Privacy of Dependent Users Against Statistical Matching

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

***Abstract*—Modern applications significantly enhance user experience by adapting to each user's individual condition and/or preferences. While this adaptation can greatly improve a user's experience or be essential for the application to work, the exposure of user data to the application presents a significant privacy threat to the users—even when the traces are anonymized—since the statistical matching of an anonymized trace to prior user behavior can identify a user and their habits. Because of the current and growing algorithmic and computational capabilities of adversaries, provable privacy guarantees as a function of the degree of anonymization and obfuscation of the traces are necessary. Our previous work has established the requirements on anonymization and obfuscation in the case that data traces are independent between users. However, the data traces of different users will be dependent in many applications, and an adversary can potentially exploit such. In this paper, we consider the negative impact of dependency between user traces on their privacy. First, we demonstrate that the adversary can readily identify the association graph of the obfuscated and anonymized version of the data, revealing which user data traces are dependent. Next, we demonstrate that the adversary can use this association graph to break user privacy with significantly shorter traces than in the case of independent users, and that obfuscating data traces independently across users is often insufficient to remedy such leakage. In other words, we have shown that inter-user dependency is disastrous to privacy, and any non-negligible dependency between users significantly reduces the effectiveness of anonymization and obfuscation schemes. Finally, we discuss how users can improve privacy by employing joint obfuscation that removes or reduces the data dependency.**

***Index Terms*—Information theoretic privacy, inter-user dependency, Internet of Things (IoT), obfuscation and anonymization, Privacy-Protection Mechanisms (PPM).**

I. Introduction

Many modern applications provide an enhanced user experience by exploiting users' characteristics, including their past choices and present states. In particular, emerging Internet of Things (IoT) applications include smart homes, healthcare, and connected vehicles that intelligently tailor their performance to their users. For instance, a typical connected vehicle application optimizes its route selection based on the current location of the vehicle, traffic conditions, and the users' preferences. For such applications to be able to provide their enhanced, user-tailored performances, they need to request their clients for potentially sensitive user information such as mobility behaviors and social preferences. Therefore, such applications trade off user privacy for enhanced utility. Previous work [2] shows that even if users' data traces are anonymized before being provided to such applications, standard statistical matching techniques can be used to leak users' private information. Thus, privacy and security threats are a major obstacle to the wide adoption of IoT applications, as demonstrated by prior studies [3]–[18].

The bulk of previous work assumes independence between the traces of different users. In [19]–[25], temporal and spatial dependencies within data traces are considered, but not cross-user dependency. In [19], an obfuscation technique is employed to achieve privacy; however, for continuous Location-Based Services (LBS) queries, there is often strong temporal dependency in the locations. Hence, [19] considers how temporal dependency of the users' obfuscated data can impact privacy, and then employs an adaptive noise level to improve privacy while still maintaining an acceptable level of utility. Liu et al. [21] show that the spatiotemporal dependency between neighboring location sets can ruin the privacy achieved using a dummy-based location-privacy preserving mechanism (LPPM); to solve this problem, they propose a spatiotemporal dependency-aware privacy protection that perturbs the spatiotemporal dependency between neighboring locations. Zhang et al. [20] employ Protecting Location Privacy (PLP) against dependency-analysis attack in crowd sensing: the potential dependency between users' data is modeled, and the data is filtered to remove the samples that disclose the user's private data. In [22], locations of a single user are temporally dependent, and δ-location set based differential privacy is proposed to achieve location privacy at every timestamp. Finally, Song et al. [26] provide privacy when there is dependency within the data of a single user. In summary, previous studies do not consider dependency between users, which is the focus of this work. We argue that for many applications, there is dependency between the traces of different users. For example, friends tend to travel together or might meet at given places, hence introducing dependency between the traces of their location information. Several previous works [27]–[33] have considered cross-user dependency; however, this only has been for protecting queries on aggregated data, which is different

This work was supported by National Science Foundation under grants CCF–1421957 and CNS–1739462. This work was presented in part in IEEE International Symposium on Information Theory (ISIT 2018) [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).

than our application scenario.

We use the notion of "perfect privacy" and "no privacy", as introduced in our prior work [34], [35], to evaluate the privacy of user traces. The "perfect privacy" notion provides an information-theoretic guarantee on privacy in the presence of a strong adversary who has complete knowledge on users' prior data traces. On the opposite extreme is the notion of 'no privacy". It means there exists an algorithm for the adversary to estimate the actual data points of users with diminishing error probability. Through a series of work [34]–[39], we have derived the degree of user anonymization and data obfuscation required to obtain perfect privacy—assuming that the data traces of different users are *independent* across users. Particularly, we evaluated the case of independent and identically distributed (i.i.d.) samples from a given user and the case when there is temporal dependency within the trace of a given user [35], [36] (but independent across users). In this work, we expand our study to the case where there is dependency between the data traces of different users. That is, we investigate how privacy is affected by the presence of dependency between the data traces of users when anonymization and obfuscation techniques are used. We show that dependency significantly reduces the privacy of users. Specifically, we show that the same anonymization and obfuscation levels that could produce perfect privacy for independent users result in no privacy for dependent users. Thus, in the presence of inter-user dependency, we need to employ much stronger anonymization and obfuscation compare to the case data traces of different users are independent.

We model dependency between user traces with an association graph, where the presence of an edge between the vertices corresponding to a pair of users indicates a non-zero dependency between their data traces. We employ standard concentration inequalities to demonstrate that the adversary can readily determine this association graph. Using this association graph and statistical data about the users, the adversary can attempt to identify users, and we demonstrate that this provides the adversary with a significant advantage versus the case when the data traces of different users are independent of one another. This suggests that, unless additional countermeasures are employed, the results of [34]–[37] for independent traces are optimistic when user traces are dependent. We next consider the effectiveness of countermeasures. First, we argue that adding independent obfuscation to user data points is often ineffective in improving the privacy of (dependent) users. Next, we demonstrate that, if users with dependent traces can jointly design their obfuscation, user privacy can be significantly improved.

A related but parallel approach to our study is graph alignment in which the edge set is sampled at random. Graph alignment is the problem of finding a matching between the vertices of the two graphs that matches, or aligns, many edges of the first graph with edges of the second graph. Shirani et. al. [40], [41] and Cullina et. al. [42] have done significant work on graph alignment. Although the graph alignment problem looks similar to our problem on the surface, there exist notable differences between the two. First, in Shirani et. al.'s work [40], [41], [43], graphs are generated using a model which is sampled at random from a probability distribution, while here the association graph is deterministic, as it is based on the dependency between data traces of users. Consequently, Shirani et. al. [40], [41], [43] employed a completely different approach and solution to de-anonymize users. In other words, they have not used the probability distribution of the data traces of each user to break anonymization, while here the probability distribution of the data trace of each user is a key characteristic which helps the adversary to break users' privacy. Finally, Shirani et. al. [40], [41] considered discrete values for the correlation between users and used them to de-anonymize the graph, while here the correlation between users have continuous values and the adversary does not have access to the exact value of them.

In [42], [44], [45], the graph alignment for two correlated graphs is considered, while here we assume the adversary has the association graph and tries to reconstruct it from the anonymized and obfuscated data traces. Thus, in our work, the adversary has two identical graphs and their goal is to identify all of the users based on the observed data and their statistical knowledge of users. Also, Cullina et. al. [42] considered fractional matching, while here the adversary can identify not only all of the users but also the data points of each user at all time with small error probability. Matching of non-identical pairs of correlated Erdös-Rényi graphs is studied in [46]–[49].

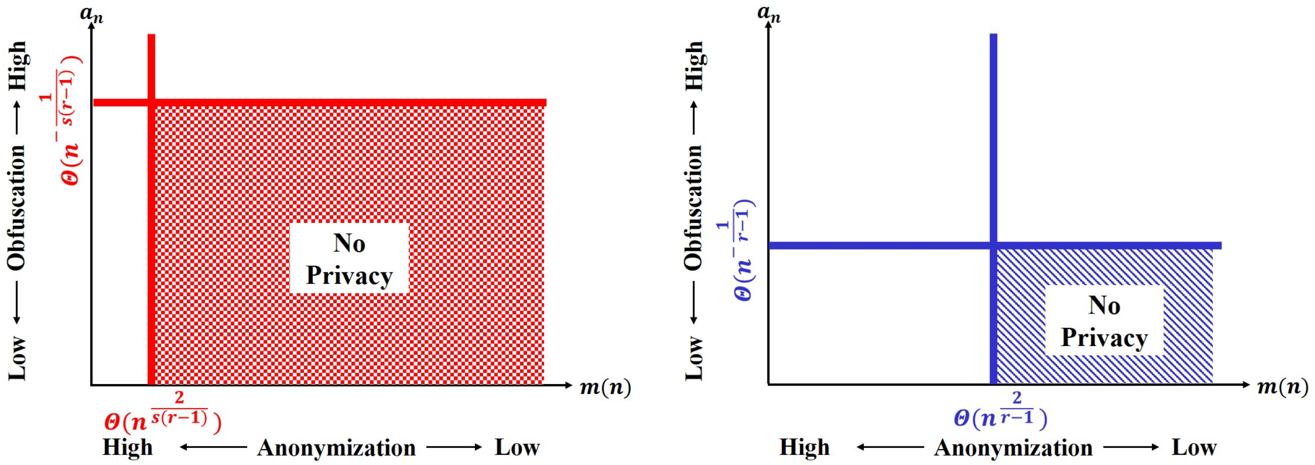
Also, graph isomorphism studied in [50]–[53] is an instance of the matching problem where the two graphs are identical copies of one another. Bollabás et al. [53] studied different algorithms such as maximum degree algorithms to match two identical graphs for the case where each edge of the graph has a fixed probability of being present or absent which is in the range of $\left[\omega\left(\log n/n^{\frac{1}{5}}\right), 1-\omega\left(\log n/n^{\frac{1}{5}}\right)\right]$, where n is the number of vertices in the graph. Here, the approach of our work is completely different, as the adversary uses probability distributions of users' data traces to reconstruct the association graph. After reconstruction of the association graph, the adversary uses the size of each disjoint group to identify all of the members.

In summary, although matching (alignment) between graphs can be considered as a part of our analysis, the analysis based on the users' data traces and the statistical knowledge of the adversary is a key part of this paper which distinguishes it from previous works on graph alignment.

The rest of this paper is organized as follows. In Section II we present the model and metrics considered in this work. In Sections III and IV, we show dependency between users' traces degrades privacy. In Section V, we discuss how our methodology can be applied to a more general setting for the association graph. In Section VI, we propose a method to improve privacy in the case when there exists inter-user dependency. Finally, Section VIII presents the conclusions and ideas for continuing work.

A. Summary of the Results

Consider a setting with n total users. As in our previous work [35], privacy depends on two parameters: (1) $m = m(n)$,



(a) The dependent case.

(b) The independent case.

Fig. 1: Representations of the no-privacy region in the case of dependent and independent users. Note that $m(n)$ is the number of the adversary's observations per user (degree of anonymization), and a_n is the amount of noise level (degree of obfuscation). Here, the size of the group of users whose data traces are dependent is equal to the finite value of s.

the number of data points after which the pseudonyms of users are changed in the anonymization technique, i.e., smaller m implies higher levels of anonymization; and (2) a_n, which indicates the amplitude of the obfuscation noise, i.e., larger a_n implies higher levels of obfuscation.

When there are a large number of users in the setting ($n \to \infty$) and each user's dataset is governed by an i.i.d. process with r possible values for each data point (e.g., r possible locations), we obtain a no-privacy region in the $m(n) - a_n$ plane. Figure 1a shows the no-privacy region for the case when there exists inter-user dependency, and Figure 1b shows the no-privacy region when the users' traces are independent across users. There exists a larger no-privacy region in the presence of inter-user dependency; therefore, we find that dependency between users weakens their privacy.

In addition, for the case where users' datasets are governed by an irreducible and aperiodic Markov chains with r states and $|E|$ edges, we obtain similar results, again showing that inter-user dependency degrades user privacy.

Note that for the case when only anonymization is employed, an initial extension in Gaussian case with known covariance matrix is also presented in [54].

II. System Model, Definitions, and Metrics

Here, we employ a similar framework to [34], [35]. The system has n users, and $X_u(k)$ is the data point of user u at time k. Our main goal is protecting $X_u(k)$ from a strong adversary who has full knowledge of the (unique) marginal probability distribution function of the data points of each user based on previous observations or other sources. In order to achieve data privacy for users, both anonymization and obfuscation techniques can be used as shown in Figure 2. In Figure 2, $Z_u(k)$ shows the (reported) data point of user u at time k after applying obfuscation, and $Y_u(k)$ shows the (reported) data point of user u at time k after applying anonymization to $Z_u(k)$. Let $m = m(n)$ be the number of data points after which the pseudonyms of users are changed using anonymization. To break obfuscation and anonymization, the adversary tries to estimate $X_u(k)$, $k = 1, 2, \cdots, m$, from m observations per user by matching the sequence of observations to the known statistical characteristics of the users. Let $\mathbf{X}_u$ be the $m \times 1$ vector containing the data points of user u, and $\mathbf{X}$ be the $m \times n$ matrix with the u^{th} column equal to $\mathbf{X}_u$:

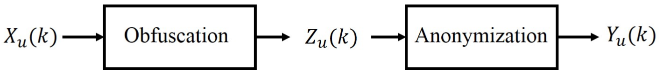


Fig. 2: Applying obfuscation and anonymization techniques to the users' data points.

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

Data Points Model: Here, we assume two different models for users' data points: in the first case, we assume the sequence of data for any individual user is modeled as i.i.d. which could 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 temporal dependency. In the second case, we assume the data trace of any individual user is governed by a Markov chain that governs how samples of a user's data are dependent over time.

We also assume users' data points can have one of r possible values $(0, 1, \cdots, r-1)$. Thus, according to a user-specific probability distribution $(\mathbf{p}_u)$, $X_u(k)$ is equal to a value in

$\{0, 1, \cdots, r-1\}$ at any time. Note $p_u(i)$ 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. Note these user-specific probability distributions, i.e., $\mathbf{p}_u$'s, are known to the adversary and form the basis upon which they perform (statistical) matching.

Association Graph: An association graph or dependency graph is an undirected graph representing dependency of the data of users with each other. Let $G(\mathcal{V}, F)$ denote the association graph with set of nodes $\mathcal{V}$, ($|\mathcal{V}| = n$), and set of edges F. Two vertices (users) are connected if their data sets are dependent. More specifically,

- $(u, u') \notin F$ if and only if $I(X_u(k); X_{u'}(k)) = 0$,
- $(u, u') \in F$ if and only if $I(X_u(k); X_{u'}(k)) > 0$,

where $I\left(X_u(k); X_{u'}(k)\right)$ is the mutual information between the k^{th} data point of user u and user u'[1].

Obfuscation Model: Obfuscation perturbs the users' data points [55]–[57]; in other words, the obfuscation can be viewed as passing data through a noisy channel. Normally, in such settings, each user has only limited knowledge of the characteristics of the overall population. Thus, usually, a simple distributed method in which the data points of each user are reported with error with a certain probability is employed [58]. Note that this probability itself is generated randomly for each user. Let $\mathbf{Z}_u$ be the vector that contains the obfuscated version of user u's data points, and $\mathbf{Z}$ be the collection of $\mathbf{Z}_u$ for all users,

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

Here, we define the *asymptotic noise level* for an obfuscation technique. Loosely speaking, the asymptotic noise level of obfuscation is the highest probable percentage of data points that are corrupted. More precisely, for a subset of users $\mathfrak{U}$, let $X_u(k)$ be the actual data point of user u at time k, $u \in \mathfrak{U}$, $k \in \{1, 2, \cdots, m\}$, and let $Z_u(k)$ be the obfuscated (noisy) version of $X_u(k)$. Define

$$A_m(u) = \frac{|\{k : Z_u(k) \neq X_u(k)\}|}{m}.$$

Then, the asymptotic noise level for user u is defined as follows:

$$a(u) = \inf\left\{\tau \geq 0 : \mathbb{P}\left(A_m(u) > \tau\right) \to 0 \text{ as } m \to \infty\right\}.$$

Also, define

$$A_m = \frac{\sum\limits_{u \in \mathfrak{U}} |\{k : Z_u(k) \neq X_u(k)\}|}{m|\mathfrak{U}|},$$

then, the asymptotic noise level for the entire dataset is

$$a = \inf\left\{\tau \geq 0 : \mathbb{P}\left(A_m > \tau\right) \to 0 \text{ as } m \to \infty\right\}.$$

Note that while the above definition is given for a general case required in Section VI, in practice we often use simple obfuscation techniques that employ i.i.d. noise sequences. Then, by the Strong Law of Large Number (SLLN),

$$\frac{|\{k : Z_u(k) \neq X_u(k)\}|}{m} \xrightarrow{\text{a.s.}} \mathbb{P}\left(Z_u(k) \neq X_u(k)\right),$$

and for any k,

$$a(u) = \mathbb{P}\left(Z_u(k) \neq X_u(k)\right).$$

Anonymization Model: In the anonymization technique, the identity of the users is perturbed [2], [59]–[64]. Anonymization is modeled as a random permutation Π on the set of n users. Let $\mathbf{Y}_u$ be the vector which contains the anonymized version of $\mathbf{Z}_u$, and $\mathbf{Y}$ is the collection of $\mathbf{Y}_u$ for all users, thus

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

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

Adversary Model: We assume the adversary has full knowledge of the marginal probability distribution function of each of the users on $\{0, 1, \ldots, r-1\}$. As discussed in the data points models in succeeding sections, 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 given hypercube. The density $f_{\mathbf{P}}(\mathbf{p}_u)$ might be unknown to the adversary, so all that is assumed here is that such a density exists. From the results of the paper, it will be evident that knowing or not knowing $f_{\mathbf{P}}(\mathbf{p}_u)$ does not change the results asymptotically.

The adversary knows the anonymization mechanism but does not know the realization of the random permutation. The adversary also knows the obfuscation mechanism but does not know the realization of the noise parameters. And finally, the adversary knows the association graph $G(\mathcal{V}, F)$, but does not necessarily know the exact nature of the dependency. That is, while the adversary knows the marginal distributions $X_u(k)$ as well as which pairs of users have strictly positive mutual information, they might not know the joint distributions or even the values of the mutual information $\mathbb{I}\left(X_u(k); X_{u'}(k)\right)$.

It is critical to note that the adversary does not have any other auxiliary information or side information about users' data.

[1] It is worth noting that the mechanism that determines the joint distribution of $X_u(k)$ and $X_{u'}(k)$ does not affect the results of this paper as long as the marginal densities of $X_u(k)$'s (i.e., $\mathbf{p}_u$'s) are drawn independently from $f_{\mathbf{P}}(\mathbf{p}_u)$.

We adopt the definitions of perfect privacy and no privacy from [34], [35]:

Definition 1. For an algorithm for the adversary that tries to estimate the actual data point of user u at time k, define the error probability as

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

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

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

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.

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

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

where $\mathbb{I}\left(X_u(k); \mathbf{Y}\right)$ denotes the mutual information between the data point of user u at time k and the collection of the adversary's observations for all of the users.

Discussion 1**:** The studied anonymization and obfuscation mechanisms improve user privacy at the cost of user utility. An 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, such frequent changes may also degrade the usability of the underlying application by concealing the temporal relation between a user's data points, e.g., for 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 also degrade the utility of the system. In this work, our goal is studying the level of anonymization and obfuscation one should employ to ensure privacy with the minimum loss in utility. In other words, we derive the optimal frequency of changing user pseudonyms during anonymization, and the optimal extent of noise added by an obfuscation mechanism while guaranteeing privacy.

