = 1, 2, \cdots, n\}$ that is drawn from the same probability distribution, which we term *statistical matching*. This is equivalent to finding the permutations of the user identities between two collections. Note that the adversary knows the anonymization mechanism; however, he/she does not know the realization of the random permutation function.

Following [17], the definition of no privacy is as follows:

Definition 1. For an algorithm of the adversary that tries to estimate the actual data point 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 $X_u(k)$ is the actual data point of user u at time k, and $\widetilde{X_u(k)}$ is the adversary's estimated data point of user u at time k. Now, define $\mathcal{E}$ as the set of all possible estimators of the adversary. 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.

In this paper, we also consider the situation in which there is perfect anonymity.

Definition 2. User u has *perfect anonymity* at time k if and only if

$$\lim_{n \to \infty} H \left(\Pi(1) | \mathbf{W}, \mathbf{Y} \right) \to +\infty,$$

where $H \left(\Pi(1) | \mathbf{W}, \mathbf{Y} \right)$ is the entropy of $\Pi(1)$ given $\mathbf{W}$ and $\mathbf{Y}$.

III. TWO-STATE I.I.D. MODEL

In this section, we assume each user's trace consists of samples from an i.i.d. random process and there are only two possible values for each user data point $X_u(k) \in \{0, 1\}$. Thus, both training traces and real data traces are governed by an i.i.d. Bernoulli distribution with parameter p_u, where p_u is probability that user u taking value of a 1, hence,

$$W_u(k) \sim Bernoulli \left(p_u \right),$$

and

$$X_u(k) \sim Bernoulli \left(p_u \right), \quad Y_u(k) \sim Bernoulli \left(p_{\Pi(u)} \right).$$

As discussed in Section II, while p_u's are unknown to the adversary, they are drawn independently from a known continuous density function ($f_P(x)$), where for all $x \in (0, 1)$, we have

$$0 < \delta_1 < f_P(x) < \delta_2 < \infty. \tag{1}$$

A. Perfect Anonymity Analysis

The following theorem states that if m or l are significantly smaller than n^2 in this two-state model, then all users have perfect anonymity.

Theorem 1. For the above two-state i.i.d. model, if $\mathbf{Y}$ is the anonymized version of $\mathbf{X}$, and $\mathbf{W}$ is the past behavior of the users as defined above, and
- at least one of m or l is less than or equal to $cn^{2-\alpha}$ for any $c, \alpha > 0$;

then, user 1 has perfect anonymity at time k.

Proof. First, consider the case $m \leq l$. Here, $\mathbf{W}$ is considered as the training set and $\mathbf{Y}$ is considered as the observed set; thus, given $\mathbf{Y}$, $\mathbf{W} \to \mathbf{P} \to \Pi(1)$ forms a Markov chain. According to the data processing inequality,

$$I\left(\Pi(1); \mathbf{W}|\mathbf{Y}\right) \leq I\left(\Pi(1); \mathbf{P}|\mathbf{Y}\right);$$

thus,

$$H(\Pi(1)|\mathbf{Y}) - H\left(\Pi(1)|\mathbf{W}, \mathbf{Y}\right) \leq H(\Pi(1)|\mathbf{Y}) - H\left(\Pi(1)|\mathbf{P}, \mathbf{Y}\right),$$

and

$$H\left(\Pi(1)|\mathbf{W}, \mathbf{Y}\right) \geq H\left(\Pi(1)|\mathbf{P}, \mathbf{Y}\right).$$

In [15, Theorem 1], it is shown that if $m = n^{2-\alpha}$, $H\left(\Pi(1)|\mathbf{P}, \mathbf{Y}\right) \to +\infty$, so, we can conclude

$$H\left(\Pi(1)|\mathbf{W}, \mathbf{Y}\right) \to +\infty,$$

as $n \to \infty$.

Now, consider the case $l \leq m$. By symmetry of the problem $\mathbf{Y}$ can be considered as the training set and $\mathbf{W}$ can be considered as the observed data. Thus, we can similarly prove the same results. □

B. No Privacy Analysis

The following theorem states that if both m and l are significantly larger than n^2 in this two-state model, then the adversary can find an algorithm to successfully estimate users' data points with arbitrarily small error probability, and as a result break users' privacy.

