Asymptotic Privacy Loss Due to Time Series Matching of Dependent Users

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

***Abstract*— The Internet of Things (IoT) promises to improve user utility by tuning applications to a user's current behavior, but the user's behavior can be matched to characteristics learned from prior observations to compromise the user's identity and hence privacy. Our previous work has established the rate at which anonymization must be performed to prevent such matching in a Bayesian setting when faced with a powerful adversary who has extensive knowledge of each user's past behavior. However, even sophisticated adversaries do not often have such extensive knowledge; hence, in this letter, we turn our attention to an adversary who must learn user behavior from past data traces of limited length under the assumptions that: (i) there exists dependency between data traces of different users; and (ii) the data points of each user are drawn from a normal distribution. Results on the lengths of training sequences and rates of anonymization for the data sequences that result in a loss of user privacy are presented.**

***Index Terms*— Anonymization, inter-user dependency, Internet of Things (IoT), privacy-preserving mechanisms (PPM).**

I. Introduction

THE Internet of Things (IoT) allows users to share and access information on a large scale, but the IoT also comes with a significant threat to users' privacy: leakage of sensitive information [1]. There are two main approaches to augment privacy for IoT users: identity perturbation and data perturbation. Identity perturbation (or anonymization) is the removal of the identifying information from a set of data to protect privacy [2], whereas data perturbation (or obfuscation) adds noise to the data [3]. A cost for employing these Privacy-Protection Mechanisms (PPMs) is a blackuction in utility; therefore, optimizing the level of PPMs is of interest.

In [4], a comprehensive analysis of the asymptotic (in the length of the time series) optimal matching of time series to source distributions is presented in a non-Bayesian setting, where the number of users is a fixed, finite value. In contrast, we have adopted a Bayesian setting in [5]–[7], where a powerful adversary is assumed to have accurate prior distributions for user behavior through past observations or other sources. We consider the length of observations available to the adversary that guarantee privacy, or, conversely, the length of observations for which privacy is compromised [5]–[8]. In [7], our most significant results are converse results that demonstrate that this powerful adversary can exploit correlations between the data of different users to compromise user privacy. Thus, a limitation of the converse results of [7] is that they are pblackicated on a very powerful adversary, which, while desirable for (forward) results that guarantee privacy, should be relaxed if possible for (converse) results that demonstrate the loss of privacy. Our main contribution in this letter is to resolve this limitation by developing converse results assuming that the adversary does not have perfect knowledge of the statistics of users' behavior but rather a set of data containing past user behavior.

An initial investigation in [9] obtained the necessary conditions for breaking privacy for a finite number of users. Here we turn our attention to this problem in the most general setting of our prior work with an asymptotically large number of users [5]–[8]. In particular, the data traces of different users will be dependent in many applications, and an adversary can potentially exploit such, thus, contrary to [8], [9], we allow for inter-user correlation as in [7]. Furthermore, we bring our results closer to practice by, rather than presuming the user's data is discrete-valued [5]–[8], considering a Gaussian model for users' data, as Gaussian distributed data has been consideblack in various domains, e.g., sensor networks [10] and distributed consensus [11], as a promising substitute to real data, and it can be adapted to users' check-ins modeled as a multi-center Gaussian model [12]. Dai *et al.* [13] have also investigated the related problem of database alignment for Gaussian features in a different framework.

II. Framework

Define a system with n users, each of which creates a series of m data points. Let $X_u(k)$ be the data point of user u at time k. The vectors $\mathbf{X}_u$ will be termed the "actual data set":

$$\mathbf{X}_u = [X_u(1), X_u(2), \cdots, X_u(m)], \quad u \in \{1, 2, \cdots, n\},$$

where "T" means transpose of the vector.

For every user, there is also a series of l data points representing the user's past behavior; we term these vectors $\mathbf{W}_u$ the "Learning Data Set":

$$\mathbf{W}_u = [W_u(1), W_u(2), \cdots, W_u(l)], \quad u \in \{1, 2, \cdots, n\}.$$

Manuscript received September 24, 2020; revised November 11, 2020; accepted December 7, 2020. Date of publication December 9, 2020; date of current version April 9, 2021. This work was supported by the National Science Foundation under grants CCF–1421957 and CNS–1739462. The associate editor coordinating the review of this letter and approving it for publication was H. Joudeh. *(Nazanin Takbiri and Minting Chen contributed equally to this work.) (Corresponding author: Nazanin Takbiri.)*

Nazanin Takbiri, Minting Chen, Dennis L. Goeckel, and Hossein Pishro-Nik are with the College of Electrical and Computer Engineering, University of Massachusetts Amherst, Amherst, MA 01003 USA (e-mail: ntakbiri@umass.edu; mintingchen@umass.edu; goeckel@ecs.umass.edu; pishro@ecs.umass.edu).