However, like similar works in privacy [2], [18], [59]–[62], we consider the quantification of utility orthogonal to our privacy evaluations for two reasons: (1) the implications of our PPMs on utility do not impact our privacy analysis, and (2) unlike privacy, the desired level of utility is application specific.

Discussion 2**:** Note that there are two kinds of dependency:

- Intra-user dependency: In this case, there are temporal and spatial dependency within data traces of one user. For example, when the data trace of a user is governed by a Markov chain model, the Markov chain characterizes temporal intra-user dependency. Thus, the adversary can benefit from this dependency and break the users' privacy. According to the results obtained in [35], when there are a large number of users in the setting ($n \to \infty$), and data traces of the users are governed by i.i.d. statistics with r possible values for each data point, users have no privacy if the number of adversary's observations per user (m) is significantly larger than $n^{\frac{2}{r-1}}$ and the amount of noise level (a_n) is significantly smaller than $n^{-\frac{1}{r-1}}$; however, if the data trace of users is governed by an irreducible and aperiodic Markov chains with r states and $|E|$ edges, users have no privacy if the number of adversary's observations per user (m) is significantly larger than $n^{\frac{2}{|E|-r}}$ and the amount of noise level (a_n) is significantly smaller than $n^{-\frac{1}{|E|-r}}$. Most of the previous work [19]–[25] that considers intra-user dependency assumes independence between the traces of different users, which is different from our work as described below.
- Inter-user dependency: Here, there exists dependency between the traces of different users. This is the main focus of our work. First, we demonstrate that the adversary can readily identify the association graph of the obfuscated and anonymized version of the data, revealing which user data traces are dependent. Next, we demonstrate that the adversary can use this association graph along with their statistical knowledge and the observed obfuscated and anonymized sequences to break user privacy with significantly shorter traces than in the case of independent users, and that obfuscating data traces independently across users is often insufficient to remedy such leakage.

Discussion 3**:** The multi-user models in classical information theory generally assume a fixed number of users and the fundamental limits of communication systems are characterized in the limit of large coding blocklength [65]–[68]. However, the emerging Internet of Things enables an ever-increasing number of users to share and access information on a large scale; in applications such as ride sharing and dining recommendation, the number of users is large. Thus, the number of users is allowed to grow with the blocklength [69]–[71], and our goal is for the asymptotic results to provide good insight into the performance of the privacy-preserving mechanisms for these applications. Moreover, 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.

III. Impact of Dependency on Privacy Using Anonymization

In this section, we consider only anonymization and thus the obfuscation block in Figure 2 is not present. In this case, the adversary's observation $\mathbf{Y}$ is the anonymized version of $\mathbf{X}$; thus

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

A. r-State i.i.d. Model

There is potentially dependency between the data of different users, but we assume here that the sequence of data for any individual user is i.i.d.. We also assume users' data points can have r possibilities $(0, 1, \cdots, r-1)$, and $p_u(i)$ is the probability of user u having the data value i, i.e., $p_u(i) = \mathbb{P}\left(X_u(k) = i\right)$, for $k = 1, 2, \cdots, m$. Note that $p_u(0) = 1 - \sum_{r=1}^{r-1} p_u(r)$; thus, the adversary can focus on a subset of size $r-1$ of the probabilities for recovering the entire probability vector of user u. As a result, we define the vectors $\mathbf{p}_u$ and $\mathbf{p}$ as

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

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. In particular, define the range of the distribution as

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

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

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

The adversary knows the values of $\mathbf{p}_u$, $u = 1, 2, \cdots, n$, and uses this knowledge to match the observed traces to the users. We will use capital letters (i.e., $\mathbf{P}_u$) when we are referring to the random variable, and use lower case (i.e., $\mathbf{p}_u$) to refer to the realization of $\mathbf{P}_u$.

A vector containing the permutation of those probabilities after anonymization is

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

where $\mathbf{W}_u = \mathbf{P}_{\Pi^{-1}(u)}$ and $\mathbf{W}_{\Pi(u)} = \mathbf{P}_u$.

In this case, we can say:

- $(u, u') \notin F$ if and only if for all $i, j \in \{0, 1, \cdots, r-1\}$, $p_{uu'}(i,j) = p_u(i) p_{u'}(j)$,
- $(u, u') \in F$ if and only if for at least one pair of $i, j \in \{0, 1, \cdots, r-1\}$, $p_{uu'}(i,j) \neq p_u(i) p_{u'}(j)$,

where $p_{uu'}(i,j) = \mathbb{P}\left(X_u(k) = i, X_{u'}(k) = j\right)$, $p_u(i) = \mathbb{P}\left(X_u(k) = i\right)$, and $p_{u'}(j) = \mathbb{P}\left(X_{u'}(k) = j\right)$. Note that the adversary knows the association graph $G(\mathcal{V}, F)$, but does not necessarily know the joint probability distribution for each specific $(u, u') \in F$. The adversary observes the anonymized version of users' data traces and combines them with their full knowledge of the marginal probability distribution of each of the users and the structure of the whole association graph to break users' privacy with arbitrarily small error probability.

In the first step, we show that the adversary can reliably reconstruct the entire association graph for *the anonymized version of the data* (i.e., the observed data traces) with relatively few observations.

Lemma 1. Consider a general association graph $G(\mathcal{V}, F)$. If the adversary obtains $m = (\log n)^3$ anonymized observations per user, they can construct $\widetilde{G} = \widetilde{G}(\widetilde{\mathcal{V}}, \widetilde{F})$, where $\widetilde{\mathcal{V}} = \{\Pi(u) : u \in \mathcal{V}\} = \mathcal{V}$, such that with high probability, for all $u, u' \in \mathcal{V}$; $(u, u') \in F$ if and only if $(\Pi(u), \Pi(u')) \in \widetilde{F}$. We write this statement as $\mathbb{P}(\widetilde{G} \simeq G) \to 1$, i.e., Graph G and Graph $\widetilde{G}$ are isomorphic with high probability.

Proof. For $u, u' \in \{1, 2, \cdots, n\}$, we normally write $v = \Pi(u)$ and $v' = \Pi(u')$. We provide an algorithm for the adversary that with high probability obtains all edges of F correctly. First, for all $v, v' \in \{1, 2, \cdots, n\}$, and all $i, j \in \{0, 1, \cdots, r-1\}$ the adversary computes $\widetilde{p_{vv'}(i,j)}$, $\widetilde{p_v(i)}$, and $\widetilde{p_{v'}(j)}$ as follows:

$$\widetilde{p_{vv'}(i,j)} = \frac{|\{k : Y_v(k) = i, Y_{v'}(k) = j\}|}{m} = \frac{\widehat{M}_{vv'}(i,j)}{m}, \quad (1)$$

$$\widetilde{p_v(i)} = \frac{|\{k : Y_v(k) = i\}|}{m} = \frac{\widehat{M}_v(i)}{m}, \quad (2)$$

$$\widetilde{p_{v'}(j)} = \frac{|\{k : Y_{v'}(k) = j\}|}{m} = \frac{\widehat{M}_{v'}(j)}{m}, \quad (3)$$

where

$$\widehat{M}_{vv'}(i,j) = |\{k : Y_v(k) = i, Y_{v'}(k) = j\}|.$$
$$\widehat{M}_v(i) = |\{k : Y_v(k) = i\}|.$$
$$\widehat{M}_{v'}(j) = |\{k : Y_{v'}(k) = j\}|.$$

After observing $m = (\log n)^3$ data points per user and computing the above expressions, the adversary constructs $\widetilde{G}$ in the following way:

- If $\left|\frac{\widehat{M}_{vv'}(i,j)}{m} - \frac{\widehat{M}_v(i)}{m}\frac{\widehat{M}_{v'}(j)}{m}\right| \leq m^{-\frac{1}{5}}$ for all $i, j \in \{0, 1, \cdots, r-1\}$, then $(v, v') \notin \widetilde{F}$.
- If $\left|\frac{\widehat{M}_{vv'}(i,j)}{m} - \frac{\widehat{M}_v(i)}{m}\frac{\widehat{M}_{v'}(j)}{m}\right| \geq m^{-\frac{1}{5}}$ for at least one pair of $i, j \in \{0, 1, \cdots, r-1\}$, then $(v, v') \in \widetilde{F}$.

We show the above method yields $\mathbb{P}(\widetilde{G} \simeq G) \to 1$ as $n \to \infty$, as follows. Note

$$\widehat{M}_{vv'}(i,j) \sim \mathrm{Binomial}(m, w_{vv'}(i,j)),$$
$$\widehat{M}_v(i) \sim \mathrm{Binomial}(m, w_v(i)),$$
$$\widehat{M}_{v'}(j) \sim \mathrm{Binomial}(m, w_{v'}(j)),$$

where $w_{vv'}(i,j) = \mathbb{P}\left(Y_v(k) = i, Y_{v'}(k) = j\right) = p_{\Pi^{-1}(v)\Pi^{-1}(v')}(i,j)$, $w_v(i) = \mathbb{P}\left(Y_v(k) = i\right) = p_{\Pi^{-1}(v)}(i)$, and $w_{v'}(j) = \mathbb{P}\left(Y_{v'}(k) = j\right) = p_{\Pi^{-1}(v')}(j)$. Now, for all $v, v' \in \{1, 2, \cdots, n\}$ and all $i, j \in \{0, 1, \cdots, r-1\}$, define

$$\mathcal{J}_{vv'}(i,j) = \left\{\left|\frac{M_{vv'}(i,j)}{m} - w_{vv'}(i,j)\right| \geq m^{-\frac{1}{4}}\right\},$$

then, for all $v, v' \in \{1, 2, \cdots, n\}$ and all $i, j \in \{0, 1, \cdots, r-1\}$, the Chernoff bound yields

$$\mathbb{P}\left(\mathcal{J}_{vv'}(i,j)\right) \leq 2\exp\left(-\frac{\sqrt{m}}{3 w_{vv'}(i,j)}\right) \leq 2\exp\left(-\frac{\sqrt{m}}{3}\right).$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2020.2985059, IEEE Transactions on Information Theory

Similarly, for all $v, v' \in \{1, 2, \cdots, n\}$ and all $i, j \in \{0, 1, \cdots, r-1\}$, define

$$\mathcal{J}_v(i) = \left\{ \left| \frac{M_v(i)}{m} - w_v(i) \right| \geq m^{-\frac{1}{4}} \right\},$$

$$\mathcal{J}_{v'}(j) = \left\{ \left| \frac{M_{v'}(j)}{m} - w_{v'}(j) \right| \geq m^{-\frac{1}{4}} \right\};$$

then, the Chernoff bound yields,

$$\mathbb{P}\left(\mathcal{J}_v(i)\right) \leq 2 \exp\left(-\frac{\sqrt{m}}{3} \right),$$

$$\mathbb{P}\left(\mathcal{J}_{v'}(j)\right) \leq 2 \exp\left(-\frac{\sqrt{m}}{3} \right),$$

Now, by employing a union bound, for all $v, v' \in \{1, 2, \cdots, n\}$ and all $i, j \in \{0, 1, \cdots, r-1\}$, we have

$$\begin{aligned} &\mathbb{P}\left(\mathcal{J}_{vv'}(i,j) \cup \mathcal{J}_v(i) \cup \mathcal{J}_{v'}(j)\right) \\ &\quad \leq 2\left(\exp\left(-\frac{\sqrt{m}}{3}\right) + \exp\left(-\frac{\sqrt{m}}{3}\right) + \exp\left(-\frac{\sqrt{m}}{3}\right) \right) \\ &\quad = 6 \exp\left(-\frac{\sqrt{m}}{3}\right). \end{aligned}$$

Then, by employing a union bound again,

$$\begin{aligned} &\mathbb{P}\left(\bigcup_{v=1}^{n} \bigcup_{v'=1}^{n} \bigcup_{i=0}^{r-1} \bigcup_{j=0}^{r-1} \left\{ \mathcal{J}_{vv'}(i,j) \cup \mathcal{J}_v(i) \cup \mathcal{J}_{v'}(j) \right\} \right) \\ &\qquad \leq \sum_{v=1}^{n} \sum_{v'=1}^{n} \sum_{i=0}^{r-1} \sum_{j=0}^{r-1} 6 \exp\left(-\frac{\sqrt{m}}{3}\right) \\ &\qquad = 6n^2 r^2 \exp\left(-\frac{\sqrt{m}}{3}\right) \\ &\qquad = 6r^2 \exp\left\{ 2\log n - \frac{(\log n)^{\frac{3}{2}}}{3} \right\}, \end{aligned} \tag{4}$$

so the right-side of (4) goes to 0 as $n \longrightarrow \infty$. Thus, (4) yields that with high probability, for all $v, v' \in \{1, 2, \cdots, n\}$ and all $i, j \in \{0, 1, \cdots, r-1\}$, we have

$$0 \leq m w_{vv'}(i,j) - m^{\frac{3}{4}} \leq \widehat{M}_{vv'}(i,j) \leq m w_{vv'}(i,j) + m^{\frac{3}{4}}. \tag{5}$$

$$0 \leq m w_v(i) - m^{\frac{3}{4}} \leq \widehat{M}_v(i) \leq m w_v(i) + m^{\frac{3}{4}}. \tag{6}$$

$$0 \leq m w_{v'}(j) - m^{\frac{3}{4}} \leq \widehat{M}_{v'}(j) \leq m w_v(j) + m^{\frac{3}{4}}. \tag{7}$$

Let us define event $A_{vv'}(i,j)$ as the event that (5), (6), and (7) are all valid, thus, as shown in (4), we have

$$\mathbb{P}\left(\bigcap_{v=1}^{n} \bigcap_{v'=1}^{n} \bigcap_{i=0}^{r-1} \bigcap_{j=0}^{r-1} \{A_{vv'}(i,j)\} \right) \longrightarrow 1, \tag{8}$$

as $n \longrightarrow \infty$. Now, if $A_{vv'}(i,j)$ is true for some $v, v' \in \{1, 2, \cdots, n\}$ and some $i, j \in \{0, 1, \cdots, r-1\}$, we have

$$\begin{aligned} &\frac{\widehat{M}_{vv'}(i,j)}{m} - \frac{\widehat{M}_v(i)}{m} \frac{\widehat{M}_{v'}(j)}{m} \\ &\quad \leq \frac{m w_{vv'}(i,j) + m^{\frac{3}{4}}}{m} - \frac{m w_v(i) - m^{\frac{3}{4}}}{m} \frac{m w_{v'}(i) - m^{\frac{3}{4}}}{m} \\ &\quad = w_{vv'}(i,j) - w_v(i) w_{v'}(j) + m^{-\frac{1}{4}} \\ &\qquad + (w_v(i) + w_{v'}(j)) m^{-\frac{1}{4}} - m^{-\frac{1}{2}} \\ &\quad \leq w_{vv'}(i,j) - w_v(i) w_{v'}(j) + m^{-\frac{1}{4}} \\ &\qquad + (w_v(i) + w_{v'}(j)) m^{-\frac{1}{4}} + m^{-\frac{1}{2}}. \end{aligned} \tag{9}$$

Similarly,

$$\begin{aligned} &\frac{\widehat{M}_{vv'}(i,j)}{m} - \frac{\widehat{M}_v(i)}{m} \frac{\widehat{M}_{v'}(j)}{m} \\ &\quad \geq \frac{m w_{vv'}(i,j) - m^{\frac{3}{4}}}{m} - \frac{m w_v(i) + m^{\frac{3}{4}}}{m} \frac{m w_{v'}(i) + m^{\frac{3}{4}}}{m} \\ &= w_{vv'}(i,j) - w_v(i) w_{v'}(j) - m^{-\frac{1}{4}} \\ &\qquad - (w_v(i) + w_{v'}(j)) m^{-\frac{1}{4}} - m^{-\frac{1}{2}}. \end{aligned} \tag{10}$$

Thus, by using (9) and (10), for some $v, v' \in \{1, 2, \cdots, n\}$ and some $i, j \in \{0, 1, \cdots, r-1\}$, we have

$$\begin{aligned} &\left| \left(\frac{\widehat{M}_{vv'}(i,j)}{m} - \frac{\widehat{M}_v(i)}{m} \frac{\widehat{M}_{v'}(j)}{m} \right) - (w_{vv'}(i,j) - w_v(i) w_{v'}(j)) \right| \\ &\qquad \leq (1 + w_v(i) + w_{v'}(j)) m^{-\frac{1}{4}} + m^{-\frac{1}{2}}. \end{aligned} \tag{11}$$

Let us define event $B_{vv'}(i,j)$ as the event that (11) is valid for v, v', i, and j. We have shown, for any $v, v' \in \{1, 2, \cdots, n\}$ and any $i, j \in \{0, 1, \cdots, r-1\}$, $A_{vv'}(i,j) \subseteq B_{vv'}(i,j)$, thus

$$\left\{ \bigcap_{v=1}^{n} \bigcap_{v'=1}^{n} \bigcap_{i=0}^{r-1} \bigcap_{j=0}^{r-1} \{A_{vv'}(i,j)\} \right\} \subseteq \left\{ \bigcap_{v=1}^{n} \bigcap_{v'=1}^{n} \bigcap_{i=0}^{r-1} \bigcap_{j=0}^{r-1} \{B_{vv'}(i,j)\} \right\},$$

and as a result,

$$\begin{aligned} &\mathbb{P}\left(\bigcap_{v=1}^{n} \bigcap_{v'=1}^{n} \bigcap_{i=0}^{r-1} \bigcap_{j=0}^{r-1} \{B_{vv'}(i,j)\} \right) \\ &\qquad \geq \mathbb{P}\left(\bigcap_{v=1}^{n} \bigcap_{v'=1}^{n} \bigcap_{i=0}^{r-1} \bigcap_{j=0}^{r-1} \{A_{vv'}(i,j)\} \right). \end{aligned}$$

Thus, by using (8), we have

$$\mathbb{P}\left(\bigcap_{v=1}^{n} \bigcap_{v'=1}^{n} \bigcap_{i=0}^{r-1} \bigcap_{j=0}^{r-1} \{B_{vv'}(i,j)\} \right) \longrightarrow 1,$$

as $n \longrightarrow \infty$. Hence, with high probability, (11) is simultaneously valid for all $v, v' \in \{1, 2, \cdots, n\}$ and all $i, j \in \{0, 1, \cdots, r-1\}$.

Now, if $(u, u') \notin F$, then for all $i, j \in \{0, 1, \cdots, r-1\}$, we have $p_{uu'}(i,j) - p_u(i) p_{u'}(j) = 0$, and as a result, $w_{vv'}(i,j) - w_v(i) w_{v'}(j) = 0$. Thus, we have

$$\left| \frac{\widehat{M}_{vv'}(i,j)}{m} - \frac{\widehat{M}_v(i)}{m} \frac{\widehat{M}_{v'}(j)}{m} \right| \leq (1 + w_v(i) + w_{v'}(j)) m^{-\frac{1}{4}} + m^{-\frac{1}{2}}, \tag{12}$$

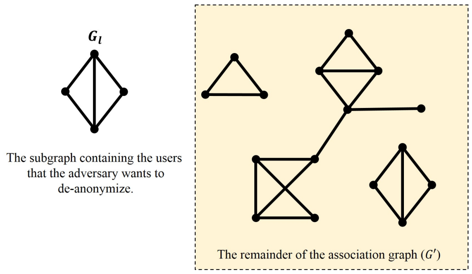

G_l
The subgraph containing the users that the adversary wants to de-anonymize.
The remainder of the association graph (G')

Fig. 3: The structure of the association graph (G): Group l with s_l vertices is disjoint from the remainder of the association graph (G').

and as a result, for large enough m,

$$\left| \frac{\widehat{M}_{vv'}(i,j)}{m} - \frac{\widehat{M}_v(i)}{m}\frac{\widehat{M}_{v'}(j)}{m} \right| \le m^{-\frac{1}{5}}. \tag{13}$$

Thus, we can conclude, $(v,v') \notin \widetilde{F}$, and in other words, $(\Pi(u),\Pi(u')) \notin \widetilde{F}$. This is true with high probability, for all $u,u' \in \{1,2,\cdots,n\}$ where $(u,u') \notin F$.