Theorem 2. For the above two-state i.i.d. model, if $\mathbf{Y}$ is the anonymized version of $\mathbf{X}$, and $\mathbf{W}$ is the past behavior of the users as defined above, and
- $m = cn^{2+\alpha}$ for any $c, \alpha > 0$;
- $l = c'n^{2+\alpha}$ for any $c', \alpha > 0$;

then, user 1 has no privacy at time k.

Proof. For $u \in \{1, 2, \cdots n\}$, define

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

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

and

$$\overline{W_u} = \frac{W_u(1) + W_u(2) + \cdots + W_u(l)}{l}.$$

We claim that for $m = cn^{2+\alpha}$, $l = c'n^{2+\alpha}$ and large enough n:

1) $\mathbb{P}\left(\left|\overline{Y_{\Pi(1)}} - \overline{W_1}\right| \leq \Delta_n\right) \to 1,$

2) $\mathbb{P}\left(\bigcup_{u=2}^{n}\left\{\left|\overline{Y_{\Pi(u)}} - \overline{W_1}\right| \leq \Delta_n\right\}\right) \to 0,$

where $\Delta_n = n^{-(1+\frac{\alpha}{4})}$. Thus, the adversary can match $\mathbf{W}_1$ to $\mathbf{Y}_{\Pi(1)}$.

First Step: We want to show

$$\mathbb{P}\left(\left|\overline{X_u} - \overline{W_u}\right| \leq \Delta_n\right) \to 1.$$

Note $\mathbb{E}[X_u(k)] = \mathbb{E}[W_u(k)] = p_u$, so as $n \to \infty$,

$$\begin{aligned}
\mathbb{P}\left(\left|\overline{X_u} - \overline{W_u}\right| \geq \Delta_n\right) &= \mathbb{P}\left(\left|\overline{X_u} - p_u - \overline{W_u} + p_u\right| \geq \Delta_n\right) \\
&\leq \mathbb{P}\left(\left|\overline{X_u} - p_u\right| + \left|\overline{W_u} - p_u\right| \geq \Delta_n\right) \\
&\leq \mathbb{P}\left(\left\{\left|\overline{X_u} - p_u\right| \geq \frac{\Delta_n}{2}\right\} \bigcup \left\{\left|\overline{W_u} - p_u\right| \geq \frac{\Delta_n}{2}\right\}\right) \\
&\leq \mathbb{P}\left(\left|\overline{X_u} - p_u\right| \geq \frac{\Delta_n}{2}\right) + \mathbb{P}\left(\left|\overline{W_u} - p_u\right| \geq \frac{\Delta_n}{2}\right) \\
&\leq 2e^{-\frac{m\Delta_n^2}{12p_u}} + 2e^{-\frac{l\Delta_n^2}{12p_u}} \\
&= 2e^{-\frac{cn^{2+\alpha} \cdot n^{-2-\frac{\alpha}{2}}}{12p_u}} + 2e^{-\frac{c'n^{2+\alpha} \cdot n^{-2-\frac{\alpha}{2}}}{12p_u}} \\
&= 2e^{-\frac{cn^{\frac{\alpha}{2}}}{12}} + 2e^{-\frac{c'n^{\frac{\alpha}{2}}}{12}} \to 0, \tag{2}
\end{aligned}$$

where the first inequality follows from the fact that $|a - b| \leq |a| + |b|$, and as a result, $\mathbb{P}\left(|a - b| \geq \Delta_n\right) \leq \mathbb{P}\left(|a| + |b| \geq \Delta_n\right)$. The union bound yields the third inequality, and the fourth inequality follows from Chernoff bounds. Now, for u=1, we have

$$\mathbb{P}\left(\left|\overline{Y_{\Pi(1)}} - \overline{W_1}\right| \leq \Delta_n\right) \to 1,$$

as $n \to \infty$.