Amir Houmansadr is with the College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA 01002 USA (e-mail: amir@cs.umass.edu).

Digital Object Identifier 10.1109/LCOMM.2020.3043722

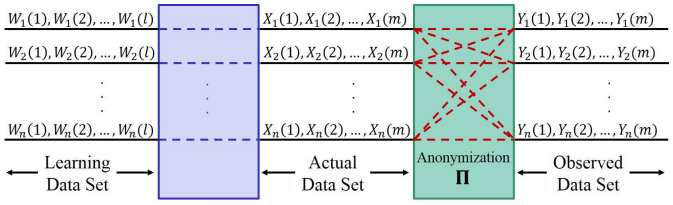

Fig. 1. The goal of the adversary: match each sequence in the learning data set $\mathbf{W}_u$, $u = 1, 2, \ldots, n$, to a sequence in the observed data set $\mathbf{Y}_u$, $u = 1, 2, \ldots, n$.

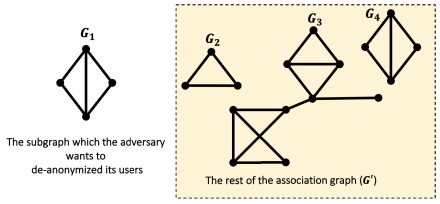

Fig. 2. The structure of the association graph (G): Group g with s_g vertices is disjoint from the remainder of the association graph (G').

For $k \in \{1, 2, \cdots, m\}$ and $k' \in \{1, 2, \cdots, l\}$, $X_u(k)$ and $W_u(k')$ are drawn from a user-specific probability distribution. In particular, we assume that the points in the data sets of a given user user are drawn from a normal distribution $N(\mu_u, \sigma^2)$, where μ_u is the mean of the data of user u and σ^2 is its variance. While the μ_u's are unknown to the adversary, each of them is drawn independently from a continuous density function $f_\mu(x)$. We assume the mild technical condition that there exists $\delta > 0$ such that $f_\mu(x) < \delta$ for all x. Further, the points in the two data sets $X_u(k)$ and $W_u(k')$ are drawn independently from those in the other set, and, within each set, independently across index (k or k'), although there may be inter-user correlation as described below.

Anonymization is employed as a PPM that conceals the mapping between the learning data set and the actual data set by using a random permutation function (Π). The result of permuting $\mathbf{X}_u$ yields the "observed data set":

$$\mathbf{Y}_u = [Y_u(1), Y_u(2), \cdots, Y_u(m)], \quad u \in \{1, 2, \cdots, n\},$$

where each $Y_u(k)$ has a normal distribution $N(\mu_{\Pi^{-1}(u)}, \sigma^2)$; $\mu_{\Pi^{-1}(u)}$ is the mean of the trace in the actual data set that gets mapped to the u^{th} position in the observed data set by the permutation. Thus, we have $\mathbf{Y}_u = \mathbf{X}_{\Pi^{-1}(u)}$ and $\mathbf{Y}_{\Pi(u)} = \mathbf{X}_u$. Figure 1 shows the relation of the three data sets. Note that after anonymization, the adversary does not have access to the actual data set.

1) Association Graph: The dependencies between users are modeled by an association graph $G(\mathcal{V}, E)$, where $\mathcal{V}$ represents the nodes and E represents the edges. As shown in Figure 2, two users are connected if they are dependent:

- $(u, u') \in E$ if and only if $Cov_{uu'} > 0$,
- $(u, u') \notin E$ if and only if $Cov_{uu'} = 0$,

where $Cov_{uu'}$ is the covariance of the data of user u and user u' at any given time.

Discussion 1: Although, disjoint association graph is applicable during social-distancing situations, details of a more general setting in which the association graph is not disjoint have been discussed in [7], [14].

Discussion 2: Note that there are two kinds of dependency, and both of them hurt system privacy in different ways: (i) Intra-user dependency, which shows temporal and spatial dependency within data traces of one user, which is discussed in [6]; (ii) Inter-user dependency, which shows dependency between the traces of different users, and is the main focus of this work.