Similarly, if $(u,u') \in F$, there exists at least one pair of $i,j \in \{0,1,\cdots,r-1\}$ with $p_{uu'}(i,j) - p_u(i)p_{u'}(j) \ge \epsilon - m^{-\frac{1}{4}}$ for a fixed value of ϵ. Thus, there exists at least one pair of $i,j \in \{0,1,\cdots,r-1\}$ with $w_{vv'}(i,j) - w_v(i)w_{v'}(j) \ge \epsilon - m^{-\frac{1}{4}}$. As a result, for large enough m, we have

$$\left| \frac{\widehat{M}_{vv'}(i,j)}{m} - \frac{\widehat{M}_v(i)}{m}\frac{\widehat{M}_{v'}(j)}{m} \right| \ge m^{-\frac{1}{5}}.$$

Thus, we can conclude, $(v,v') \in \widetilde{F}$, and in other words, $(\Pi(u),\Pi(u')) \in \widetilde{F}$. Again, this is true with high probability, for all $u,u' \in \{1,2,\cdots,n\}$ where $(u,u') \in F$.

Now, we can conclude, for large enough n, we have $\mathbb{P}\left(\widetilde{G} \simeq G\right) \to 1$, so the adversary can reconstruct the association graph of the anonymized version of the data with an arbitrarily small error probability. Note that reconstruction of the association graph does not require the adversary's knowledge about user statistics (i.e., the values of $\mathbf{p}_u$'s). □

The structure of the association graph (G) can leak a lot of information. For the rest of this section, we consider a graph structure shown in Figure 3. In this structure, G_l, the subgraph consisting of the users the adversary wants to de-anonymize, has s_l vertices and is disjoint from the remainder of the association graph. So, we can write $G_l(\mathcal{V}_l, F_l)$, where $|\mathcal{V}_l| = s_l$. Note that we assume s_l is finite. In particular, the subgraph G_l can be thought of as a group of "friends" or "associates" such that their data sets are dependent. In Section V, we discuss how our methodology can be applied to the settings where the subgraph G_l is not disjoint from the remainder of the graph (G') [72]–[81].

The following theorem states that if the number of observations per user (m) is significantly larger than $n^{\frac{2}{s(r-1)}}$ in the r-state i.i.d. model, where s is the size of a group, then the adversary can successfully de-anonymize all of the data at different time samples ($k = 1,2,\cdots,m$) for all of the users in that group as the number of users in the network (n) goes to infinity.

Theorem 1. For the above r-state model, if $\mathbf{Y}$ is the anonymized version of $\mathbf{X}$ as defined above, the size of the group including User 1 is s, and

- $m = \Omega\left(n^{\frac{2}{s(r-1)}+\alpha}\right)$, for any $\alpha > 0$;

then, User 1 has no privacy at time k.

Discussion 4: It is insightful to compare this result to [34, Theorem 2], where it is stated that if the users are not dependent, then all users have perfect privacy as long as the number of adversary's observations per user (m) is smaller than $O(n^{\frac{2}{r-1}-\alpha})$. Here, Theorem 1 states that with much smaller m, the adversary can de-anonymize the users. Therefore, we see that dependency can significantly reduce the privacy of users.

Proof of Theorem 1:

Proof. As shown in Figure 4, the proof of Theorem 1 consists of three parts:

- **First Step:** Showing the adversary can reconstruct the association graph of the anonymized version of the data with an arbitrarily small error probability (as shown in Figure 4a).
- **Second Step:** Showing the adversary can uniquely identify Group 1 with an arbitrarily small error probability (as shown in Figure 4b).
- **Third Step:** Showing the adversary can individually identify all the members within Group 1 with an arbitrarily small error probability (as shown in Figure 4c).

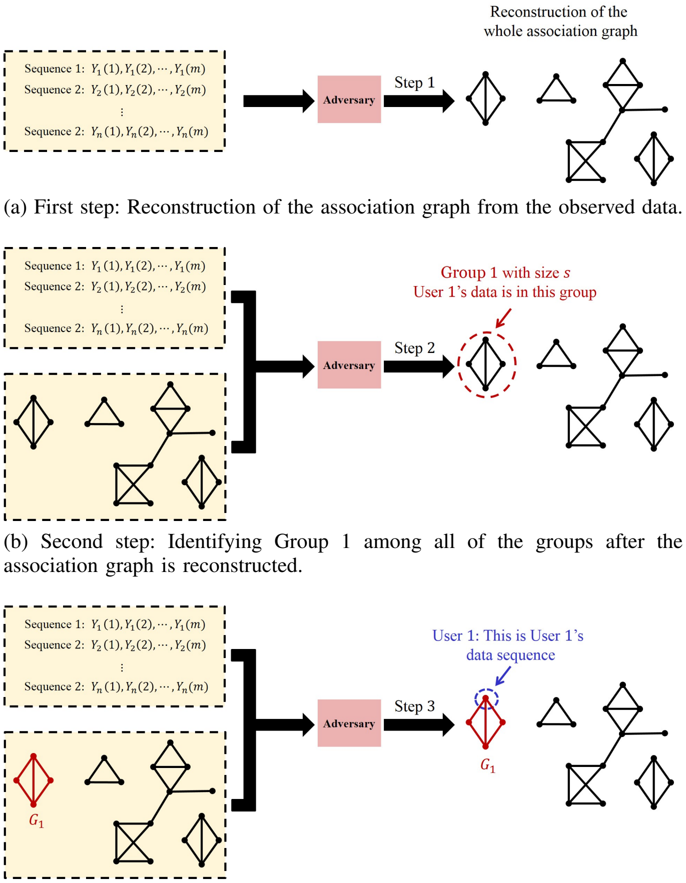
The first part of the proof exploits the fact that the adversary can readily reconstruct the association graph of the anonymized data in $m = (\log n)^3$. It is the second and third parts that give rise to the condition $m = \Omega\left(n^{\frac{2}{s(r-1)}+\alpha}\right)$, and it is in the second part where we see the mechanism for the effectiveness of the adversary's algorithm relative to the case where user traces are independent. In particular, due to the dependence between users breaking them into groups, the key search for the adversary now involves finding a set of users corresponding to a length-s vector of probabilities rather than searching for a single user associated with a given probability.

First step: Reconstruction of the association graph from the observed data: In this step, we use Lemma 1. More specifically, since $n^{\frac{2}{s(r-1)}} > (\log n)^3$ for large enough n, we can use Lemma 1 to conclude that the adversary can reconstruct the association graph with arbitrarily small error probability.

Second step: Identifying Group 1 among all of the groups: Now, assume the size of Group 1 is s. Without loss of generality, suppose the members of Group 1 are users $\{1,2,\cdots,s\}$. Note that there are at most $\frac{n}{s}$ isolated groups of size s in the association graph. We call these Groups $1,2,\cdots,\frac{n}{s}$. The adversary needs to first identify Group 1 among all of these groups.

First, for all $u \in \{1,2,\cdots,n\}$ and all $i \in \{1,2,\cdots,r-1\}$, the adversary computes $\widetilde{p_u(i)}$ as:

$$\widetilde{p_u(i)} = \frac{|\{k : Y_u(k) = i\}|}{m} = \frac{\widehat{M}_u(i)}{m}, \tag{14}$$


Reconstruction of the whole association graph
Sequence 1: $Y_1(1), Y_1(2), \cdots, Y_1(m)$
Sequence 2: $Y_2(1), Y_2(2), \cdots, Y_2(m)$
⋮
Sequence 2: $Y_n(1), Y_n(2), \cdots, Y_n(m)$
Adversary
Step 1
Group 1 with size s
User 1's data is in this group
Step 2
User 1: This is User 1's data sequence
Step 3
G_1

(a) First step: Reconstruction of the association graph from the observed data.

(b) Second step: Identifying Group 1 among all of the groups after the association graph is reconstructed.

(c) Third step: Identifying User 1 among all of the members of Group 1 after Group 1 is uniquely identified.

Fig. 4: The algorithm of the adversary to estimate data points of User 1 with vanishing error probability.

and as a result,

$$\widetilde{p_{\Pi(u)}(i)} = \frac{|\{k : X_u(k) = i\}|}{m} = \frac{M_u(i)}{m}, \tag{15}$$

where $\widehat{M_u}(i) = |\{k : Y_u(k) = i\}|$ and $M_u(i) = |\{k : X_u(k) = i\}|$. Let $\widetilde{\mathbf{p}_u}$ be the collection of $\widetilde{p_u(i)}$ and $\widetilde{\mathbf{p}_{\Pi(u)}}$ be the collection of $\widetilde{p_{\Pi(u)}(i)}$ for all $i \in \{1, 2, \cdots, r-1\}$:

$$\widetilde{\mathbf{p}_u} = \begin{bmatrix} \widetilde{p_u(1)} \\ \widetilde{p_u(2)} \\ \vdots \\ \widetilde{p_u(r-1)} \end{bmatrix}, \quad \widetilde{\mathbf{p}_{\Pi(u)}} = \begin{bmatrix} \widetilde{p_{\Pi(u)}(1)} \\ \widetilde{p_{\Pi(u)}(2)} \\ \vdots \\ \widetilde{p_{\Pi(u)}(r-1)} \end{bmatrix}.$$

Now, define Σ_s as the set of all permutations on s elements; for $\sigma \in \Sigma_s$, $\sigma : \{1, 2, \cdots, s\} \to \{1, 2, \cdots, s\}$ is a one-to-one mapping.

First, we provide the definition of a distance measure $D(\mathbf{\Phi}, \mathbf{\Psi})$ for vectors

$$\mathbf{\Phi} = [\mathbf{\Phi}_1 \;\; \mathbf{\Phi}_2 \;\; \cdots \;\; \mathbf{\Phi}_s],$$

$$\mathbf{\Psi} = [\mathbf{\Psi}_1 \;\; \mathbf{\Psi}_2 \;\; \cdots \;\; \mathbf{\Psi}_s],$$

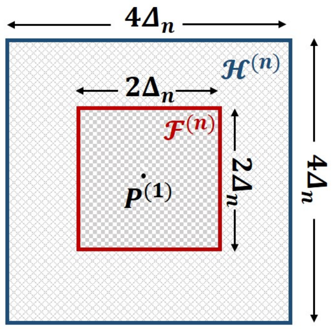


Fig. 5: $\mathbf{P}^{(1)}$, sets $\mathcal{F}^{(n)}$ and $\mathcal{H}^{(n)}$ for the case $r = s = 2$.

where $\mathbf{\Phi}_u \in \mathbb{R}^{r-1}$ and $\mathbf{\Psi}_u \in \mathbb{R}^{r-1}$. We need to (slightly) redefine our distance measure as

$$D(\mathbf{\Phi}, \mathbf{\Psi}) = \min_{\sigma \in \Sigma_s} \{d_\infty(\mathbf{\Phi}, \mathbf{\Psi}_\sigma)\},$$

where

$$d_\infty(\mathbf{\Phi}, \mathbf{\Psi}_\sigma) = \max_{u \in \{1,2,\cdots,s\}} \;\; \max_{i \in \{1,2,\cdots,r-1\}} \left\{\Phi_u(i), \Psi_{\sigma(u)}(i)\right\}.$$

Here, let $\mathbf{P}^{(l)}$ be a vector which contains probability distributions of users belonging to Group l, and $\widetilde{\mathbf{P}^{(l)}_{\Pi}}$ be a vector which contains the estimate of the adversary about the probability distribution of users belong to Group l. For example, for Group 1, we have

$$\mathbf{P}^{(1)} = [\mathbf{p}_1 \;\; \mathbf{p}_2 \;\; \cdots \;\; \mathbf{p}_s],$$

and

$$\widetilde{\mathbf{P}^{(1)}_{\Pi}} = \left[\widetilde{\mathbf{p}_{\Pi(1)}} \;\; \widetilde{\mathbf{p}_{\Pi(2)}} \;\; \cdots \;\; \widetilde{\mathbf{p}_{\Pi(s)}}\right].$$

Now, we claim for $m = cn^{\frac{2}{s(r-1)}+\alpha}$ and large enough n,

- $\mathbb{P}\left(D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}^{(1)}_{\Pi}}\right) \leq \Delta_n\right) \to 1,$
- $\mathbb{P}\left(\bigcup_{l=2}^{\frac{n}{s}} \left\{D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}^{(l)}_{\Pi}}\right) \leq \Delta_n\right\}\right) \to 0\,,$

where $\Delta_n = n^{-\frac{1}{s(r-1)}-\frac{\alpha}{4}}$.

Define the hypercubes of $\mathcal{F}^{(n)}$ and $\mathcal{H}^{(n)}$ as

$$\mathcal{F}^{(n)} = \Big\{(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_s) \in \left(\mathbb{R}^{(r-1)}\right)^s : \max_u\{|\mathbf{x}_u - \mathbf{p}_u|\} \leq \Delta_n, u = 1, 2, \cdots, s\Big\},$$

$$\mathcal{H}^{(n)} = \Big\{(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_s) \in \left(\mathbb{R}^{(r-1)}\right)^s : \max_u\{|\mathbf{x}_u - \mathbf{p}_u|\} \leq 2\Delta_n, u = 1, 2, \cdots, s\Big\},$$

Figure 5 shows sets $\mathcal{F}^{(n)}$ and $\mathcal{H}^{(n)}$ in the case $r = s = 2$.

First, we prove $\widetilde{\mathbf{P}_{\Pi}^{(1)}}$ is in set $\mathcal{F}^{(n)}$, thus, $D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}_{\Pi}^{(1)}}\right) \leq \Delta_n$. Note that for all $u \in \{1, 2, \cdots, n\}$ and all $i \in \{1, 2, \cdots, r-1\}$, a Chernoff bound yields

$$\begin{aligned}\mathbb{P}\left(\left|\frac{M_u(i)}{m} - p_u(i)\right| \geq \Delta_n\right) &\leq 2\exp\left(-\frac{m\Delta_n^2}{3p_u}\right)\\ &= 2\exp\left(-\left(cn^{\frac{2}{s(r-1)}+\alpha}\right)\left(\frac{1}{n^{\frac{1}{s(r-1)}+\frac{\alpha}{4}}}\right)^2\left(\frac{1}{3p_u}\right)\right)\\ &\leq 2\exp\left(-\frac{c}{3}n^{\frac{\alpha}{2}}\right). \qquad (16)\end{aligned}$$

Thus, for all $u \in$ Group 1 and all $i \in \{1, 2, \cdots, r-1\}$, (16) and the union bound yield

$$\begin{aligned}\mathbb{P}\left(D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}_{\Pi}^{(1)}}\right) \geq \Delta_n\right) &\leq \sum_{u=1}^{s}\sum_{i=1}^{r-1}\mathbb{P}\left(\left|\frac{M_u(i)}{m} - p_u(i)\right| \geq \Delta_n\right)\\ &\leq 2s(r-1)\exp\left(-\frac{c}{3}n^{\frac{\alpha}{2}}\right). \qquad (17)\end{aligned}$$

The right-side of (17) goes to 0 as $n \to \infty$. As a result, $D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}_{\Pi}^{(1)}}\right) \leq \Delta_n$ with high probability.

In the next step, we prove $\mathbb{P}\left(\bigcup_{l=2}^{\frac{n}{s}}\left\{D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}_{\Pi}^{(l)}}\right) \leq \Delta_n\right\}\right) \to 0$. Note that for all groups other than Group 1, we have

$$(4\Delta_n)^{s(r-1)}\delta_1 \leq \mathbb{P}\left(\mathbf{P}^{(l)} \in \mathcal{H}^{(n)}\right) \leq (4\Delta_n)^{s(r-1)}\delta_2,$$

and as a result,

$$\begin{aligned}\mathbb{P}\left(\mathbf{P}^{(l)} \in \mathcal{H}^{(n)}\right) &\leq \delta_2(4\Delta_n)^{s(r-1)}\\ &= \delta_2 4^{s(r-1)}\frac{1}{n^{1+\frac{\alpha}{4}s(r-1)}}.\end{aligned}$$

Similarly, for any $\sigma \in \Sigma_s$,

$$\begin{aligned}\mathbb{P}\left(\mathbf{P}_{\sigma}^{(l)} \in \mathcal{H}^{(n)}\right) &\leq \delta_2(4\Delta_n)^{s(r-1)}\\ &= \delta_2 4^{s(r-1)}\frac{1}{n^{1+\frac{\alpha}{4}s(r-1)}},\end{aligned}$$

and since $|\Sigma_s| = s!$, by a union bound, we have

$$\begin{aligned}\mathbb{P}\left(\bigcup_{l=2}^{\frac{n}{s}}\left\{\bigcup_{\sigma\in\Sigma_s}\left\{\mathbf{P}_{\sigma}^{(l)} \in \mathcal{H}^{(n)}\right\}\right\}\right) &\leq \sum_{l=2}^{\frac{n}{s}}\sum_{\sigma\in\Sigma_s}\mathbb{P}\left(\mathbf{P}_{\sigma}^{(l)} \in \mathcal{H}^{(n)}\right)\\ &\leq \frac{n}{s}s!\delta_2 4^{s(r-1)}\frac{1}{n^{1+\frac{\alpha}{4}s(r-1)}}\\ &= (s-1)!4^{s(r-1)}\delta_2 n^{-\frac{\alpha}{4}s(r-1)}. \qquad (18)\end{aligned}$$

The right-side of (18) goes to 0 as $n \to \infty$. Thus, all $\mathbf{P}^{(l)}$'s are outside of $\mathcal{H}^{(n)}$ with high probability.