Second Step: First, we show as $n \to \infty$,

$$\mathbb{P}\left(\bigcup_{u=2}^{n}\left\{\left|p_u - p_1\right| \leq 4\Delta_n\right\}\right) \to 0.$$

According to (1), for all $u \in \{2, 3, \cdots, n\}$, we have

$$\mathbb{P}\left(\left|p_u - p_1\right| \leq 4\Delta_n\right) \leq 8\Delta_n\delta_2,$$

and according to the union bound,

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{u=2}^{n}\left\{\left|p_u - p_1\right| \leq 4\Delta_n\right\}\right) &\leq \sum_{u=2}^{n}\mathbb{P}\left(\left|p_u - p_1\right| \leq 4\Delta_n\right) \\
&\leq 8n\Delta_n\delta_2 \\
&= 8n^{-\frac{\alpha}{4}}\delta_2 \to 0,
\end{aligned}$$

as $n \to \infty$. Thus, for $u \in \{2, 3, \cdots, n\}$, the distance between p_u and p_1 is bigger than $4\Delta_n$ with high probability.

Next, we show as $n \to \infty$,

$$\mathbb{P}\left(\bigcup_{u=2}^{n}\left\{\left|\overline{W_u} - \overline{W_1}\right| \le 2\Delta_n\right\}\right) \to 0.$$

Note for all $u \in \{1, 2, \cdots, n\}$, Chernoff bounds yields:

$$\mathbb{P}\left(\left|\overline{W_u} - p_u\right| \ge \Delta_n\right) \le 2e^{-\frac{l\Delta_n^2}{3p_u}} \le 2e^{-\frac{l\Delta_n^2}{3}}. \tag{3}$$

As a result, for $u = 1$, we have

$$\mathbb{P}\left(\left|\overline{W_1} - p_1\right| \ge \Delta_n\right) \le 2e^{-\frac{c'n^{\frac{\alpha}{2}}}{3}} \to 0,$$

as $n \to \infty$. In other words, with high probability, the distance between $\overline{W_1}$ and p_1 is less than Δ_n.

Now, given the fact that the distance between all p_u's and p_1 is bigger than $4\Delta_n$, and the fact that the distance between $\overline{W_1}$ and p_1 is less than Δ_n, for all $u \in \{2, 3, \cdots, n\}$, we have

$$\mathbb{P}\left(\left|\overline{W_u} - \overline{W_1}\right| \le 2\Delta_n\right) \le \mathbb{P}\left(\left|\overline{W_u} - p_u\right| \ge \Delta_n\right)$$

$$\le 2e^{-\frac{l\Delta_n^2}{3}}.$$

Thus,

$$\mathbb{P}\left(\bigcup_{u=2}^{n}\left\{\left|\overline{W_u} - \overline{W_1}\right| \le 2\Delta_n\right\}\right) \le \sum_{u=2}^{n}\mathbb{P}\left(\left|\overline{W_u} - \overline{W_1}\right| \le 2\Delta_n\right)$$

$$\le 2ne^{-\frac{l\Delta_n^2}{3}}$$

$$= 2ne^{-\frac{c'n^{\frac{\alpha}{2}}}{3}} \to 0,$$

as $n \to \infty$.

Now, we claim that given the fact that the distances between each of the $\overline{W_u}$'s and $\overline{W_1}$ are bigger than $2\Delta_n$, we have

$$\mathbb{P}\left(\bigcup_{u=2}^{n}\left\{\left|\overline{X_u} - \overline{W_1}\right| \le \Delta_n\right\}\right) \to 0.$$

Note, using (2), we have

$$\mathbb{P}\left(\left|\overline{X_u} - \overline{W_1}\right| \le \Delta_n\right) = \mathbb{P}\left(\left|\overline{X_u} - \overline{W_u}\right| \ge \Delta_n\right)$$

$$\le 2e^{-\frac{cn^{\frac{\alpha}{2}}}{12}} + 2e^{-\frac{c'n^{\frac{\alpha}{2}}}{12}}.$$

Thus, by using union bound, we have

$$\mathbb{P}\left(\bigcup_{u=2}^{n}\left\{\left|\overline{X_u} - \overline{W_1}\right| \le \Delta_n\right\}\right) \le \sum_{u=2}^{n}\mathbb{P}\left(\left|\overline{X_u} - \overline{W_u}\right| \ge \Delta_n\right)$$

$$\le 2ne^{-\frac{cn^{\frac{\alpha}{2}}}{12}} + 2ne^{-\frac{c'n^{\frac{\alpha}{2}}}{12}} \to 0,$$

as $n \to \infty$.