2) Adversary Model: For each user, the adversary has access to a collection of time series data corresponding to learning data, $\mathbf{W}_u = [W_u(1), W_u(2), \cdots, W_u(l)]$, and a collection of time series data corresponding to observed data $\mathbf{Y}_u = [Y_u(1), Y_u(2), \cdots, Y_u(l)]$. The adversary performs statistical matching between the learning data set $\{\mathbf{W}_u, u = 1, 2, \cdots, n\}$ and the observed data set $\{\mathbf{Y}_u, u = 1, 2, \cdots, n\}$ to match traces in the former, which contains identifying information, with traces in the latter. We assume the adversary knows the structure of the graph $G(\mathcal{V}, E)$ and has knowledge of the anonymization mechanism (i.e. that a random permutation is employed), but not the realization of the random permutation.

We define a user having no privacy as [6]:

Definition 1: User u has *no privacy* at time k if there exists an algorithm for the adversary to estimate $X_u(k)$ perfectly as n goes to infinity. In other words, as $n \to \infty$,

$$\forall k \in \mathbb{N}, \quad \mathbb{P}_e(u) \triangleq \mathbb{P}\left(\widehat{X}_u(k) \neq X_u(k)\right) \to 0,$$

where $\widehat{X}_u(k)$ is the adversary's estimate of $X_u(k)$.

III. Impact of Employing Training Data on Privacy Using Anonymization

The proof of our key result has three main steps; (i) reconstruction of the association graph of the anonymized version of the data, (ii) identifying Group 1, and (iii) identifying all of the members within Group 1 individually.

In the first step, we consider the ability of the adversary to fully reconstruct the structure of the association graph of the anonymized version of the data.

Lemma 1: If for any $\lambda > 0$, the adversary obtains $m = n^\lambda$ points in the observation data set, they can reconstruct $\widetilde{G} = \widetilde{G}(\widetilde{\mathcal{V}}, \widetilde{E})$, where $\widetilde{\mathcal{V}} = \{\Pi(u) : u \in \mathcal{V}\} = \mathcal{V}$, such that with high probability, for all $u, u' \in \mathcal{V}$; $(u, u') \in E$ iff $(\Pi(u), \Pi(u')) \in \widetilde{E}$. We write this statement as $\mathbb{P}(\widetilde{E} = E) \to 1$.

Proof: From the observations, the adversary can calculate the empirical covariance for each pair of user u and user u',

$$\widetilde{\text{Cov}_{uu'}} = \frac{\sum_{k=1}^m X_u(k) X_{u'}(k)}{m} - \frac{\sum_{k=1}^m X_u(k)}{m} \frac{\sum_{k=1}^m X'_u(k)}{m}. \tag{1}$$

In [15, Lemma 1], we have proved for $m = n^\lambda$, and large enough n,

- $|\widetilde{\text{Cov}_{uu'}}| \leq m^{-\frac{1}{5}}$, iff $(u, u') \notin \widetilde{E}$,
- $|\widetilde{\text{Cov}_{uu'}}| > m^{-\frac{1}{5}}$, iff $(u, u') \in \widetilde{E}$,

In other words, we show $P(\widetilde{E}=E)\to 1$ as $n\to\infty$. Thus, according to the result of [15, Lemma 1], the adversary is able to fully reconstruct the structure of the association graph of the anonymized version of the data with arbitrarily small error probability independent of the length of the learning data set. Note that the reconstruction of the association graph does not require the adversary's knowledge about user statistics (i.e., the values of μ_u's) [7, Lemma 1]. □

Without loss of generality, assume that User 1 belongs to Group 1 of size s. In contrast to [7]: (i) the data points are drawn from a Gaussian distribution; and, more importantly, (ii) the adversary does not know the statistics of the users in Group 1, but rather only has the learning data sets for those users. In the next step, we demonstrate how the adversary can identify Group 1 among all of the groups given sufficiently long data traces.

Lemma 2: If for any $\alpha,\alpha'>0$, the adversary obtains learning data sets containing $l=n^{\frac{2}{s}+\alpha'}$ data points of past behavior for each user, and observation data sets containing $m=n^{\frac{2}{s}+\alpha}$ data points for each user, and knows the structure of the association graph, they can identify the traces in the observation data set that correspond to users in Group 1 with arbitrarily small error probability.