Now, given the fact that all $\mathbf{P}^{(l)}$'s are outside of $\mathcal{H}^{(n)}$, we prove $\mathbb{P}\left(\bigcup_{l=2}^{\frac{n}{s}}\left\{D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}_{\Pi}^{(l)}}\right) \leq \Delta_n\right\}\right) \to 0$ as $n \to \infty$. We show that $\widetilde{\mathbf{P}_{\Pi}^{(l)}}$'s are close to $\mathbf{P}^{(l)}$'s, and as a result, they will be outside of $\mathcal{F}^{(n)}$. For all $u \in$ Group l and all $i \in \{1, 2, \cdots, r-1\}$, (16) and the union bound yield,

$$\begin{aligned}\mathbb{P}\left(D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}_{\Pi}^{(l)}}\right) \leq \Delta_n\right) &= \mathbb{P}\left(D\left(\mathbf{P}^{(l)}, \widetilde{\mathbf{P}_{\Pi}^{(l)}}\right) \geq \Delta_n\right)\\ &\leq \sum_{u=1}^{s}\sum_{i=1}^{r-1}\mathbb{P}\left(\left|\frac{M_u(i)}{m} - p_u(i)\right| \geq \Delta_n\right)\\ &\leq 2s(r-1)\exp\left(-\frac{c}{3}n^{\frac{\alpha}{2}}\right).\end{aligned}$$

Now by using a union bound, again, we have

$$\begin{aligned}\mathbb{P}\left(\bigcup_{l=2}^{\frac{n}{s}}\left\{D\left(\mathbf{P}^{(l)}, \widetilde{\mathbf{P}_{\Pi}^{(l)}}\right) \geq \Delta_n\right\}\right) &\leq \sum_{l=2}^{\frac{n}{s}}\mathbb{P}\left(D\left(\mathbf{P}^{(l)}, \widetilde{\mathbf{P}_{\Pi}^{(l)}}\right) \geq \Delta_n\right)\\ &\leq \frac{n}{s}2s(r-1)\exp\left(-\frac{c}{3}n^{\frac{\alpha}{2}}\right)\\ &= 2n(r-1)\exp\left(-\frac{c}{3}n^{\frac{\alpha}{2}}\right). \qquad (19)\end{aligned}$$

The right-side of (19) goes to 0 as $n \to \infty$. Thus, for all $l \in \{2, 3, \cdots, \frac{n}{s}\}$, $\widetilde{\mathbf{P}_{\Pi}^{(l)}}$'s are close to $\mathbf{P}^{(l)}$'s, thus, they will be outside of $\mathcal{F}^{(n)}$ with high probability. Now, we can conclude as $n \to \infty$,

$$\mathbb{P}\left(\bigcup_{l=2}^{\frac{n}{s}}\left\{D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}_{\Pi}^{(l)}}\right) \leq \Delta_n\right\}\right) \to 0.$$

This means that with high probability, for all $l \in \{2, 3, \cdots, \frac{n}{s}\}$, $\widetilde{\mathbf{P}_{\Pi}^{(l)}}$'s are outside of $\mathcal{F}^{(n)}$, so the adversary can successfully identify Group 1 among all of the groups.

Third step: Identifying User 1 among all of the members of Group 1: In this step, we prove that, after identifying Group 1, the adversary can correctly identify all of members of Group 1. This step can be done using a similar approach to the one above. We define two sets $\mathcal{B}^{(n)}$ and $\mathcal{C}^{(n)}$ around $\mathbf{p}_1$. We will show that with high probability, the true estimated value of $\mathbf{p}_1$ (shown as $\widetilde{\mathbf{p}_1}$) is inside of $\mathcal{B}^{(n)}$. Also, all $\mathbf{p}_u$'s of other members of Group 1 are outside of $\mathcal{C}^{(n)}$, and since their estimated values are close to $\mathbf{p}_u$'s, the estimated values will be outside of $\mathcal{B}^{(n)}$. Therefore, the adversary can successfully invert the permutation Π within Group 1 and identify all of the members. Below are the details.

From (14) and (15), for all $u \in \{1, 2, \cdots, s\}$ and all $i \in \{1, 2, \cdots, r-1\}$, we have

$$\widetilde{p_u(i)} = \frac{|\{k : Y_u(k) = i\}|}{m}, \qquad (20)$$

and as a result,

$$\widetilde{p_{\Pi(u)}(i)} = \frac{|\{k : X_u(k) = i\}|}{m} = \frac{M_u(i)}{m}, \qquad (21)$$

where $M_u(i) = |\{k : X_u(k) = i\}|$. Let's define sets $\mathcal{B}^{(n)}$ and $\mathcal{C}^{(n)}$ as

$$\begin{aligned}\mathcal{B}^{(n)} = \Big\{&(x_1, x_2, \cdots, x_{r-1}) \in \mathcal{R}_{\mathbf{P}} : |x_i - p_1(i)| \leq \Delta_n,\\ &i = 1, 2, \cdots, r-1\Big\},\end{aligned}$$

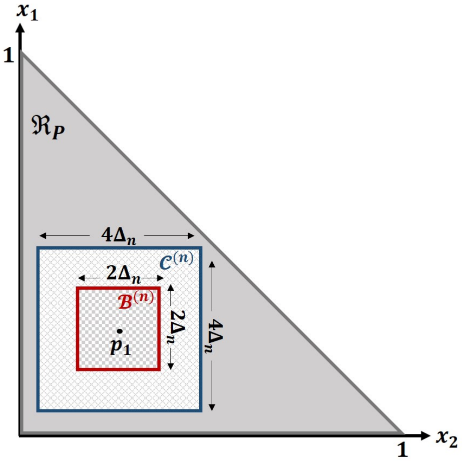


Fig. 6: $\mathbf{p}_1$, sets $\mathcal{B}^{(n)}$ and $C^{(n)}$ in $\mathcal{R}_{\mathbf{P}}$ for case $r = 3$.

$$C^{(n)} = \Big\{(x_1, x_2, \cdots, x_{r-1}) \in \mathcal{R}_{\mathbf{P}} : |x_i - p_1(i)| \leq 2\Delta_n, \quad i = 1, 2, \cdots, r-1\Big\},$$

where $\Delta_n = n^{-\frac{1}{s(r-1)} - \frac{\alpha}{4}}$. Figure 6 shows $\mathbf{p}_1$, sets $\mathcal{B}^{(n)}$ and $C^{(n)}$ in range of $\mathcal{R}_{\mathbf{P}}$ for case $r = 3$. Now, we claim for $m = cn^{\frac{2}{s(r-1)}+\alpha}$,

1) $\mathbb{P}\left(\widetilde{\mathbf{p}_{\Pi(1)}} \in \mathcal{B}^{(n)}\right) \to 1$,
2) $\mathbb{P}\left(\bigcup\limits_{u=2}^{s} \left\{\widetilde{\mathbf{p}_{\Pi(u)}} \in \mathcal{B}^{(n)}\right\}\right) \to 0$,

as $n \to \infty$. Thus, the adversary can identify $\Pi(1)$ by examining $\widetilde{\mathbf{p}_u}$'s and choosing the only one that belongs to $\mathcal{B}^{(n)}$.

In the first part, we want to show that as n goes to infinity,

$$\mathbb{P}\left(\widetilde{\mathbf{p}_{\Pi(1)}} \in \mathcal{B}^{(n)}\right) \to 1.$$

For all $i \in \{1, 2, \cdots, r-1\}$, By using (16) and the union bound, we have

$$\begin{aligned}\mathbb{P}\left(\widetilde{\mathbf{p}_{\Pi(1)}} \notin \mathcal{B}^{(n)}\right) &\leq \sum_{i=1}^{r-1} \mathbb{P}\left(\left|\frac{M_1(i)}{m} - p_1(i)\right| \geq \Delta_n\right) \\ &\leq (r-1)\left(2\exp\left(-\frac{c}{3}n^{\frac{\alpha}{2}}\right)\right);\end{aligned}$$

thus,

$$\mathbb{P}\left(\widetilde{\mathbf{p}_{\Pi(1)}} \in \mathcal{B}^{(n)}\right) \geq 1 - 2(r-1)\exp\left(-\frac{c}{3}n^{\frac{\alpha}{2}}\right). \tag{22}$$

The right-side of (22) goes to 1 as $n \to \infty$.

In the second part, we need to show that as n goes to infinity,

$$\mathbb{P}\left(\bigcup_{u=2}^{s} \left\{\widetilde{\mathbf{p}_{\Pi(u)}} \in \mathcal{B}^{(n)}\right\}\right) \to 0.$$

We show as n goes to infinity,

$$\mathbb{P}\left(\bigcup_{u=2}^{s} \left\{\mathbf{p}_u \in C^{(n)}\right\}\right) \to 0.$$

Note

$$4\left(\Delta_n\right)^{r-1}\delta_1 < \mathbb{P}\left(\mathbf{p}_u \in C^{(n)}\right) < 4\left(\Delta_n\right)^{r-1}\delta_2,$$

and according to the union bound, we have

$$\begin{aligned}\mathbb{P}\left(\bigcup_{u=2}^{s} \left\{\mathbf{p}_u \in C^{(n)}\right\}\right) &\leq \sum_{u=2}^{s} \mathbb{P}\left(\mathbf{p}_u \in C^{(n)}\right) \\ &\leq 4s\left(\Delta_n\right)^{r-1}\delta_2 \\ &\leq 4s\frac{1}{n^{\frac{1}{s}+\frac{\alpha(r-1)}{4}}}\delta_2.\end{aligned} \tag{23}$$

The right-side of (23) goes to 0 as $n \to \infty$, and as a result, all $\mathbf{p}_u$'s are outside of $C^{(n)}$ with high probability.

Now, we claim that given all $\mathbf{p}_u$'s are outside of $C^{(n)}$, $\mathbb{P}\left(\widetilde{\mathbf{p}_{\Pi(u)}} \in \mathcal{B}^{(n)}\right)$ is small. Note, for all $i \in \{1, 2, \cdots, r-1\}$, by using (16) and the union bounds, we have

$$\begin{aligned}\mathbb{P}\left(\widetilde{\mathbf{p}_{\Pi(u)}} \in \mathcal{B}^{(n)}\right) &\leq \mathbb{P}\left(\left|\widetilde{\mathbf{p}_{\Pi(u)}} - \mathbf{p}_u\right| \geq \Delta_n\right) \\ &\leq \sum_{i=1}^{r-1} \mathbb{P}\left(\left|\frac{M_u(i)}{m} - p_u(i)\right| \geq \Delta_n\right) \\ &\leq 2(r-1)\exp\left(-\frac{c}{3}n^{\frac{\alpha}{2}}\right).\end{aligned}$$

As a result, by using another union bound,

$$\mathbb{P}\left(\bigcup_{u=2}^{s} \left\{\left|\widetilde{\mathbf{p}_{\Pi(u)}} - \mathbf{p}_u\right| \geq \Delta_n\right\}\right) \leq s\left(2(r-1)\exp\left(-\frac{c}{3}n^{\frac{\alpha}{2}}\right)\right). \tag{24}$$

The right-side of (24) goes to 0 as $n \to \infty$. Thus, for all $u \in \{2, 3, \cdots, s\}$, $\widetilde{\mathbf{p}_{\Pi(u)}}$'s are close to $\mathbf{p}_u$'s; thus, they will be outside of $\mathcal{B}^{(n)}$. Now, we can conclude that:

$$\mathbb{P}\left(\bigcup_{u=2}^{s} \left\{\widetilde{\mathbf{p}_{\Pi(u)}} \in \mathcal{B}^{(n)}\right\}\right) \to 0,$$

as $n \to \infty$.

Thus, we have proved that if $m = cn^{\frac{2}{s(r-1)}+\alpha}$, there exists an algorithm for the adversary to successfully recover User 1. Remember, the adversary identifies the members of Group 1 independent of the structure of the subgraph. □

Discussion 5: Complexity of the adversary's algorithm shown in Figure 4:

1) Reconstruction of the association graph from the observed data: According to (1), for each of n^2 points, the adversary needs to perform m operations; thus, the complexity for the adversary is $O(mn^2)$. In addition, according to (2), for each of n points, the adversary needs to perform m operations, so the complexity for the adversary of (2) is $O(mn)$. Then, the thresholding requires $O\left(mn^2\right)$ operations. Since m can be as small as $(\log n)^3$ (according to Lemma 1), the minimum complexity of the adversary for the first step is $O\left((\log n)^3 n^2\right)$.
2) Identifying Group 1 among all of the groups: According to (15), for each of n points, the adversary needs to perform m operations; thus, the complexity for the adversary is $O(mn)$.Then, adversary has to search all $\frac{n}{s}$ groups of size s to find the one for which Group 1's vector is close. Since matching to an $s-$length group is $O(1)$ (since s is finite), the complexity of the adversary in the second step is $O\left(mn\right)$.

3) Identifying User 1 among all of the members of Group 1: According to (21), for each of s points, the adversary needs to perform m operations; thus, the complexity for the adversary is $O(ms)$. Since s is finite, the complexity of the adversary in the third step is $O(m)$.

Thus, the overall complexity of the adversary's algorithm could be $O(mn^2)$. Now, if the adversary was not overly concerned with complexity, they would need the m from the theorem statement put into step 1 results above ($O(mn^2)$). If they were more concerned about their complexity and did not use all m points from the Theorem statement to build the data dependency graph, but only the $O((\log n)^3)$ that they need in the first step, often the complexity would be $\max\left\{O((\log n)^3 n^2), O(mn)\right\}$, which for $s(r-1) > 2$ would be $O((\log n)^3 n^2)$.

B. r-State Markov Chain Model

In Sections III-A, we assumed each user's data pattern was i.i.d.; however, in this section, users' data patterns are modeled using Markov chains in which each user's data points are dependent over time. In this model, we again assume there are r possibilities for each users' data point, i.e., $X_u(k) \in \{0, 1, \cdots, r-1\}$. More specifically, each user's data set is modeled as a Markov chain with r states. It is assumed that the Markov chains of all users have the same structure but have different transition probabilities. Let E be the set of edges in the assumed transition graph, so, $(i, j) \in E$ if there exists an edge from state i to state j, meaning that $p_u(i, j) = \mathbb{P}(X_u(k+1) = j | X_u(k) = i) > 0$. The transition matrix is a square matrix used to describe the transitions of a Markov chain; thus, different users can have different transition probability matrices. Note for each state i, we have $\sum\limits_{j=1}^{r-1} p_u(i, j) = 1$, so, the adversary can focus on a subset of size $|E| - r$ of the transition probabilities for recovering the entire transition matrix. Let $\mathbf{p}_u$ be the vector that contains these transition probabilities for user u. We write

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

We also consider all $p_u(i)$'s are drawn independently from some continuous density function, $f_{\mathbf{P}}(\mathbf{p}_u)$, on the $(0, 1)^{|E|-r}$ hypercube. Define the range of distribution as

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

and as before, we assume there are $\delta_1, \delta_2 > 0$, such that

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

Now, we can repeat the similar steps as the previous sections to prove the following theorem.

Theorem 2. For an irreducible, aperiodic Markov chain model, if $\mathbf{Y}$ is the anonymized version of $\mathbf{X}$ as defined above, the size of the group including User 1 is s, and

- $m = \Omega\left(n^{\frac{2}{s(|E|-r)} + \alpha}\right)$, for any $\alpha > 0$;

then, User 1 has no privacy at time k.

Proof. The basic ideas behind the proof of Theorem 2 are similar to the ones for Theorem 1; thus, we highlight only the differences and key ideas.

Define the random variable $M_u(i)$ as the total number of visits by user u to state i, for all $u \in \{1, 2, \cdots, n\}$ and $i \in \{0, 1, \cdots, r-1\}$. Since the Markov chain is irreducible and aperiodic, and $m \to \infty$, all $\frac{M_i(u)}{m}$ converge to their stationary values [82]. Given $M_u(i) = m_u(i)$, the transitions from state i to state j for user u has a multinomial distribution with probabilities $p_u(i, j)$. Now, considering the fact that the vector $\mathbf{p}_u$ uniquely determines the user u, the adversary can invert the anonymization permutation function in a similar way to the i.i.d. case by focusing on $\mathbf{p}_u$'s. Let

$$\mathbf{\Phi} = [\mathbf{\Phi}_1 \quad \mathbf{\Phi}_2 \quad \cdots \quad \mathbf{\Phi}_s],$$
$$\mathbf{\Psi} = [\mathbf{\Psi}_1 \quad \mathbf{\Psi}_2 \quad \cdots \quad \mathbf{\Psi}_s],$$

where $\mathbf{\Phi}_u \in \mathbb{R}^{|E|-r}$ and $\mathbf{\Psi}_u \in \mathbb{R}^{|E|-r}$. We need to (slightly) re-define our distance measure as

$$D(\mathbf{\Phi}, \mathbf{\Psi}) = \min_{\sigma \in \Sigma_s} \{d_\infty(\mathbf{\Phi}, \mathbf{\Psi}_\sigma)\},$$

where

$$d_\infty(\mathbf{\Phi}, \mathbf{\Psi}_\sigma) = \max_{u \in \{1,2,\cdots,s\}} \max_{i \in \{1,2,\cdots,|E|-r\}} \left\{\Phi_u(i), \Psi_{\sigma(u)}(i)\right\}.$$

We claim for $m = cn^{\frac{2}{s(|E|-r)} + \alpha}$ and large enough n,

- $\mathbb{P}\left(D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}_{\Pi}^{(1)}}\right) \leq \Delta'_n\right) \to 1,$
- $\mathbb{P}\left(\bigcup\limits_{l=2}^{\frac{n}{s}} \left\{D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}_{\Pi}^{(l)}}\right) \leq \Delta'_n\right\}\right) \to 0,$

where $\Delta'_n = n^{-\frac{1}{s(|E|-r)} - \frac{\alpha}{4}}$. This can be shown similar to the proof of Theorem 1. First, define $\mathcal{F}'^{(n)}$ and $\mathcal{H}'^{(n)}$ as

$$\mathcal{F}'^{(n)} = \Big\{(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_s) \in \left(\mathbb{R}^{(|E|-r)}\right)^s : \\ \max_u\{|\mathbf{x}_u - \mathbf{p}_u|\} \leq \Delta'_n, u = 1, 2, \cdots, s\Big\};$$

$$\mathcal{H}'^{(n)} = \Big\{(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_s) \in \left(\mathbb{R}^{(|E|-r)}\right)^s : \\ \max_u\{|\mathbf{x}_u - \mathbf{p}_u|\} \leq 2\Delta'_n, u = 1, 2, \cdots, s\Big\};$$

then, it is straightforward to show that the adversary can identify Group 1 successfully.

In the next step, the adversary has to identify each member of Group 1 correctly. Define sets $\mathcal{B}'^{(n)}$ and $C'^{(n)}$ as

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

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

where $\Delta'_n = n^{-\frac{1}{s(|E|-r)}-\frac{\alpha}{4}}$. Now, we claim for $m = cn^{\frac{2}{s(|E|-r)}+\alpha}$,

1) $\mathbb{P}\left(\widetilde{\mathbf{p}_{\Pi(1)}} \in \mathcal{B}'^{(n)}\right) \to 1$,
2) $\mathbb{P}\left(\bigcup\limits_{u=2}^{s} \left\{\widetilde{\mathbf{p}_{\Pi(u)}} \in \mathcal{B}'^{(n)}\right\}\right) \to 0$,

as $n \longrightarrow \infty$. This can be shown in a similar manner to the proof of Theorem 1, so the adversary can successfully recover the data traces of User 1. □

Discussion 6: Note that the i.i.d. case can also be written as a Markov chain with a transition matrix with identical rows; then, $|E| = r^2$. However, for the i.i.d. case, if the adversary knows $r-1$ elements of a row, they know that row and all of the others. In other words, if we restrict the users' data models to i.i.d., then we are using a different model where $p_u(i)$'s are restricted in a way to create an i.i.d. sequence. This is a different model and is not compatible to our model for the Markov chain where $p_u(i)$'s are drawn independently from some continuous density function, $f_{\mathbf{P}}(\mathbf{p}_u)$, on the $(0,1)^{|E|-r}$ hypercube. Thus, the results of Theorem 2 cannot be applied to the i.i.d. case.

IV. Impact of Dependency on Privacy using Anonymization and Obfuscation

Here, we consider the case when both anonymization and obfuscation techniques are employed, as shown in Figure 2.

The mechanism that we employ is called randomized response in privacy literature [58] . In randomized response, the answer is changed randomly with some small probability. Note that the randomized response method has been analyzed frequently [83]–[93]; however, in those studies, a different approach to privacy termed differential privacy has been considered. Differential privacy is mainly used when there is a statistical database of users' sensitive information and the goal is to protect an individual's data while publishing aggregate information about the database [57], [94]–[97]. In contrast, here we consider the fundamental limits of a similar obfuscation technique for providing privacy in the long time series of emerging applications.