After completing the first and second steps, we can conclude if $m = cn^{2+\alpha}$ and $l = c'n^{2+\alpha}$, users have no privacy as $n \to \infty$. $\square$

IV. r-STATE I.I.D. MODEL

In this section, we assume each user's trace consists of samples from an i.i.d. random process, and users' data points can have r possibilities, where $X_u(k) \in \{0, 1, \cdots, r-1\}$. Thus, both training traces and real data traces are governed by an i.i.d. multinoulli distribution with parameter $\mathbf{p}_u$, and

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

where $p_u(i)$ is the probability that a datum of user u has value i.

As discussed in Section II, while $\mathbf{p}_u$'s are unknown to the adversary, they are drawn independently from a known continuous density function $f_{\mathbf{P}}(\mathbf{x})$, where for all $\mathbf{x} \in \mathcal{R}_{\mathbf{p}}$,

$$\mathcal{R}_{\mathbf{p}} = \Big\{(x_1, x_2, \cdots, x_{r-1}) \in (0, 1)^{r-1} :$$

$$x_i > 0, x_1 + x_2 + \cdots + x_{r-1} < 1, \ i = 1, 2, \cdots, r-1\Big\},$$

we have

$$0 < \delta_1 < f_{\mathbf{P}}(\mathbf{x}) < \delta_2 < \infty. \tag{4}$$

A. Perfect Anonymity Analysis

The following theorem states that if m or l are significantly smaller than $n^{\frac{2}{r-1}}$ in this r-state model, then all users have perfect anonymity.

Theorem 3. For the above r-state i.i.d. model, if $\mathbf{Y}$ is the anonymized version of $\mathbf{X}$, and $\mathbf{W}$ is the past behavior of the users as defined above, and

- at least one of m or l is less than or equal to $cn^{\frac{2}{r-1}-\alpha}$ for any $c, \alpha > 0$;

then, user 1 has perfect anonymity at time k.

Proof. We can now repeat the similar reasoning as Theorem 1; then, by using [15, Theorem 2], the proof is complete. $\square$

B. No Privacy Analysis

The following theorem states that if both m and l are significantly larger than $n^{\frac{2}{r-1}}$ in this r-state model, then the adversary can find an algorithm to successfully estimate users' data points with arbitrarily small error probability, and as a result break users' privacy.

Theorem 4. For the above r-state i.i.d. model, if $\mathbf{Y}$ is the anonymized version of $\mathbf{X}$, and $\mathbf{W}$ is the past behavior of the users as defined above, and

- $m = cn^{\frac{2}{r-1}+\alpha}$ for any $c, \alpha > 0$;
- $l = c'n^{\frac{2}{r-1}+\alpha}$ for any $c', \alpha > 0$;

then, user 1 has no privacy at time k.

Proof. The proof of Theorem 4 is similar to the proof of Theorem 2, so we just provide the general idea. We similarly define the empirical probability that the user with pseudonym u has data sample i as follows:

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

and

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

We also have

$$\overline{W_u(i)} = \frac{\left|\{k \in \{1, 2, \cdots, l\} : W_u(k) = i\}\right|}{l}.$$

The difference from the proof of Theorem 2 is that, for each $u \in \{1, 2, \cdots, n\}$, $\overline{\mathbf{Y}_u}$ and $\overline{\mathbf{W}_u}$ are vectors of length $r - 1$. In other words,

$$\overline{\mathbf{Y}_u} = \left[\overline{Y_u(1)}, \overline{Y_u(2)}, \cdots, \overline{Y_u(r-1)}\right]^T, \quad u \in \{1, 2, \cdots, n\},$$

$$\overline{\mathbf{W}_u} = \left[\overline{W_u(1)}, \overline{W_u(2)}, \cdots, \overline{W_u(r-1)}\right]^T, \quad u \in \{1, 2, \cdots, n\},$$

and we claim for $m = cn^{\frac{2}{r-1}+\alpha}$, $l = c'n^{\frac{2}{r-1}+\alpha}$, and large enough n,

1) $\mathbb{P}\left(\left|\overline{\mathbf{Y}_{\Pi(1)}} - \overline{\mathbf{W}_1}\right| \leq \Delta'_n\right) \to 1,$

2) $\mathbb{P}\left(\bigcup_{u=2}^{n} \left\{\left|\overline{\mathbf{Y}_{\Pi(u)}} - \overline{\mathbf{W}_1}\right| \leq \Delta'_n\right\}\right) \to 0,$

where $\Delta'_n = n^{-\left(\frac{1}{r-1}+\frac{\alpha}{4}\right)}$. $\quad\square$

V. r-STATE MARKOV CHAIN MODEL

In Section III and IV, the data trace of each user is governed by an i.i.d. random process, while here the data trace of each user is governed by an irreducible and aperiodic r-state Markov chain where E is the set of edges. Let us define the transition probability from state i to state j as:

$$p_u(i, j) = \mathbb{P}\left(X_u(k+1) = j | X_u(k) = i\right);$$

thus, $(i, j) \in E$ if and only if $p_u(i, j) > 0$.

Here, we assume the same Markov chain structure for all of the users, but different users have different transition matrices. Note that a subset of the transition probabilities with size $|E| - r$ is sufficient for recovering the whole transition matrix. Let this subset be called $\mathbf{p}_u$, so

$$\mathbf{p}_u = \left[p_u(1), p_u(2), \cdots p_u(|E|-r)\right]^T, \quad \mathbf{p} = \left[\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n\right].$$

where $p_u(i)$ is the probability that a datum of user u has value i. As discussed in Section II, while $\mathbf{p}_u$'s are unknown to the adversary, they are drawn independently from a known continuous density function $f_{\mathbf{P}}(\mathbf{x})$, where for all $\mathbf{x} \in \mathcal{R}_{\mathbf{p}}$,

$$\mathcal{R}_{\mathbf{p}} = \Big\{(x_1, x_2, \cdots, x_{|E|-r}) \in (0, 1)^{|E|-r} :$$

$$x_i > 0, x_1 + x_2 + \cdots + x_{|E|-r} < 1, \ i = 1, 2, \cdots, |E| - r\Big\},$$

we have

$$0 < \delta_1 < f_{\mathbf{P}}(\mathbf{x}) < \delta_2 < \infty. \tag{5}$$

A. Perfect Anonymity Analysis

The following theorem states that if m or l are significantly smaller than $n^{\frac{2}{|E|-r}}$ in this r-state Markov chain model, then all users have perfect anonymity.

Theorem 5. For the above r-state Markov chain model, if $\mathbf{Y}$ is the anonymized version of $\mathbf{X}$, and $\mathbf{W}$ is the past behavior of the users as defined above, and

- at least one of m or l is less than or equal to $cn^{\frac{2}{|E|-r}-\alpha}$ for any $c, \alpha > 0$;

then, user 1 has perfect anonymity at time k.

Proof. We can now repeat the similar reasoning as Theorem 1; then, by using [15, Theorem 3], the proof is complete. $\quad\square$

B. No Privacy Analysis

The following theorem states that if both m and l are significantly larger than $n^{\frac{2}{|E|-r}}$, then the adversary can find an algorithm to successfully estimate users' data points with arbitrarily small error probability, and as a result, break users' privacy.