Here, we assume similar obfuscation to [35]. To obfuscate the users' data points, 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 is the probability that the user's data point is changed to a different value by obfuscation, and a_n is termed the "noise level' of the system. Let $\mathbf{Z}_u$ be the vector that contains the obfuscated version of user u's data points, and $\mathbf{Z}$ be the collection of $\mathbf{Z}_u$ for all users,

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

Thus, the adversary's observation $\mathbf{Y}$ is the anonymized version of $\mathbf{Z}$;

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

A. *r*-State i.i.d. Model

Now, assume users' data points can have r possibilities $(0, 1, \cdots, r-1)$. Similar to Section III-A, we 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, and $\mathbf{p}_u$, $f_{\mathbf{P}}(\mathbf{p}_u)$, and $\mathcal{R}_{\mathbf{P}}$ are defined as in Section III-A.

To create a noisy version of the data samples, for each user u, we independently generate a random variable R_u that is uniformly distributed between 0 and a_n, per above $a_n \in (0, 1]$ is the noise level[2]. Then, the obfuscated data is obtained by passing the users' data through an r-ary symmetric channel with a random error probability R_u [58], so for $j \in \{0, 1, \cdots, r-1\}$:

$$\mathbb{P}(Z_u(k) = j | X_u(k) = i) = \begin{cases} 1 - R_u, & \text{for } j = i. \\ \frac{R_u}{r-1}, & \text{for } j \ne i. \end{cases}$$

The effect of the obfuscation is to alter the probability distribution function of each user's data points 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 data points of length m, thus providing a very low-weight obfuscation algorithm.

Now, define

$$Q_u(i) = \mathbb{P}\left(Z_u(k) = i\right),$$

where

$$\begin{aligned} Q_u(i) &= P_u(i)(1 - R_u) + (1 - P_u(i))R_u \\ &= P_u(i) + (1 - 2P_u(i))R_u. \end{aligned} \tag{25}$$

The vectors $\mathbf{Q}_u$ and $\mathbf{Q}$ which contain the obfuscated probabilities are defined as below:

$$\mathbf{Q}_u = \begin{bmatrix} Q_u(1) \\ Q_u(2) \\ \vdots \\ Q_u(r-1) \end{bmatrix}, \quad \mathbf{Q} = [\mathbf{Q}_1 \;\; \mathbf{Q}_2 \;\; \cdots \;\; \mathbf{Q}_n] ,$$

[2] It is desirable that our results are true over the largest set of strategies that users can employ. In fact, our results would apply to a general set of distributions and are true for any random noise with support that extends out to the maximum amount of a_n. The reason that we have used a uniformly random noise is that we want to have a similar mechanism as [35] to have a good comparison between the results of this paper and [35] to show that dependency is a significant detriment to the privacy of users.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2020.2985059, IEEE Transactions on Information Theory

and the vector containing the permutation of those probabilities after anonymization is $\mathbf{W}$. Thus,

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

The following theorem states that if the number of observations per user m is significantly larger than $n^{\frac{2}{s(r-1)}}$ and the amount of noise level a_n is significantly smaller that $n^{-\frac{1}{s(r-1)}}$ in the r-state i.i.d. model, where s is the size of a group, then the adversary can successfully de-anonymize and de-obfuscate all of the data at different time samples $(k = 1, 2, \cdots, m)$ for all of the users in that group as the number of users in the network (n) goes to infinity.

Theorem 3. 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 above, the size of the group including User 1 is s, and

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

then, User 1 has no privacy at time k.

Discussion 7: It is insightful to compare this result to [35, Theorem 2]. We can see that when users' traces are dependent, the required level of obfuscation and anonymization to achieve privacy is significantly higher. Therefore, we see that dependency can significantly reduce the privacy of users. However, note that the asymptotic noise level is still zero in this case. Specifically, if $A_m(u) = \frac{|\{k : Z_u(k) \neq X_u(k)\}|}{m}$, then as $n \to \infty$,

$$\mathbb{E}[A_m(u)] = \mathbb{E}[A_m] = O\left(n^{-\frac{1}{s(r-1)}-\beta}\right) \to 0.$$

Proof of Theorem 3:

Proof. The proof of Theorem 3 is similar to the proof of Theorem 1 and consists of three parts:

- **First step:** Showing the adversary can reconstruct the association graph of the obfuscated and anonymized version of the data with an arbitrarily small error probability.
- **Second step:** Showing the adversary can uniquely identify Group 1 with an arbitrarily small error probability.
- **Third step:** Showing the adversary can successfully identify all of the members of Group 1 with an arbitrarily small error probability.

First step: Reconstruction of the association graph: In Lemma 1, we show that for the case of anonymization, the adversary can reconstruct the entire association graph of the anonymized data with an arbitrarily small error probability if the number of the adversary's observations per user (m) is bigger than $(\log n)^3$. Since obfuscation is done independently (from other users' obfuscation and from users' data), it does not change the association graph. Therefore, since $n^{\frac{2}{s(r-1)}} > (\log n)^3$, we can use Lemma 1 to show the adversary can reconstruct the association graph of the obfuscated and anonymized data with an arbitrarily small error probability.

Second step: Identifying Group 1 among all of the groups: Now, assume the size of Group 1 is s. Without loss of generality, suppose the members of Group 1 are users $\{1, 2, \cdots, s\}$, so there are at most $\frac{n}{s}$ groups of size s. We call these Groups $1, 2, \cdots, \frac{n}{s}$. The adversary needs to first identify the Group 1 among all of these groups.

As in Section III-A, Σ_s is defined as the set of all permutation on s elements, $\mathbf{P}^{(l)}$ is a vector which contains the probability distributions of users belong to Group l, and $\widetilde{\mathbf{P}_{\Pi}^{(l)}}$ is a vector which contains the estimate of the adversary about the probability distribution of users belonging to Group l. For example, For Group 1, we have

$$\mathbf{P}^{(1)} = \begin{bmatrix}\mathbf{p}_1 & \mathbf{p}_2 & \cdots & \mathbf{p}_s\end{bmatrix},$$

and

$$\widetilde{\mathbf{P}_{\Pi}^{(1)}} = \begin{bmatrix}\widetilde{\mathbf{p}_{\Pi(1)}} & \widetilde{\mathbf{p}_{\Pi(2)}} & \cdots & \widetilde{\mathbf{p}_{\Pi(s)}}\end{bmatrix}.$$

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

- $\mathbb{P}\left(D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}_{\Pi}^{(1)}}\right) \leq \Delta_n\right) \to 1,$
- $\mathbb{P}\left(\bigcup_{l=2}^{\frac{n}{s}}\left\{D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}_{\Pi}^{(l)}}\right) \leq \Delta_n\right\}\right) \to 0\ ,$

where $\Delta_n = n^{-\frac{1}{s(r-1)}-\frac{\alpha}{4}}$. As in Section III-A,

$$\begin{aligned}\mathcal{F}^{(n)} =& \Big\{(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_s) \in \left(\mathbb{R}^{(r-1)}\right)^s :\\ &\qquad \max_u\{|\mathbf{x}_u - \mathbf{p}_u|\} \leq \Delta_n, u = 1, 2, \cdots, s\Big\},\end{aligned}$$

$$\begin{aligned}\mathcal{H}^{(n)} =& \Big\{(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_s) \in \left(\mathbb{R}^{(r-1)}\right)^s :\\ &\qquad \max_u\{|\mathbf{x}_u - \mathbf{p}_u|\} \leq 2\Delta_n, u = 1, 2, \cdots, s\Big\}.\end{aligned}$$

In the first step, we prove, for large enough n,

$$\mathbb{P}\left(D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}_{\Pi}^{(1)}}\right) \leq \Delta_n\right) \to 1.$$

Note that for all $u \in \{1, 2, \cdots, n\}$ and all $i \in \{1, 2, \cdots, r-1\}$, the adversary computes $\widetilde{p_u(i)}$ as follows:

$$\widetilde{p_u(i)} = \frac{|\{k : Y_u(k) = i\}|}{m}, \tag{26}$$

and as a result,

$$\widetilde{p_{\Pi(u)}(i)} = \frac{|\{k : Z_u(k) = i\}|}{m} = \frac{\breve{M}_u(i)}{m}, \tag{27}$$

where $\breve{M}_u(i) = |\{k : Z_u(k) = i\}|$. Now, for all $u \in \{1, 2, \cdots, n\}$ and all $i \in \{0, 1, \cdots, r-1\}$, we have

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{\breve{M}_u(i)}{m} - p_u(i)\right| \leq \Delta_n\right)\\ &= \mathbb{P}\left(p_u(i) - \Delta_n \leq \frac{\breve{M}_u(i)}{m} \leq p_u(i) + \Delta_n\right)\\ &= \mathbb{P}\Bigg(p_u(i) - \Delta_n - q_u(i) \leq \frac{\breve{M}_u(i)}{m} - q_u(i)\\ &\qquad\qquad \leq p_u(i) + \Delta_n - q_u(i)\Bigg).\end{aligned}$$

Note that according to (25), for all $u \in \{1, 2, \cdots, n\}$ and all $i \in \{0, 1, \cdots, r-1\}$, we have

$$\begin{aligned} |p_u(i) - q_u(i)| &= |1 - 2p_u(i)| R_u \\ &\leq R_u \leq a_n, \end{aligned}$$

so, we can conclude for all $u \in \{1, 2, \cdots, n\}$ and all $i \in \{0, 1, \cdots, r-1\}$,

$$\begin{aligned} &\mathbb{P}\left(\left|\frac{\breve{M}_u(i)}{m} - p_u(i)\right| \leq \Delta_n\right) \\ &\geq \mathbb{P}\left(-\Delta_n + a_n \leq \frac{\breve{M}_u(i)}{m} - q_u(i) \leq -a_n + \Delta_n\right) \\ &= \mathbb{P}\left(\left|\frac{\breve{M}_u(i)}{m} - q_u(i)\right| \leq \Delta_n - a_n\right). \end{aligned} \tag{28}$$

By employing a Chernoff bound, we have

$$\begin{aligned} &\mathbb{P}\left(\left|\frac{\breve{M}_u(i)}{m} - q_u(i)\right| \leq \Delta_n - a_n\right) \geq 1 - 2\exp\left(-\frac{m(\Delta_n - a_n)^2}{3q_u(i)}\right) \\ &\geq 1 - 2\exp\left(-\frac{1}{3}\left(cn^{\frac{2}{s(r-1)}+\alpha}\right)\left(\frac{1}{n^{\frac{1}{s(r-1)}+\frac{\alpha}{4}}} - \frac{c'}{n^{\frac{1}{s(r-1)}+\beta}}\right)^2\right) \end{aligned} \tag{29}$$

Now from (28) and (29), we can conclude for all $u \in \{1, 2, \cdots, n\}$ and all $i \in \{0, 1, \cdots, r-1\}$,

$$\begin{aligned} &\mathbb{P}\left(\left|\frac{\breve{M}_u(i)}{m} - p_u(i)\right| \leq \Delta_n\right) \\ &\geq 1 - 2\exp\left(-\frac{1}{3}\left(cn^{\frac{2}{s(r-1)}+\alpha}\right)\left(\frac{1}{n^{\frac{1}{s(r-1)}+\frac{\alpha}{4}}} - \frac{c'}{n^{\frac{1}{s(r-1)}+\beta}}\right)^2\right), \end{aligned}$$

and

$$\begin{aligned} &\mathbb{P}\left(\left|\frac{\breve{M}_u(i)}{m} - p_u(i)\right| \geq \Delta_n\right) \\ &\leq 2\exp\left(-\frac{1}{3}\left(cn^{\frac{2}{s(r-1)}+\alpha}\right)\left(\frac{1}{n^{\frac{1}{s(r-1)}+\frac{\alpha}{4}}} - \frac{c'}{n^{\frac{1}{s(r-1)}+\beta}}\right)^2\right). \end{aligned} \tag{30}$$

Now, for all $u \in$ Group 1, all $i \in \{1, 2, \cdots, r-1\}$ and any $\beta > \frac{\alpha}{4}$, (30) and the union bound yield

$$\begin{aligned} &\mathbb{P}\left(D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}_{\Pi}^{(1)}}\right) \geq \Delta_n\right) \leq \sum_{u=1}^{s}\sum_{i=1}^{r-1}\mathbb{P}\left(\left|\frac{\breve{M}_u(i)}{m} - p_u(i)\right| \geq \Delta_n\right) \\ &\leq 2s(r-1)\exp\left(-\frac{1}{3}\left(cn^{\frac{2}{s(r-1)}+\alpha}\right)\left(\frac{1}{n^{\frac{1}{s(r-1)}+\frac{\alpha}{4}}} - \frac{c'}{n^{\frac{1}{s(r-1)}+\beta}}\right)^2\right). \end{aligned} \tag{31}$$

The right-side of (31) goes to 0 as $n \to \infty$. As a result,

$$\mathbb{P}\left(D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}_{\Pi}^{(1)}}\right) \leq \Delta_n\right) \to 1,$$

as $n \to \infty$.

In the next step, we prove $\mathbb{P}\left(\bigcup_{l=2}^{\frac{n}{s}}\left\{D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}_{\Pi}^{(l)}}\right) \leq \Delta_n\right\}\right) \to 0$. For all groups other than Group 1, we have

$$(4\Delta_n)^{s(r-1)}\delta_1 \leq \mathbb{P}\left(\mathbf{P}^{(l)} \in \mathcal{H}'^{(n)}\right) \leq (4\Delta_n)^{s(r-1)}\delta_2,$$

and as a result,

$$\begin{aligned} \mathbb{P}\left(\mathbf{P}^{(l)} \in \mathcal{H}^{(n)}\right) &\leq \delta_2(4\Delta_n)^{s(r-1)} \\ &= \delta_2 4^{s(r-1)}\frac{1}{n^{1+\frac{\alpha}{4}s(r-1)}}. \end{aligned}$$

Similarly, for any $\sigma \in \Sigma_s$,

$$\begin{aligned} \mathbb{P}\left(\mathbf{P}_{\sigma}^{(l)} \in \mathcal{H}'^{(n)}\right) &\leq \delta_2(4\Delta_n)^{s(r-1)} \\ &= \delta_2 4^{s(r-1)}\frac{1}{n^{1+\frac{\alpha}{4}s(r-1)}}, \end{aligned}$$

and since $|\Sigma_s| = s!$, by a union bound,

$$\begin{aligned} \mathbb{P}\left(\bigcup_{l=2}^{\frac{n}{s}}\left\{\bigcup_{\sigma\in\Sigma_s}\left\{\mathbf{P}_{\sigma}^{(l)} \in \mathcal{H}^{(n)}\right\}\right\}\right) &\leq \sum_{l=2}^{\frac{n}{s}}\sum_{\sigma\in\Sigma_s}\mathbb{P}\left(\mathbf{P}_{\sigma}^{(l)} \in \mathcal{H}^{(n)}\right) \\ &\leq \frac{n}{s}s!\delta_2 4^{s(r-1)}\frac{1}{n^{1+\frac{\alpha}{4}s(r-1)}} \\ &= (s-1)!4^{s(r-1)}\delta_2 n^{-\frac{\alpha}{4}s(r-1)}. \end{aligned} \tag{32}$$

The right-side of (32) goes to 0 as $n \to \infty$. Thus, all $\mathbf{P}^{(l)}$'s are outside of $\mathcal{H}^{(n)}$ with high probability.

Now, we claim that given all $\mathbf{P}^{(l)}$'s are outside of $\mathcal{H}^{(n)}$, $\mathbb{P}\left(\bigcup_{l=2}^{\frac{n}{s}}\left\{D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}_{\Pi}^{(l)}}\right) \leq \Delta_n\right\}\right)$ is arbitrarily small. In other words, by using a Chernoff bound, it is shown $\widetilde{\mathbf{P}^{(l)}}$'s are close to $\mathbf{P}^{(l)}$'s, and they will be outside of $\mathcal{F}^{(n)}$. Thus, for all $u \in$ Group l and all $i \in \{1, 2, \cdots, r-1\}$, (30) and the union bound yield

$$\begin{aligned} &\mathbb{P}\left(D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}_{\Pi}^{(l)}}\right) \leq \Delta_n\right) = \mathbb{P}\left(D\left(\mathbf{P}^{(l)}, \widetilde{\mathbf{P}_{\Pi}^{(l)}}\right) \geq \Delta_n\right) \\ &\leq \sum_{u=1}^{s}\sum_{i=1}^{r-1}\mathbb{P}\left(\left|\frac{\breve{M}_u(i)}{m} - p_u(i))\right| \geq \Delta_n\right) \\ &\leq 2s(r-1)\exp\left(-\frac{1}{3}\left(cn^{\frac{2}{s(r-1)}+\alpha}\right)\left(\frac{1}{n^{\frac{1}{s(r-1)}+\frac{\alpha}{4}}} - \frac{c'}{n^{\frac{1}{s(r-1)}+\beta}}\right)^2\right). \end{aligned}$$

Now, by using a union bound again, we can conclude that, for any $\beta > \frac{\alpha}{4}$,

$$\begin{aligned} &\mathbb{P}\left(\bigcup_{l=2}^{\frac{n}{s}}\left\{D\left(\mathbf{P}^{(l)}, \widetilde{\mathbf{P}_{\Pi}^{(l)}}\right) \geq \Delta_n\right\}\right) \leq \sum_{l=2}^{\frac{n}{s}}\mathbb{P}\left(D\left(\mathbf{P}^{(l)}, \widetilde{\mathbf{P}_{\Pi}^{(l)}}\right) \geq \Delta_n\right) \\ &\leq 2n(r-1)\exp\left(-\frac{1}{3}\left(cn^{\frac{2}{s(r-1)}+\alpha}\right)\left(\frac{1}{n^{\frac{1}{s(r-1)}+\frac{\alpha}{4}}} - \frac{c'}{n^{\frac{1}{s(r-1)}+\beta}}\right)^2\right). \end{aligned} \tag{33}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2020.2985059, IEEE Transactions on Information Theory

The right-side of (33) goes to 0 as $n \longrightarrow \infty$. Thus, we have shown that for all $l \in \{1, 2, \cdots, \frac{n}{s}\}$, $\widetilde{\mathbf{P}^{(l)}}$'s are close to $\mathbf{P}^{(l)}$, which are outside of set $\mathcal{F}^{(n)}$. As a result, as $n \to \infty$,

$$\mathbb{P}\left(\bigcup_{l=2}^{\frac{n}{s}} \left\{D\left(\mathbf{P}^{(1)}, \widetilde{\mathbf{P}_{\Pi}^{(l)}}\right) \leq \Delta_n\right\}\right) \to 0.$$

Third step: Identifying User 1 **among all of the members of Group** 1**:** In this step, we need to prove that after identifying Group 1, the adversary can correctly identify each member. In other words, the adversary should identify the permutation of Group 1.

From (26) and (27), for all $u \in \{1, 2, \cdots, s\}$ and all $i \in \{1, 2, \cdots, r-1\}$, we have

$$\widetilde{p_u(i)} = \frac{|\{k : Y_u(k) = i\}|}{m},$$

and as a result,

$$\widetilde{p_{\Pi(u)}(i)} = \frac{|\{k : Z_u(k) = i\}|}{m} = \frac{\breve{M}_u(i)}{m},$$

where $\breve{M}_u(i) = |\{k : Z_u(k) = i\}|$.