Theorem 6. For the above r-state Markov chain model, if $\mathbf{Y}$ is the anonymized version of $\mathbf{X}$, and $\mathbf{W}$ is the past behavior of the users as defined above, and

- $m = cn^{\frac{2}{|E|-r}+\alpha}$ for any $c, \alpha > 0$;
- $l = c'n^{\frac{2}{|E|-r}+\alpha}$ for any $c', \alpha > 0$;

then, user 1 has no privacy at time k.

Proof. The proof of Theorem 6 is similar to the proof of Theorem 2, so we just provide the general idea. For each $u \in \{1, 2, \cdots, n\}$, we similarly define $\overline{\mathbf{Y}_u}$ and $\overline{\mathbf{W}_u}$ as vectors of length $|E| - r$:

$$\overline{\mathbf{Y}_u} = \left[\overline{Y_u(1)}, \overline{Y_u(2)}, \cdots, \overline{Y_u(|E|-r)}\right]^T, \quad u \in \{1, 2, \cdots, n\}.$$

$$\overline{\mathbf{W}_u} = \left[\overline{W_u(1)}, \overline{W_u(2)}, \cdots, \overline{W_u(|E|-r)}\right]^T, \quad u \in \{1, 2, \cdots, n\}.$$

We claim that for $m = cn^{\frac{2}{|E|-r}+\alpha}$, $l = c'n^{\frac{2}{|E|-r}+\alpha}$, and large enough n,

1) $\mathbb{P}\left(\left|\overline{\mathbf{Y}_{\Pi(1)}} - \overline{\mathbf{W}_1}\right| \leq \Delta''_n\right) \to 1,$

2) $\mathbb{P}\left(\bigcup_{u=2}^{n} \left\{\left|\overline{\mathbf{Y}_{\Pi(u)}} - \overline{\mathbf{W}_1}\right| \leq \Delta''_n\right\}\right) \to 0,$

where $\Delta''_n = n^{-\left(\frac{1}{|E|-r}+\frac{\alpha}{4}\right)}$. $\quad\square$

VI. CONCLUSION

In this paper, we have derived the theoretical bounds on user privacy in situations in which user traces are matchable to prior user behavior despite anonymization protection. In particular, the adversary employs statistical matching of the user traces to previous behavior of users within a network to compromise their privacy.

As shown in Figure 2, which displays the characterized privacy limits for the i.i.d. case, we demonstrated that the parameter plane, with coordinates length of learning set (l) and

length of observed set (m), can be divided into two regions: in the first region, all users have perfect anonymity and in the second region no user has any privacy whatsoever. Specifically, we showed that if either l or m is significantly smaller than $n^{\frac{2}{r-1}}$, users have perfect anonymity and the adversary cannot identify the permutation function (**II**), and, if both of them are significantly larger than $n^{\frac{2}{r-1}}$, users have no privacy. It is worth noting that in the case the adversary has the accurate prior information, which is discussed in [15], [16] and is shown in Figure 3, users have no privacy as long as number of adversary observations per user m is larger than $n^{\frac{2}{r-1}}$.

For the case where the users' data points are governed by an irreducible and aperiodic r-state Markov chain with $|E|$ edges, we demonstrated similar results: if either l or m is significantly smaller than $n^{\frac{2}{|E|-r}}$, users have perfect anonymity, and, if both of them are significantly larger than $n^{\frac{2}{|E|-r}}$, users have no privacy.
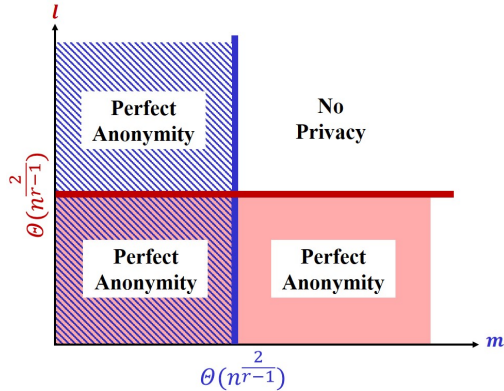

Fig. 2: Limits of privacy in the entire $m-l$ plane in the case the adversary does not have the accurate prior distribution. Here, both training data traces and observed data traces are governed by an i.i.d. multinoulli distribution. l is the length of the learning set, m is the length of the observed data, and r is the number of possible values for each user's data point.
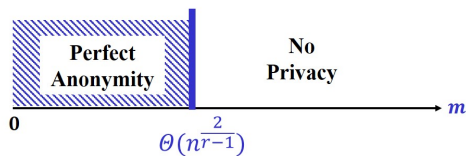

Fig. 3: Limits of privacy in the case the adversary has an accurate prior distribution. Here, the observed data traces are governed by an i.i.d. multinoulli distribution, m is the length of the observed data and r is the number of possible values for each user's data point.