As in Section III-A, we define sets $\mathcal{B}^{(n)}$ and $C^{(n)}$ as

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

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

where $\Delta_n = n^{-\frac{1}{s(r-1)} - \frac{\alpha}{4}}$. We claim that for $m = cn^{\frac{2}{s(r-1)}+\alpha}$ and $a_n = c'n^{-\left(\frac{1}{s(r-1)}+\beta\right)}$,

1) $\mathbb{P}\left(\widetilde{\mathbf{p}_{\Pi(1)}} \in \mathcal{B}^{(n)}\right) \to 1,$
2) $\mathbb{P}\left(\bigcup\limits_{u=2}^{s} \left\{\widetilde{\mathbf{p}_{\Pi(u)}} \in \mathcal{B}^{(n)}\right\}\right) \to 0,$

as $n \longrightarrow \infty$. Thus, the adversary can identify $\Pi(1)$ by examining $\widetilde{\mathbf{p}_u}$'s and choosing the only one that belongs to $\mathcal{B}^{(n)}$.

In the first step, we show that as n goes to infinity,

$$\mathbb{P}\left(\widetilde{\mathbf{p}_{\Pi(1)}} \in \mathcal{B}^{(n)}\right) \longrightarrow 1.$$

According to (30) and the union bound, for all $u \in$ Group 1 and all $i \in \{1, 2, \cdots, r-1\}$, we have

$$\mathbb{P}\left(\widetilde{\mathbf{p}_{\Pi(1)}} \notin \mathcal{B}^{(n)}\right) \leq \sum_{i=1}^{r-1} \mathbb{P}\left(\left|\frac{\breve{M}_1(i)}{m} - p_1(i)\right| \geq \Delta_n\right)$$
$$\leq (r-1)\left(2\exp\left(-\frac{1}{3}\left(cn^{\frac{2}{s(r-1)}+\alpha}\right)\left(\frac{1}{n^{\frac{1}{s(r-1)}+\frac{\alpha}{4}}} - \frac{c'}{n^{\frac{1}{s(r-1)}+\beta}}\right)^2\right)\right).$$

The right-side of the above equation goes to 1 as $n \longrightarrow \infty$, thus we can conclude,

$$\mathbb{P}\left(\widetilde{\mathbf{p}_{\Pi(1)}} \in \mathcal{B}^{(n)}\right) \longrightarrow 1. \tag{34}$$

Now, in the next step, we need to show that as n goes to infinity,

$$\mathbb{P}\left(\bigcup_{u=2}^{s} \left\{\widetilde{\mathbf{p}_{\Pi(u)}} \in \mathcal{B}^{(n)}\right\}\right) \to 0.$$

We show that as n goes to infinity,

$$\mathbb{P}\left(\bigcup_{u=2}^{s} \left\{\mathbf{p}_u \in C'^{(n)}\right\}\right) \to 0.$$

Note for all $u \in \{2, 3, \cdots, s\}$,

$$4\left(\Delta_n\right)^{r-1}\delta_1 < \mathbb{P}\left(\mathbf{p}_u \in C'^{(n)}\right) < 4\left(\Delta_n\right)^{r-1}\delta_2,$$

and according to the union bound,

$$\begin{aligned}\mathbb{P}\left(\bigcup_{u=2}^{s} \left\{\mathbf{p}_u \in C^{(n)}\right\}\right) &\leq \sum_{u=2}^{s} \mathbb{P}\left(\mathbf{p}_u \in C^{(n)}\right) \\ &\leq 4s\left(\Delta_n\right)^{r-1}\delta_2 \\ &\leq 4s\frac{1}{n^{\frac{1}{s}+\frac{\alpha(r-1)}{4}}}\delta_2. \end{aligned} \tag{35}$$

The right-side of (35) goes to 0 as $n \longrightarrow \infty$. Thus, all $\mathbf{p}_u$'s are outside of $C^{(n)}$ with high probability.

Now, we claim that given all $\mathbf{p}_u$'s are outside of $C^{(n)}$, $\mathbb{P}\left(\widetilde{\mathbf{p}_{\Pi(u)}} \in \mathcal{B}^{(n)}\right)$ is arbitrarily small. Note that for all $u \in \{2, 3, \cdots, s\}$ and all $i \in \{1, 2, \cdots, r-1\}$, (30) and the union bounds yield

$$\mathbb{P}\left(\widetilde{\mathbf{p}_{\Pi(u)}} \in \mathcal{B}^{(n)}\right) \leq \mathbb{P}\left(\left|\widetilde{\mathbf{p}_{\Pi(u)}} - \mathbf{p}_u\right| \geq \Delta_n\right)$$
$$\leq \sum_{i=1}^{r-1} \mathbb{P}\left(\left|\widetilde{p_{\Pi(u)}(i)} - p_u(i)\right| \geq \Delta_n\right)$$
$$\leq 2(r-1)\exp\left(-\frac{1}{3}\left(cn^{\frac{2}{s(r-1)}+\alpha}\right)\left(\frac{1}{n^{\frac{1}{s(r-1)}+\frac{\alpha}{4}}} - \frac{c'}{n^{\frac{1}{s(r-1)}+\beta}}\right)^2\right).$$

As a result, by using a union bound again,

$$\mathbb{P}\left(\bigcup_{u=2}^{s} \left\{\left|\widetilde{\mathbf{p}_{\Pi(u)}} - \mathbf{p}_u\right| \geq \Delta_n\right\}\right)$$
$$\leq 2s(r-1)\exp\left(-\frac{1}{3}\left(cn^{\frac{2}{s(r-1)}+\alpha}\right)\left(\frac{1}{n^{\frac{1}{s(r-1)}+\frac{\alpha}{4}}} - \frac{c'}{n^{\frac{1}{s(r-1)}+\beta}}\right)^2\right). \tag{36}$$

Now, the right-side of (36) goes to 0 as $n \to \infty$. Thus, for all $u \in \{2, 3, \cdots, s\}$, $\widetilde{\mathbf{p}_{\Pi(u)}}$'s are close to $\mathbf{p}_u$'s, so they will be outside of $\mathcal{B}^{(n)}$. Now, we can conclude as $n \to \infty$:

$$\mathbb{P}\left(\bigcup_{u=2}^{s} \left\{\widetilde{\mathbf{p}_{\Pi(u)}} \in \mathcal{B}^{(n)}\right\}\right) \to 0.$$

Hence, the adversary can successfully recover $Z_1(k)$. Since $Z_1(k) = X_1(k)$ with probability $1-R_u = 1-o(1)$, the adversary can recover $X_1(k)$ with vanishing error probability. □

B. r-State Markov Chain Model

In this section, users' data patterns are modeled using Markov chains and there are r possibilities for users' data samples. Similar to Section III-B, we assume $p_u(i)$'s are drawn independently from some continuous density function, $f_{\mathbf{P}}(\mathbf{p}_u)$, on the $(0,1)^{|E|-r}$ hypercube, and $\mathbf{p}_u$, $f_{\mathbf{P}}(\mathbf{p}_u)$, and $\mathcal{R}_{\mathbf{P}}$ are defined as in Section III-B.

By using the general idea stated in Section III-B, we can now repeat similar reasoning as in the proof of Theorem 3 to show the following theorem.

Theorem 4. For an irreducible, aperiodic Markov chain model, if $\mathbf{Z}$ is the obfuscated version of $\mathbf{X}$, and $\mathbf{Y}$ is the anonymized version of $\mathbf{Z}$ as defined above, the size of the group including User 1 is s, and

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

then, User 1 has no privacy at time k.

V. More General Setting for the Association Graph

The association graph structure that we have studied so far was general except for one aspect: we assumed that people in a group have dependency but that they are completely independent from members of other groups. But, in practice there could be weak dependency between members of each group and outside members. Here we discuss how to apply our results to this more general setting.

Similar to [72]–[81], we consider a community structure with strong intra-community connections and weak inter-community connections. In the community structure, the nodes of the network can be grouped into sets of users such that each set of users is densely connected internally as shown in Figure 7a. Here, we also assume that the adversary has some knowledge about the correlation between users in addition to the marginal probability distributions: the adversary know whether the value of each correlation is less than or higher than a specific threshold. We show that the adversary can reliably reconstruct the entire association graph for *the anonymized version of the data* (i.e., the observed data traces) with relatively few observations.

Let $G(\mathcal{V}, F)$ denote the association graph with set of nodes $\mathcal{V}$, ($|\mathcal{V}| = n$), and set of edges F. In this case, we use an association graph based on a threshold as follows: we assume two vertices (users) are connected if their data sets are strongly correlated, and are not connected if their data sets are weakly correlated. More specifically,

- $(u, u') \notin F$ if and only if $\text{Cov}\left(X_u(k); X_{u'}(k)\right) \leq \epsilon_1$,
- $(u, u') \in F$ if and only if $\text{Cov}\left(X_u(k); X_{u'}(k)\right) \geq \epsilon_2$,

where $\text{Cov}\left(X_u(k); X_{u'}(k)\right)$ is the covariance between the k^{th} data point of user u and user u'.

Lemma 2. Consider a general association graph, $G(\mathcal{V}, F)$, based on the threshold as described above. If the adversary obtains $m = (\log n)^3$ anonymized observations per user, they can construct $\widetilde{G} = \widetilde{G}(\widetilde{\mathcal{V}}, \widetilde{F})$, where $\widetilde{\mathcal{V}} = \{\Pi(u) : u \in \mathcal{V}\} = \mathcal{V}$, such that with high probability, for all $u, u' \in \mathcal{V}$; $(u, u') \in F$ if and only if $(\Pi(u), \Pi(u')) \in \widetilde{F}$. We write this statement as $\mathbb{P}(\widetilde{G} \simeq G) \to 1$.

Proof. Note that for $u, u' \in \{1, 2, \cdots, n\}$, we write $v = \Pi(u)$ and $v' = \Pi(u')$. We provide an algorithm for the adversary that with high probability obtains all edges of F correctly. For each pair w and w', the adversary computes $\widetilde{Cov_{vv'}}$ as follows:

$$\widetilde{Cov_{vv'}} = \frac{\sum\limits_{i=1}^{r-1}\sum\limits_{j=1}^{r-1} ij\widehat{M}_{vv'}(i,j)}{m} - \frac{\sum\limits_{i=1}^{r-1} i\widehat{M}_v(i)}{m}\frac{\sum\limits_{i=1}^{r-1} i\widehat{M}_{v'}(i)}{m} \tag{37}$$

where

$$\widehat{M}_{vv'}(i,j) = |\{k : Y_v(k) = i, Y_{v'}(k) = j\}|.$$

$$\widehat{M}_v(i) = |\{k : Y_v(k) = i\}|.$$

$$\widehat{M}_{v'}(j) = |\{k : Y_{v'}(k) = j\}|.$$

After observing $m = (\log n)^3$ data points per user and computing the above expressions, the adversary constructs $\widetilde{G}$ in the following way:

- If $|\widetilde{Cov_{vv'}}| \leq \epsilon_1$, then $(v, v') \notin \widetilde{F}$.
- If $|\widetilde{Cov_{vv'}}| \geq \epsilon_2$, then $(v, v') \in \widetilde{F}$.

We show the above method yields $\mathbb{P}(\widetilde{G} \simeq G) \to 1$ as $n \to \infty$, as follows. Note

$$\widehat{M}_{vv'}(i,j) \sim \text{Binomial}(m, w_{vv'}(i,j)),$$

$$\widehat{M}_v(i) \sim \text{Binomial}(m, w_v(i)),$$

$$M_{v'}(i) \sim \text{Binomial}(m, w_{v'}(i)),$$

where $w_{vv'}(i,j) = \mathbb{P}\left(Y_v(k) = i, Y_{v'}(k) = j\right)$, $w_v(i) = \mathbb{P}\left(Y_v(k) = i\right)$, and $w_{v'}(i) = \mathbb{P}\left(Y_{v'}(k) = i\right)$. From the proof of Lemma 1, by using (4), for all $v, v' \in \{1, 2, \cdots, n\}$ and all $i, j \in \{0, 1, \cdots, r-1\}$, we have

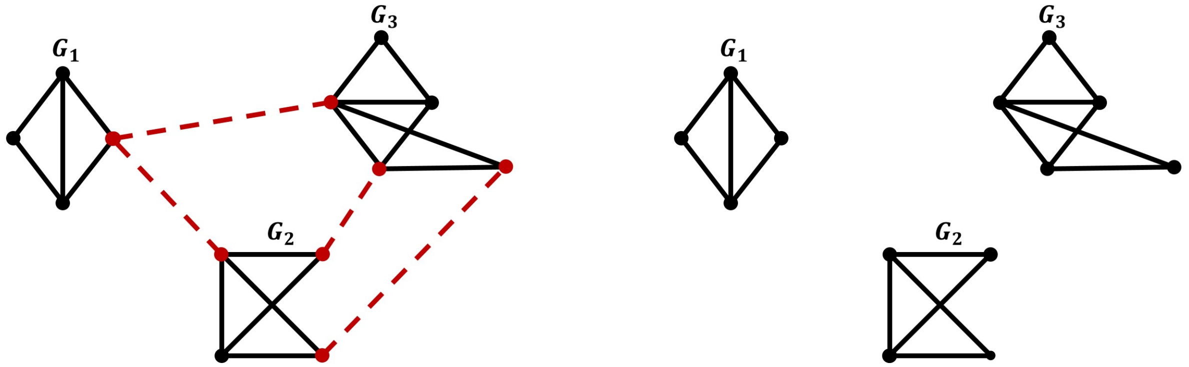
$$0 \leq mw_{vv'}(i,j) - m^{\frac{3}{4}} \leq \widehat{M}_{vv'}(i,j) \leq mw_{vv'}(i,j) + m^{\frac{3}{4}}. \tag{38}$$

$$0 \leq mw_v(i) - m^{\frac{3}{4}} \leq \widehat{M}_v(i) \leq mw_v(i) + m^{\frac{3}{4}}. \tag{39}$$

$$0 \leq mw_{v'}(i) - m^{\frac{3}{4}} \leq \widehat{M}_{v'}(i) \leq mw_v(i) + m^{\frac{3}{4}}. \tag{40}$$

$A_{vv'}(i,j)$ is defined as the event that (38), (39), and (40) are all valid; thus, based on proof of Lemma 1, we have

$$\mathbb{P}\left(\bigcap_{v=1}^{n}\bigcap_{v'=1}^{n}\bigcap_{i=0}^{r-1}\bigcap_{j=0}^{r-1}\{A_{vv'}(i,j)\}\right) \to 1, \tag{41}$$



(a) A sketch of a small network displaying community structure.

(b) The association graph consists of disjoint subgraphs.

Fig. 7: The adversary uses their prior knowledge to break inter-community edges.

as $n \longrightarrow \infty$. Let us define $C_{vv'} = \bigcap_{i=0}^{r-1} \bigcap_{j=0}^{r-1} \{A_{vv'}(i,j)\}$. Now, if $C_{vv'}$ is true for some $v, v' \in \{1, 2, \cdots, n\}$, according to (37), we have

$$\begin{aligned}
\widetilde{Cov_{vv'}} &= \frac{\sum_{i=1}^{r-1}\sum_{j=1}^{r-1} ij\widehat{M}_{vv'}(i,j)}{m} - \frac{\sum_{i=1}^{r-1} i\widehat{M}_{v}(i)}{m}\frac{\sum_{i=1}^{r-1} i\widehat{M}_{v'}(i)}{m} \\
&\leq \frac{\sum_{i=1}^{r-1}\sum_{j=1}^{r-1} ij\left(mw_{vv'}(i,j) + m^{\frac{3}{4}}\right)}{m} \\
&\quad - \frac{\sum_{i=1}^{r-1} i\left(mw_{v}(i) - m^{\frac{3}{4}}\right)}{m}\frac{\sum_{i=1}^{r-1} i\left(mw_{v'}(i) - m^{\frac{3}{4}}\right)}{m} \\
&= \sum_{i=1}^{r-1}\sum_{j=1}^{r-1} ijw_{vv'}(i,j) - \sum_{i=1}^{r-1} iw_{v}(i)\sum_{i=1}^{r-1} iw_{v'}(i) \\
&\quad + \frac{r^2(r-1)^2}{4}m^{-\frac{1}{4}} + \frac{r(r-1)}{2}\sum_{i=1}^{r-1} i(w_v(i) + w_{v'}(i))m^{-\frac{1}{4}} \\
&\quad - \frac{r^2(r-1)^2}{4}m^{-\frac{1}{2}} \\
&\leq Cov_{vv'} + \frac{r^2(r-1)^2}{4}m^{-\frac{1}{4}} \\
&\quad + \frac{r(r-1)}{2}\sum_{i=1}^{r-1} i(w_v(i) + w_{v'}(i))m^{-\frac{1}{4}} + \frac{r^2(r-1)^2}{4}m^{-\frac{1}{2}},
\end{aligned} \tag{42}$$

where $Cov_{vv'} = \sum_{i=1}^{r-1}\sum_{j=1}^{r-1} ijw_{vv'}(i,j) - \sum_{i=1}^{r-1} iw_{v}(i)\sum_{i=1}^{r-1} iw_{v'}(i)$. Similarly,

$$\begin{aligned}
\widetilde{Cov_{vv'}} &= \frac{\sum_{i=1}^{r-1}\sum_{j=1}^{r-1} ij\widehat{M}_{vv'}(i,j)}{m} - \frac{\sum_{i=1}^{r-1} i\widehat{M}w(i)}{m}\frac{\sum_{i=1}^{r-1} i\widehat{M}_{v'}(i)}{m} \\
&\geq \frac{\sum_{i=1}^{r-1} i\left(mw_{vv'}(i) - m^{\frac{3}{4}}\right)}{m} \\
&\quad - \frac{\sum_{i=1}^{r-1} i\left(mw_{v}(i) + m^{\frac{3}{4}}\right)}{m}\frac{\sum_{i=1}^{r-1} i\left(mw_{v'}(i) + m^{\frac{3}{4}}\right)}{m} \\
&= Cov_{vv'} - \frac{r^2(r-1)^2}{4}m^{-\frac{1}{4}} \\
&\quad - \frac{r(r-1)}{2}\sum_{i=1}^{r-1} i(w_v(i) + w_{v'}(i))m^{-\frac{1}{4}} - \frac{r^2(r-1)^2}{4}m^{-\frac{1}{2}}.
\end{aligned} \tag{43}$$

Now, by using (42) and (43), for some $v, v' \in \{1, 2, \cdots, n\}$, we have

$$\begin{aligned}
&\left|\widetilde{Cov_{vv'}} - Cov_{vv'}\right| \\
&\leq \frac{r^2(r-1)^2}{4}m^{-\frac{1}{4}} + \frac{r(r-1)}{2}\sum_{i=1}^{r-1} i(w_v(i) + w_{v'}(i))m^{-\frac{1}{4}} \\
&\quad + \frac{r^2(r-1)^2}{4}m^{-\frac{1}{2}}.
\end{aligned} \tag{44}$$

Let us define event $D_{vv'}$ as the event that (44) is valid; thus, we have shown, for any $v, v' \in \{1, 2, \cdots, n\}$, $C_{vv'} \subseteq D_{vv'}$, and consequently,

$$\left\{\bigcap_{v=1}^{n}\bigcap_{v'=1}^{n}\{C_{vv'}\}\right\} \subseteq \left\{\bigcap_{v=1}^{n}\bigcap_{v'=1}^{n}\{D_{vv'}\}\right\}.$$

As a result,

$$\mathbb{P}\left(\bigcap_{v=1}^{n}\bigcap_{v'=1}^{n}\{D_{vv'}\}\right) \geq \mathbb{P}\left(\bigcap_{v=1}^{n}\bigcap_{v'=1}^{n}\{C_{vv'}\}\right).$$

Thus, by using (41), we have

$$\mathbb{P}\left(\bigcap_{v=1}^{n}\bigcap_{v'=1}^{n}\{D_{vv'}\}\right) \longrightarrow 1,$$

as $n \longrightarrow \infty$. Hence, with high probability, (44) is simultaneously valid for all $v, v' \in \{1, 2, \cdots, n\}$. Thus, we can conclude, with high probability, for all $v, v' \in \{1, 2, \cdots, n\}$, $\widetilde{Cov_{vv'}}$'s are close to $Cov_{vv'}$'s.

Now, if (u, u') is an inter-community edge, the adversary knows $Cov_{uu'} \leq \epsilon_1$, and as a result, $Cov_{vv'} \leq \epsilon_1$; thus, the adversary removes that edge. Now, we can conclude $(v, v') \notin \widetilde{F}$, and in other words, $(\Pi(u), \Pi(u')) \notin \widetilde{F}$. This is true with high probability, simultaneously for all $u, u' \in \{1, 2, \cdots, n\}$ where (u, u') is an inter-community edge.

In addition, if (u, u') is an intra-community edge, the adversary knows $Cov_{uu'} \geq \epsilon_2$, and as a result, $Cov_{vv'} \geq \epsilon_2$. Now, we can conclude $(v, v') \in \widetilde{F}$, and in other words, $(\Pi(u), \Pi(u')) \in \widetilde{F}$. This is true with high probability, simultaneously for all $u, u' \in \{1, 2, \cdots, n\}$ where (u, u') is an intra-community edge.

As a result, for large enough n, we have $\mathbb{P}\left(\widetilde{G} \simeq G\right) \to 1$, so the adversary can reconstruct the association graph of the anonymized version of the data which is based on a threshold with an arbitrarily small error probability. □

Now, the adversary has observed the graph structure shown in Figure 7b, where subgraph G_1 is a connected graph with s_1 vertices which is disjoint from the remainder of the association graph $(G' = G - G_1)$. In other words,

$$G = G_1 \cup G'.$$

Now, we can repeat the same reasoning as that in the proof of Theorem 1, Theorem 2, Theorem 3, and Theorem 4 to obtain the same results for this case.

Discussion 8: The stochastic block model is a generative model for random graphs [98]–[104]. Note that there are two key differences between the stochastic block model and the work here. First, in the stochastic block model, the edge set is sampled at random and the probability distributions of edges are the key part of the work, while here the analysis is based on the users' data traces, and the statistical knowledge of the adversary is a key part. Second, in the stochastic block model, nodes within a community connect to nodes in other communities in an equivalent way. In other words, any two vertices $u \in C_i$ and $v \in C_j$ are connected by an edge with probability p_{ij}, where C_i and C_j are different blocks, so all edges between two communities have the same weights or strengths. Here, as shown in Figure 7a, there is no need that an inter-community edge corresponding to the correlation of nodes in separate communities, has the same value as others; in other words, there is no need for the nodes in a community to connect to the nodes in other communities in an equivalent way. In our work, for each of the intra-community edges, $Cov_{uu'} \geq \epsilon_2$, and for each of the inter-community edges, $Cov_{uu'} \leq \epsilon_1$; thus, edges can have different weights.

VI. Improving Privacy in the Presence of Dependency

In the previous parts of this paper, we argued and demonstrated that inter-user dependency degrades the privacy provided by standard privacy-preserving mechanisms (PPMs). In this section, we discuss how to design PPMs considering inter-user dependency in order to better preserve privacy. First, note that independent obfuscation alone cannot be sufficient even at a high noise level, because it does not change the association graph, and thus, the adversary can still reconstruct the association graph with a small number of observations. To mitigate this issue, we suggest that associated users collaborate in applying the noise when deploying a PPM.

For clarity, we focus on the two-state i.i.d. case $(r = 2)$. In the first part, we also focus on the case when the association graph consists of subgraphs with the size of each of them less than or equal to 2 $(s_l \leq 2)$. Thus, as shown in Figure 8, there are some connected users and there are also some isolated users. First, we state the following lemma.

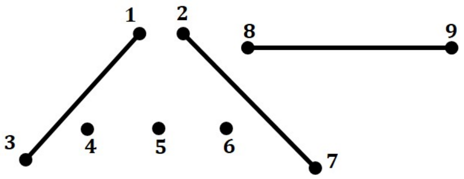


Fig. 8: Graph G consists of some subgraphs (G_l) with $s_l \leq 2$.

Lemma 3. Let $X_u(k) \sim \text{Bernoulli}(p_u)$ and $X_{u'}(k) \sim \text{Bernoulli}(p_{u'})$; then, there exists an obfuscation technique with a noise level equal to

$$\check{a}(u, u') = \frac{\text{Cov}(X_u(k), X_{u'}(k))}{\max\{p_u, p_{u'}, 1 - p_u, 1 - p_{u'}\}},$$

for the dataset of user u and user u' such that $\check{Z}_u(k)$ and $\check{Z}_{u'}(k)$ are independent from each other. Note $\check{Z}_u(k)$ and $\check{Z}_{u'}(k)$ are the k^{th} (reported) data point of user u and u', respectively, after applying obfuscation with the noise level equal to $\check{a}(u, u')$.

Proof. Let $X_u(k) \sim \text{Bernoulli}\,(p_u)$ and $X_{u'}(k) \sim \text{Bernoulli}\,(p_{u'})$. Then, to make these two sequences independent, it suffices if $\check{Z}_u(k)|\check{Z}_{u'}(k) = 0$ has the same distributions as $\check{Z}_u(k)|\check{Z}_{u'}(k) = 1$. We provide the proof for the case $\max\{p_u, p_{u'}, 1 - p_u, 1 - p_{u'}\} = 1 - p_{u'}$; the proofs of the other cases are similar. Now, If $X_u(k) = 1$ and $X_{u'}(k) = 1$, we pass $X_u(k)$ through a $BSC(\Upsilon)$ in order to obtain $\check{Z}_u(k)$. Thus,

$$\frac{\mathbb{P}\left(X_u(k) = 1, X_{u'}(k) = 0\right)}{1 - p_{u'}} = \frac{\mathbb{P}\left(X_u(k) = 1, X_{u'}(k) = 1\right)(1 - \Upsilon)}{p_{u'}},$$

and Υ can be calculated as

$$\Upsilon = 1 - \frac{p_{u'}}{1 - p_{u'}} \frac{\mathbb{P}\left(X_u(k) = 1, X_{u'}(k) = 0\right)}{\mathbb{P}\left(X_u(k) = 1, X_{u'}(k) = 1\right)}.$$

Now, we can conclude,

$$\begin{aligned}
\check{a}(u, u') &= \Upsilon \mathbb{P}\left(X_u(k) = 1, X_{u'}(k) = 1\right) \\
&= \left(1 - \frac{p_{u'}}{1 - p_{u'}} \frac{\mathbb{P}\left(X_u(k) = 1, X_{u'}(k) = 0\right)}{\mathbb{P}\left(X_u(k) = 1, X_{u'}(k) = 1\right)}\right) \\
&\qquad \cdot \mathbb{P}\left(X_u(k) = 1, X_{u'}(k) = 1\right) \\
&= \mathbb{P}\left(X_u(k) = 1, X_{u'}(k) = 1\right) \\
&\qquad - \frac{p_{u'}}{1 - p_{u'}} \mathbb{P}\left(X_u(k) = 1, X_{u'}(k) = 0\right) \\
&= \frac{\mathbb{P}\left(X_u(k) = 1, X_{u'}(k) = 1\right) - p_u p_{u'}}{1 - p_{u'}} \\
&= \frac{\text{Cov}(X_u(k), X_{u'}(k))}{\max\{p_u, p_{u'}, 1 - p_u, 1 - p_{u'}\}}.
\end{aligned}$$

Next, we explain the idea behind this lemma by an example.

Example 1. Let $X_u(k) \sim \text{Bernoulli}\left(\frac{3}{5}\right)$ and $X_{u'}(k) \sim \text{Bernoulli}\left(\frac{1}{5}\right)$, with the joint probability mass function of $X_u(k)$ and $X_{u'}(k)$ as given in Table I

TABLE I: Joint probability mass function of $X_u(k)$ and $X_{u'}(k)$.

$X_u(k)$ \ $X_{u'}(k)$	0	1
0	$\frac{7}{20}$	$\frac{1}{20}$
1	$\frac{9}{20}$	$\frac{3}{20}$

TABLE II: The expected results of $X_u(k)$ and $X_{u'}(k)$ according to Table I after observing 2000 bits of data.

$X_u(k)$ \ $X_{u'}(k)$	0	1
0	700	100
1	900	300

TABLE III: The desired results to make $\check{Z}_u(k)$ and $\check{Z}_{u'}(k)$ independent from each other after observing 2000 bits of data.

$\check{Z}_u(k)$ \ $\check{Z}_{u'}(k)$	0	1
0	700	175
1	900	225

If we observe 2000 bits of data, Table II shows the expected results according to Table I.

Then, to make $\check{Z}_u(k)$ and $\check{Z}_{u'}(k)$ independent, it is sufficient for $\check{Z}_u(k)|\check{Z}_{u'}(k) = 0$ to have the same distribution as $\check{Z}_u(k)|\check{Z}_{u'}(k) = 1$. This means we should have

$$\frac{\mathbb{P}\left(\check{Z}_u(k) = 1, \check{Z}_{u'}(k) = 1\right)}{\mathbb{P}\left(\check{Z}_{u'}(k) = 1\right)} = \frac{\mathbb{P}\left(\check{Z}_u(k) = 1, \check{Z}_{u'}(k) = 0\right)}{\mathbb{P}\left(\check{Z}_{u'}(k) = 0\right)};$$

thus, according to Table III,

$$\frac{300(1-\Upsilon)}{100+300} = \frac{900}{700+900} \rightarrow \Upsilon = \frac{1}{4},$$

where Υ is the portion of data points $X_u(k)$ that need to be changed in the fourth region of Table II (i.e., the region $X_u(k) = 1, X_{u'}(k) = 1$). Now, we need to change $\frac{1}{4} \cdot 300 = 75$ of the data bits. As a result, if $X_u(k) = 1$ and $X_{u'}(k) = 1$, then we pass $X_u(k)$ through a $BSC(\frac{1}{4})$, and obtain $\check{Z}_u(k)$. Hence, the asymptotic noise level is equal to

$$\check{a}(u,u') = \frac{3}{20} \cdot \frac{1}{4} = 3.75\%.$$

It is easy to check that the asymptotic noise level will be given by the equation in Lemma 3. Specifically, for the above example,

$$\begin{aligned} \mathrm{Cov}(X_u(k), X_{u'}(k)) &= \mathbb{P}(X_u(k) = 1, X_{u'}(k) = 1) \\ &\quad - \mathbb{P}(X_u(k) = 1)\mathbb{P}(X_{u'}(k) = 1) \\ &= \frac{3}{20} - \frac{3}{5} \cdot \frac{1}{5} = \frac{3}{100}, \end{aligned} \tag{45}$$

and we have

$$\check{a}(u,u') = \frac{\mathrm{Cov}(X_u(k), X_{u'}(k))}{\max\{p_u, p_{u'}, 1-p_u, 1-p_{u'}\}} = \frac{\frac{3}{100}}{\frac{4}{5}} = 3.75\%.$$

□

Lemma 3 provides a method to convert correlated data to independent traces. The remaining task is to show that we can achieve perfect privacy after applying such a method. As shown in Figure 9, two stages of obfuscation and one stage of anonymization are employed to achieve perfect privacy for users. Note that the first stage of obfuscation is due to Lemma 3 and the second stage (as will be explained in Theorem 5) is the same obfuscation technique given in Theorem 1 of [35]. In Figure 9, $\check{Z}_u(k)$ shows the (reported) data point of user u at time k after applying the first stage of obfuscation with the noise level equal to

$$\check{a}(u,u') = \frac{\mathrm{Cov}(X_u(k), X_{u'}(k))}{\max\{p_u, p_{u'}, 1-p_u, 1-p_{u'}\}}$$

for the dataset of user u and user u', $Z_u(k)$ shows the (reported) data point of user u at time k after applying the second stage of obfuscation with the noise level equal to

$$a_n = c' n^{-\left(\frac{1}{s} - \beta\right)},$$

and $Y_u(k)$ shows the (reported) data point of user u at time k after applying anonymization.

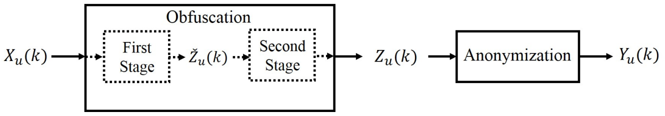


Fig. 9: Applying obfuscation and anonymization techniques to the users' data points.

Consider $G(\mathcal{V}, F)$, where $s_l \leq 2$. We have the same model for p_u as in the previous sections: p_u is chosen from some density $f_P(p_u)$ such that, for $\delta_1, \delta_2 > 0$:

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

Also, if $(u,u') \in F$, $\rho_{uu'}$ is chosen according to some density $f_P(\rho_{uu'}|p_u, p_{u'})$ with range of $\left[0, \min\left\{\sqrt{\frac{p_u(1-p_{u'})}{p_{u'}(1-p_u)}}, \sqrt{\frac{p_{u'}(1-p_u)}{p_u(1-p_{u'})}}\right\}\right]$. The following theorem states that we can indeed achieve perfect privacy if we allow collaboration between users.

Theorem 5. For the two-state model, if $\mathbf{Z}$ is the obfuscated version of $\mathbf{X}$, $\mathbf{Y}$ is the anonymized version of $\mathbf{Z}$, and the size of all subgraphs are less than or equal to 2, there exists an anonymization/obfuscation scheme such that for all $(u,u') \in F$, the asymptotic noise level for users u and u' is at most

$$a(u,u') = \frac{\mathrm{Cov}(X_u(k), X_{u'}(k))}{\max\{p_u, p_{u'}, 1-p_u, 1-p_{u'}\}},$$

to achieve perfect privacy for all users. The anonymization parameter $m = m(n)$ can be made arbitrarily large.

Proof. There are two main steps.

Step 1: De-correlate based on Lemma 3. In particular, note that for at least half of the users, no noise is added in this step. More specifically, define

$$\begin{aligned}\mathcal{U} &= \text{Set of unaffected users}\\ &= \{u : \text{no noise is added to user } u \text{ in this step}\}.\end{aligned}$$

Then after step 1, we have $\check{Z}_u(k) \sim \text{Bernoulli}(\check{q}_u)$. As a result,

- For $u \in \mathcal{U}$; $\check{Z}_u(k) = X_u(k)$ and $\check{q}_u = p_u$.
- For $u \in \{1, 2, \cdots, n\} - \mathcal{U}$; $\check{Z}_u(k) \neq X_u(k)$ and $\check{q}_u \neq p_u$.

Note $|\mathcal{U}| \geq \frac{n}{2}$, because the main graph consists of subgraphs with $s_l \leq 2$.

Step 2: Assume $\check{q}_u$'s are known to the adversary. The setup is now very similar to Theorem 1 in [35], where perfect privacy is proved for the i.i.d. data. But there is a difference here. Specifically, although the users' data $\check{Z}_u(k)$ are now independent, the distribution of $\check{q}_u$'s are not, since they are the result of the data-dependent obfuscation technique of Lemma 3. Luckily, this issue can be easily resolved so that we can show perfect privacy for User 1. The main idea is to use the fact that, as stated above, at least $\frac{n}{2}$ of the users are not impacted by the de-correlation step. As we see below, these users will be sufficient to ensure perfect privacy for User 1 (which may or may not be in the set $\mathcal{U}$).

Let's explore the distributions of $\check{Q}_u = \check{q}_u$ for users in the set $\mathcal{U}$. For any correlated pair of users, the method of Lemma 3 leaves the one whose p_u is farthest from $\frac{1}{2}$ intact. Since p_u's are chosen independently from each other and each user is correlated with only one user, it is easy to see that for users in the set $\mathcal{U}$, the $\check{q}_u$'s are i.i.d. with the following probability density function

$$f_{\check{Q}}(\check{q}_u) = 2 f_P(\check{q}_u) \int_{\min(\check{q}_u, 1-\check{q}_u)}^{\max(\check{q}_u, 1-\check{q}_u)} f_P(x) dx.$$

Therefore, the setup is the same as Theorem 1 in [35] where we want to prove perfect privacy for User 1, and we have $\frac{n}{2}$ users who are independent from User 1 and their parameter $\check{q}_u$ is chosen i.i.d. according to a density function. However, we need to check that the density function $f_{\check{Q}}(\check{q})$ satisfies the condition $\check{\delta}_1 < f_{\check{Q}}(\check{q}_u) < \check{\delta}_2$ for some $\check{\delta}_1$ and $\check{\delta}_2$ on a neighborhood $\check{q}_u \in [p_u - \epsilon', p_u + \epsilon']$. First, note that

$$\begin{aligned}f_{\check{Q}}(\check{q}_u) &= 2 f_P(\check{q}_u) \int_{\min(\check{q}_u, 1-\check{q}_u)}^{\max(\check{q}_u, 1-\check{q}_u)} f_P(x) dx.\\ &< 2\delta_2^2 = \check{\delta}_2.\end{aligned}$$

Next,

$$\begin{aligned}f_{\check{Q}}(\check{q}_u) &= 2 f_P(\check{q}_u) \int_{\min(\check{q}_u, 1-\check{q}_u)}^{\max(\check{q}_u, 1-\check{q}_u)} f_P(x) dx.\\ &> 2\delta_1^2 |1 - 2\check{q}_u| = \check{\delta}_1.\end{aligned}$$

Thus, as long as $p_u \neq \frac{1}{2}$, the condition is satisfied [3]. Therefore, we can show perfect privacy for User 1. Note that here, in the second step, we need to apply a second stage of obfuscation and apply anonymization according to Theorem 1 in [35]. Nevertheless, since the noise level $a_n \to 0$ for this second stage, the asymptotic noise level will stay the same as that for step 1, i.e.

$$a(u, u') = \check{a}(u, u') = \frac{|\text{Cov}(X_u(k), X_{u'}(k))|}{\max\{p_u, p_{u'}, 1 - p_u, 1 - p_{u'}\}}.$$

□

Now, the above method can be readily extended to the case where $s_l > 2$. Consider $s_l = 3$. From Figure 10, there are two different situations in this case:

1) **Case 1:** As shown in Figure 10b, User 1 and user 2 are correlated, and user 2 and user 3 are correlated. In the first step, we de-correlate user 2 and user 3 based on Lemma 3. Now, we face a similar situation as that in the case $s_l = 2$ (as shown in Figure 10a), and we de-correlate them based on Lemma 3. Hence, we can make all of the users independent from each other and then, according to Theorem 5, we can achieve perfect privacy for all of them.
2) **Case 2:** As shown in Figure 10c, all three users are correlated to each other. In the first step, we use Lemma 3 to make User 1 and user 3 uncorrelated. Now, we have a similar situation as case 1, so we can make all the users independent from each other and then, according to Theorem 5, we can achieve perfect privacy for all of them.

Discussion 9: Note that obfuscating data by adding non-zero asymptotic noise may degrade utility significantly. Therefore, in practice, it is usually not possible to de-correlate *all* dependent users without imposing substantial utility degradation. In addition, in order to convert correlated data to independent data, users should collaborate together and disclose their private data to each other, which degrades privacy unless users trust each other. In such a setting, a possible approach in applying our technique is to only add de-correlation noise to the data of highly-dependent users (e.g., spouses and close friends), and leave data of less-dependent users (e.g., co-workers) unchanged.