REFERENCES

[1] J. Bausch. (2016) The internet of things forecast of 50 billion connected devices by 2020 is grossly over-estimated and entirely misleading. [Online]. Available: https://www.electronicproducts.com/Internet_of_Things/Research/ The_Internet_of_Things_forecast_of_50_billion_connected_devices_ by_2020_is_grossly_over_estimated_and_entirely_misleading.aspx

[2] Federal Trade Commission Staff, "Internet of things: Privacy and security in a connected world," 2015.

[3] A. Ukil, S. Bandyopadhyay, and A. Pal, "IoT-privacy: To be private or not to be private," in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. Toronto, ON, Canada: IEEE, 2014, pp. 123–124.

[4] S. Hosseinzadeh, S. Rauti, S. Hyrynsalmi, and V. Leppänen, "Security in the internet of things through obfuscation and diversification," in *IEEE Conference on Computing, Communication and Security (ICCCS)*. Pamplemousses, Mauritius: IEEE, 2015, pp. 1–5.

[5] B. Hoh and M. Gruteser, "Protecting location privacy through path confusion," in *First International Conference on Security and Privacy for Emerging Areas in Communications Networks (SecureComm)*. Pamplemousses, Mauritius: IEEE, 2005, pp. 194–205.

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

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

[8] R. Soltani, D. Goeckel, D. Towsley, and A. Houmansadr, "Towards provably invisible network flow fingerprints," in *51th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, 2017.

[9] R. Soltani, D. Goeckel, D. F. Towsley, and A. Houmansadr, "Fundamental limits of invisible flow fingerprinting," *CoRR*, vol. abs/1809.08514, 2018.

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

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

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

[13] J. Unnikrishnan, "Asymptotically optimal matching of multiple sequences to source distributions and training sequences," *IEEE Transactions on Information Theory*, vol. 61, no. 1, pp. 452–468, 2014.

[14] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Limits of locatin privacy under anonymization and obfuscation," in *International Symposium on Information Theory (ISIT)*. Aachen, Germany: IEEE, 2017, pp. 764–768.

[15] Z. Montazeri, A. Houmansadr, and H. Pishro-Nik, "Achieving Perfect Location Privacy in Wireless Devices Using Anonymization," *IEEE Transaction on Information Forensics and Security*, vol. 12, no. 11, pp. 2683–2698, 2017.

[16] N. Takbiri, A. Houmansadr, D. Goeckel, and H. Pishro-Nik, "Fundamental limits of location privacy using anonymization," in *51st Annual Conference on Information Science and Systems (CISS)*. Baltimore, MD, USA: IEEE, 2017.

[17] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Matching anonymized and obfuscated time series to users' profiles," *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 724–741, 2019.

[18] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Privacy against statistical matching: Inter-user correlation," in *International Symposium on Information Theory (ISIT)*. Vail, Colorado, USA: IEEE, 2018, pp. 1036–1040.

[19] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Privacy of dependent users against statistical matching," *submitted to IEEE Transactions on Information Theory, Available at https://arxiv.org/abs/1710.00197*.

[20] N. Takbiri, R. Soltani, D. Goeckel, A. Houmansadr, and H. Pishro-Nik, "Asymptotic loss in privacy due to dependency in gaussian traces," in *IEEE Wireless Communications and Networking Conference (WCNC)*. Marrakech, Morocco: IEEE, 2019.

[21] K. Li, H. Pishro-Nik, and D. Goeckel, "Bayesian time series matching and privacy," in *51th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, 2017.