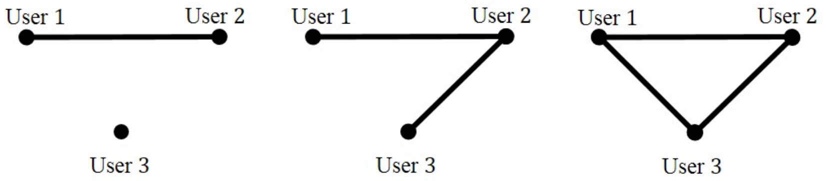
VII. Acknowledgment

The authors would like to thank Dr. Farhad Shirani Chaharsooghi of the New York University for his valuable suggestions and discussions regarding graph alignments.

VIII. Conclusion

Resourceful adversaries can leverage statistical matching based on the prior behavior of users in order to break the privacy provided by PPMs. Our previous work has considered the requirements on anonymization and obfuscation for "perfect" user privacy when traces are independent between users. However, in practice users have correlated data traces, as relationships between users establish dependence in their behavior. In this paper, we demonstrated that such dependency degrades the privacy of PPMs, as the anonymization employed must be significantly increased to preserve perfect privacy, and often no degree of independent obfuscation of the traces

[3] The case $p_u = \frac{1}{2}$ has zero probability, and thus need not be considered. Nevertheless, the result can be shown for $p_u = \frac{1}{2}$, as all we require is a number of users proportional to the length of the interval in the vicinity of p_u.



(a) $s_l = 2$. (b) $s_l = 3$: Case 1. (c) $s_l = 3$: Case 2.

Fig. 10: Three different ways which 3 users can be correlated to each other.

TABLE IV: Conditions on the number of adversary's observations per user (m) for "no privacy" in the case the anonymization is employed as a PPM. Here, s is the size of group of users whose data traces are dependent, r is the number of possible values for each user's data point, $|E|$ is the size of set of edges in the Markov chain, and the results hold for any $\alpha > 0$.

Users' data model	Independent users [34]	Dependent users
	m	m
Two-state i.i.d. model	$\Omega\left(n^{2+\alpha}\right)$	$\Omega\left(n^{\frac{2}{s}+\alpha}\right)$
r-state i.i.d. model	$\Omega\left(n^{\frac{2}{r-1}+\alpha}\right)$	$\Omega\left(n^{\frac{2}{s(r-1)}+\alpha}\right)$
r-state Markov chain model	$\Omega\left(n^{\frac{2}{\lvert E\rvert-r}+\alpha}\right)$	$\Omega\left(n^{\frac{2}{s(\lvert E\rvert-r)}+\alpha}\right)$

TABLE V: Conditions on the number of adversary's observations per user (m) and the amount of noise level (a_n) for "no privacy" in the case both obfuscation and anonymization are combined to be employed as a PPM. Here, s is the size of group of users whose data traces are dependent, r is the number of possible values for each user's data point, $|E|$ is the size of set of edges in the Markov chain, and the results hold for any $\alpha > 0$.

Users' data model	Independent users [35]		Dependent users	
	m	a_n	m	a_n
Two-state i.i.d. model	$\Omega\left(n^{2+\alpha}\right)$	$O\left(n^{-1-\beta}\right)$	$\Omega\left(n^{\frac{2}{s}+\alpha}\right)$	$O\left(n^{-\frac{1}{s}-\beta}\right)$
r-state i.i.d. model	$\Omega\left(n^{\frac{2}{r-1}+\alpha}\right)$	$O\left(n^{-\frac{1}{r-1}-\beta}\right)$	$\Omega\left(n^{\frac{2}{s(r-1)}+\alpha}\right)$	$O\left(n^{-\frac{1}{s(r-1)}-\beta}\right)$
r-state Markov chain model	$\Omega\left(n^{\frac{2}{\lvert E\rvert-r}+\alpha}\right)$	$O\left(n^{-\frac{1}{\lvert E\rvert-r}-\beta}\right)$	$\Omega\left(n^{\frac{2}{s(\lvert E\rvert-r)}+\alpha}\right)$	$O\left(n^{-\frac{1}{s(\lvert E\rvert-r)}-\beta}\right)$

can be effective. The summary of the results is shown in Tables IV and V. We have also presented preliminary results on dependent obfuscation to improve users' privacy.

References

[1] 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, 2018.

[2] 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.

[3] A.-R. Sadeghi, C. Wachsmann, and M. Waidner, "Security and privacy challenges in industrial internet of things," in *52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*. IEEE, 2015, pp. 1–6.

[4] H. Lin and N. W. Bergmann, "Iot privacy and security challenges for smart home environments," *Information*, vol. 7, no. 3, p. 44, 2016.

[5] F. Dalipi and S. Y. Yayilgan, "Security and privacy considerations for IoT application on smart grids: Survey and research challenges," in *IEEE International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*. IEEE, 2016, pp. 63–68.

[6] E. A. Alkeem, C. Y. Yeun, and M. J. Zemerly, "Security and privacy framework for ubiquitous healthcare IoT devices," in *10th International Conference for Internet Technology and Secured Transactions (ICITST)*. IEEE, 2015, pp. 70–75.

[7] A. F. Harris, H. Sundaram, and R. Kravets, "Security and privacy in public IoT spaces," in *25th International Conference on Computer Communication and Networks (ICCCN)*. IEEE, 2016, pp. 1–8.

[8] A. Ukil, S. Bandyopadhyay, and A. Pal, "Privacy for IoT: Involuntary privacy enablement for smart energy systems," in *2015 IEEE International Conference on Communications (ICC)*. IEEE, 2015, pp. 536–541.

[9] M. Vosoughi and S. Köse, "Combined distinguishers to enhance the accuracy and success of side channel analysis," in *IEEE International Symposium on Circuits and Systems*, May 2019, pp. 1–5.

[10] V. Sivaraman, H. H. Gharakheili, A. Vishwanath, R. Boreli, and O. Mehani, "Network-level security and privacy control for smart-home IoT devices," in *IEEE 11th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 2015, pp. 163–167.

[11] F. Kazemi, E. Karimi, A. Heidarzadeh, and A. Sprintson, "Single-server single-message online private information retrieval with side information," *CoRR*, vol. abs/1901.07748, 2019. [Online]. Available: http://arxiv.org/abs/1901.07748

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

[13] 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 Computing*, vol. 3, no. 2, pp. 36–45, 2016.

[14] 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.

[15] 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.

[16] N. Apthorpe, D. Reisman, and N. Feamster, "A Smart Home is No Castle: Privacy Vulnerabilities of Encrypted IoT Traffic," in *Workshop on Data and Algorithmic Transparency*, 2016.

[17] H. Wang and F. du Pin Calmon, "An estimation-theoretic view of privacy," *CoRR*, vol. abs/1710.00447, 2017. [Online]. Available: http://arxiv.org/abs/1710.00447

[18] 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.

[19] R. Al-Dhubhani and J. Cazalas, "Correlation analysis for geo-indistinguishability based continuous lbs queries," in *2nd International Conference on Anti-Cyber Crimes (ICACC)*. Abha, Saudi Arabia: IEEE, 2017.

[20] S. Zhang, Q. Ma, T. Zhu, K. Liu, L. Zhang, W. He, and Y. Liu, "Plp: Protecting location privacy against correlation-analysis attack in crowdsensing," in *44th International Conference on Parallel Processing (ICPP)*. Beijing, China: IEEE, 2015.

[21] H. Liu, X. Li, and H. Li, "Spatiotemporal correlation-aware dummy-based privacy protection scheme for location-based services," in *IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2017.

[22] Y. Xiao and L. Xiong, "Protecting locations with differential privacy under temporal correlations," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1298–1309.

[23] D. Cullina, P. Mittal, and N. Kiyavash, "Fundamental limits of database alignment," *CoRR*, vol. abs/1805.03829, 2018. [Online]. Available: http://arxiv.org/abs/1805.03829

[24] O. E. Dai, D. Cullina, and N. Kiyavash, "Database alignment with gaussian features," *CoRR*.

[25] F. Shirani, S. Garg, and E. Erkip, "A concentration of measure approach to database de-anonymization," *CoRR*, vol. abs/1901.07655, 2019. [Online]. Available: http://arxiv.org/abs/1901.07655

[26] S. Song, Y. Wang, and K. Chaudhuri, "Pufferfish privacy mechanisms for correlated data," in *Proceedings of the 2017 ACM International Conference on Management of Data*. Chicago, Illinois, USA: ACM, 2017, pp. 1291–1306.

[27] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. Athens, Greece: ACM, 2011, pp. 193–204.

[28] T. Zhu, P. Xiong, G. Li, and W. Zhou, "Correlated differential privacy: Hiding information in non-iid data set," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 2, pp. 229–242, 2015.

[29] L. Ou, Z. Qin, Y. Liu, H. Yin, Y. Hu, and H. Chen, "Multi-user location correlation protection with differential privacy," in *IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE], year=2016, address= Wuhan, China.

[30] B. Yang, I. Sato, and H. Nakagawa, "Bayesian differential privacy on correlated data," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. Melbourne, Victoria, Australia: ACM, 2015, pp. 747–762.

[31] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "Privbayes: Private data release via bayesian networks," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pp. 1423–1434.

[32] D. Kifer and A. Machanavajjhala, "Pufferfish: A framework for mathematical privacy definitions," vol. 39, 2014.

[33] C. Liu, P. Mittal, and S. Chakraborty, "Dependence makes you vulnerable: Differential privacy under dependent tuples," in *23nd Annual Network and Distributed System Security Symposium*, San diego, CA, USA, 2016.

[34] 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.

[35] 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.

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

[37] 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.

[38] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Statistical matching in the presence of anonymization and obfuscation: Non-asymptotic results in the discrete case," in *52nd Annual Conference on Information Science and Systems (CISS)*. Princeton,NJ, USA: IEEE, 2018.

[39] N. Takbiri, A. Houmansadr, D. Goeckel, and H. Pishro-Nik, "Asymptotic limits of privacy in bayesian time series matching," in *53rd Annual Conference on Information Science and Systems (CISS)*. Baltimore, MD, USA: IEEE, 2019.

[40] F. Shirani, S. Garg, and E. Erkip, "Seeded graph matching: Efficient algorithms and theoretical guarantees," in *51st Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, 2017, pp. 253–257.

[41] ——, "Typicality matching for pairs of correlated graphs," *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 221–225, 2018.

[42] D. Cullina and N. Kiyavash, "Improved achievability and converse bounds for erdos-renyi graph matching," in *SIGMETRICS*, 2016.

[43] F. Shirani, S. Garg, and E. Erkip, "Matching graphs with community structure: A concentration of measure approach," *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1028–1035, 2018.

[44] D. Cullina and N. Kiyavash, "Exact alignment recovery for correlated erdos renyi graphs," *CoRR*, vol. abs/1711.06783, 2017. [Online]. Available: http://arxiv.org/abs/1711.06783

[45] O. E. Dai, D. Cullina, N. Kiyavash, and M. Grossglauser, "On the performance of a canonical labeling for matching correlated erdős-rényi graphs," *CoRR*, vol. abs/1804.09758, 2018. [Online]. Available: http://arxiv.org/abs/1804.09758

[46] E. Kazemi, "Network alignment: Theory, algorithms, and applications," 2016.

[47] L. Yartseva and M. Grossglauser, "On the performance of percolation graph matching," in *Proceedings of the first ACM conference on Online social networks*. Boston, Massachusetts, USA: ACM, 2013, pp. 119–130.

[48] P. Pedarsani, D. R. Figueiredo, and M. Grossglauser, "A bayesian method for matching two similar graphs without seeds," in *51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. Monticello, IL, USA: ACM, 2013.

[49] S. Ji, W. Li, M. Srivatsa, and R. A. Beyah, "Structural data de-anonymization: Quantification, practice, and implications," in *ACM Conference on Computer and Communications Security*, 2014, pp. 1040–1053.

[50] L. Babai, P. Erdo"s, and S. M. Selkow, "Random graph isomorphism," *SIAM Journal on Computing*, vol. 9, no. 3, pp. 628–635, 1977.

[51] D. G. Corneil and D. G. Kirkpatrick, "A theoretical analysis of various heuristics for the graph isomorphism problem," *SIAM Journal on Computing*, vol. 9, no. 2, pp. 281–297, 1980.

[52] T. Czajka and G. Pandurangan, "Improved random graph isomorphism," *J. Discrete Algorithms*, vol. 6, pp. 85–92, 2008.

[53] B. Bollobás, "Random graphs." Cambridge Studies in Advanced Mathematics, 2001.

[54] 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.

[55] 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*. Raleigh, North Carolina, USA: ACM, 2012, pp. 617–627.

[56] 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.

[57] 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.

[58] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

[59] G. P. Corser, H. Fu, and A. Banihani, "Evaluating location privacy in vehicular communications and applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 9, pp. 2658–2667, 2016.

[60] 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.

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

[62] Z. Ma, F. Kargl, and M. Weber, "A location privacy metric for v2x communication systems," in *IEEE Sarnoff Symposium*. Princeton, NJ, USA: IEEE, 2009, pp. 1–6.

[63] R. Shokri, G. Theodorakopoulos, G. Danezis, J.-P. Hubaux, and J. Y. Le Boudec, "Quantifying location privacy: the case of sporadic location exposure," in *International Symposium on Privacy Enhancing Technologies*. Waterloo, ON, Canada: Springers, 2011, pp. 57–76.

[64] R. Soltani, D. Goeckel, D. Towsley, and A. Houmansadr, "Towards provably invisible network flow fingerprints," in *2017 51st Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, 2017, pp. 258–262.

[65] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Transactions on information theory*, vol. 46, no. 2, pp. 388–404, 2000.

[66] S. Verdú and S. Shamai, "Spectral efficiency of cdma with random spreading," *IEEE Transactions on Information theory*, vol. 45, no. 2, pp. 622–640, 1999.

[67] D. Guo and S. Verdú, "Randomly spread cdma: Asymptotics via statistical physics," *arXiv preprint cs/0503063*, 2005.

[68] K. Li, H. Pishro-Nik, and D. L. Goeckel, "Bayesian time series matching and privacy," in *2017 51st Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2017, pp. 1677–1681.

[69] X. Chen and D. Guo, "Many-access channels: The gaussian case with random user activities," in *2014 IEEE International Symposium on Information Theory*. IEEE, 2014, pp. 3127–3131.

[70] X. Chen, T.-Y. Chen, and D. Guo, "Capacity of gaussian many-access channels," *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 3516–3539, 2017.

[71] D. Guo, Y. Zhu, and M. L. Honig, "Co-channel interference mitigation in multiuser systems with unknown channels," in *Proceedings of XXIXth URSI General Assembly*, 2008.

[72] M. Girvan and M. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99(12), pp. 7821–7826, 2002.

[73] J. Geng, A. Bhattacharya, and D. Pati, "Probabilistic community detection with unknown number of communities," *Journal of the American Statistical Association*, pp. 1–13, 2018.

[74] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics Reports*, vol. 656, pp. 1–44, 2016.

[75] X. Wu, Z. Hu, X. Fu, L. Fu, X. Wang, and S. Lu, "Social network de-anonymization with overlapping communities: Analysis, algorithm and experiments," *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pp. 1151–1159, 2018.

[76] S. Nilizadeh, A. Kapadia, and Y.-Y. Ahn, "Community-enhanced de-anonymization of online social networks," in *ACM Conference on Computer and Communications Security*, 2014.

[77] X. Fu, Z. Hu, Z. Xu, L. Fu, and X. Wang, "De-anonymization of social networks with communities: When quantifications meet algorithms," *CoRR*, vol. abs/1703.09028, 2017.

[78] E. Onaran, S. Garg, and E. Erkip, "Optimal de-anonymization in random graphs with community structure," in *2016 IEEE 37th Sarnoff Symposium*, Newark, NJ, USA, 2016, pp. 1–2.

[79] E. Kazemi, L. Yartseva, and M. Grossglauser, "When can two unlabeled networks be aligned under partial overlap?" in *53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Monticello, IL, USA, 2015.

[80] D. Cullina, K. Singhal, N. Kiyavash, and P. Mittal, "On the simultaneous preservation of privacy and community structure in anonymized networks," *CoRR*, vol. abs/1603.08028, 2016. [Online]. Available: http://arxiv.org/abs/1603.08028

[81] K. Singhal, D. Cullina, and N. Kiyavash, "Significance of side information in the graph matching problem," *CoRR*, vol. abs/1706.06936, 2017. [Online]. Available: http://arxiv.org/abs/1706.06936

[82] B. C. Levy, *Principles of Signal Detection and Parameter Estimation*, 2008.

[83] Y. Wang, X. Wu, and D. Hu, "Using randomized response for differential privacy preserving data collection," in *EDBT/ICDT Workshops*, vol. 1558, 2016.

[84] J. Domingo-Ferrer and J. Soria-Comas, "Connecting randomized response, post-randomization, differential privacy and t-closeness via deniability and permutation," *arXiv preprint arXiv:1803.02139*, 2018.

[85] G. Cormode, T. Kulkarni, and D. Srivastava, "Marginal release under local differential privacy," in *Proceedings of the 2018 International Conference on Management of Data*. ACM, 2018, pp. 131–146.

[86] A. Waseda and R. Nojima, "Analyzing randomized response mechanisms under differential privacy," in *International Conference on Information Security*. Springer, 2016, pp. 271–282.

[87] P. Barbosa, A. Brito, and H. Almeida, "A technique to provide differential privacy for appliance usage in smart metering," *Information Sciences*, vol. 370, pp. 355–367, 2016.

[88] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in *26th {USENIX} Security Symposium ({USENIX} Security 17)*, 2017, pp. 729–745.

[89] Y. Sei and A. Ohsuga, "Differential private data collection and analysis based on randomized multiple dummies for untrusted mobile crowdsensing," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 4, pp. 926–939, 2016.

[90] Z. Huang and W. Du, "Optrr: Optimizing randomized response schemes for privacy-preserving data mining," in *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*. IEEE Computer Society, 2008, pp. 705–714.

[91] Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta, "Amplification by shuffling: From local to central differential privacy via anonymity," in *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2019, pp. 2468–2479.

[92] T. Wang, N. Li, and S. Jha, "Locally differentially private heavy hitter identification," *IEEE Transactions on Dependable and Secure Computing*, 2019.

[93] G. Cormode, T. Kulkarni, and D. Srivastava, "Answering range queries under local differential privacy," *Proceedings of the VLDB Endowment*, vol. 12, no. 10, pp. 1126–1138, 2019.

[94] J. Lee and C. Clifton, "Differential identifiability," in *Proceedings of the 18th International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. Beijing, China: ACM, 2012, pp. 1041–1049.

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

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

[97] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory meets practice on the map," in *4th International Conference on Data Engineering*. Cancun, Mexico: IEEE, 2008, pp. 277–286.

[98] E. Abbe and C. Sandon, "Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery," 2015, pp. 670–688.

[99] E. Abbe, "Community detection and stochastic block models: recent developments," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6446–6531, 2017.

[100] C. Wan, S. Peng, C. Wang, and Y. Yuan, "Communities detection algorithm based on general stochastic block model in mobile social networks," 2016, pp. 178–185.

[101] H. Saad and A. Nosratinia, "Community detection with side information: Exact recovery under the stochastic block model," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, pp. 944–958, 2018.

[102] V. S. Jog and P.-L. Loh, "Recovering communities in weighted stochastic block models," 2015, pp. 1308–1315.

[103] M. Lelarge, L. Massoulié, and J. Xu, "Reconstruction in the labelled stochastic block model," *IEEE Transactions on Network Science and Engineering*, vol. 2, no. 4, pp. 152–163, 2015.

[104] J. Scarlett and V. Cevher, "Partial recovery bounds for the sparse stochastic block model," in *2016 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2016, pp. 1904–1908.

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 and Computer 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.