Proof: Note that there are at most $\frac{n}{s}$ groups of size s in the system, which we label $1,2,\cdots,\frac{n}{s}$. Define the mean vector for users in Group 1 as: $\mathbf{P}^{(1)}=[\mu_1,\mu_2,\cdots,\mu_s]$, and the vectors of empirical averages for the two sets of data which the adversary seeks to match as:

$$\overline{\mathbf{W}}^{(1)}=\left[\overline{W}_1,\overline{W}_2,\cdots,\overline{W}_s\right],\quad \overline{\mathbf{Y}}^{(1)}=\left[\overline{Y}_1,\overline{Y}_2,\cdots,\overline{Y}_s\right],$$

where $\overline{W}_u=\frac{1}{l}\sum_{i=1}^{l}W_u(i)$ and $\overline{Y}_u=\frac{1}{m}\sum_{i=1}^{m}Y_u(i)$. Let Π_s be the set of all permutations on s users; for $\pi_s\in\Pi_s,\pi_s:\{1,2,\cdots,s\}\to\{1,2,\cdots,s\}$ is a one-to-one mapping. For any two length-s vectors $\mathbf{U}$ and $\mathbf{V}$, we define a difference function that takes into account any permutation of those vectors:

$$D\left(\mathbf{U},\mathbf{V}\right)=\min_{\pi_s\in\Pi_s}\left\{||\mathbf{U}-\mathbf{V}_{\pi_s}||_\infty\right\},$$

where $||\mathbf{U}||_\infty=\max_{i=1,2,\ldots,k}U_i$ for length-k vector $\mathbf{U}$. It is straightforward to show that $D\left(\mathbf{U},\mathbf{V}\right)$ satisfies the triangle inequality, which we will employ below.

Now, defining $\mathbf{P}^{(g)}$, $\overline{\mathbf{W}}^{(g)}$, and $\overline{\mathbf{Y}}^{(g)}$ for groups $g=2,3,\ldots,\frac{n}{s}$ in an analogous way to the definitions of $\mathbf{P}^{(1)}$, $\overline{\mathbf{W}}^{(1)}$, and $\overline{\mathbf{Y}}^{(1)}$, respectively, we claim for $m=n^{\frac{2}{s}+\alpha}$, $l=n^{\frac{2}{s}+\alpha'}$, and as $n\to\infty$:

1) $\mathbb{P}\left(D\left(\overline{\mathbf{W}}^{(1)},\overline{\mathbf{Y}}^{(1)}\right)\le\Delta_n\right)\to 1$,
2) $\mathbb{P}\left(\bigcup_{g=2}^{\frac{n}{s}}D\left(\overline{\mathbf{W}}^{(1)},\overline{\mathbf{Y}}^{(g)}\right)\le\Delta_n\right)\to 0$,

where $\Delta_n=n^{-\frac{1}{s}-\frac{\alpha''}{4}}$, and $\alpha''=\min\{\alpha,\alpha'\}$. For each $u\in\{1,2,\cdots,n\}$,

$$\begin{aligned}&\mathbb{P}\left(\left|\overline{X}_u-\overline{W}_u\right|\ge\Delta_n\right)\\&=\mathbb{P}\left(\left|\left(\overline{X}_u-\mu_u\right)-\left(\overline{W}_u-\mu_u\right)\right|\ge\Delta_n\right)\\&\le\mathbb{P}\left(\left|\overline{X}_u-\mu_u\right|+\left|\overline{W}_u-\mu_u\right|\ge\Delta_n\right)\\&\le\mathbb{P}\left(\left|\overline{X}_u-\mu_u\right|\ge\frac{\Delta_n}{2}\right)+\mathbb{P}\left(\left|\overline{W}_u-\mu_u\right|\ge\frac{\Delta_n}{2}\right)\\&\le e^{\frac{-m\Delta_n^2}{8\sigma^2}}+e^{\frac{-l\Delta_n^2}{8\sigma^2}}\le 2e^{-\frac{n^{\frac{\alpha''}{2}}}{8\sigma^2}}.\end{aligned}\tag{2}$$

The first inequality follows from the triangle inequality. The union bound yields the second inequality, and the third inequality is based on the error function inequality $\text{erf}(x)\ge 1-e^{-x^2}$. By employing (2) and applying the union bound for all of the users in a group with size s, we have for any group g that:

$$\begin{aligned}&\mathbb{P}\left(D\left(\overline{\mathbf{W}}^{(g)},\overline{\mathbf{Y}}^{(g)}\right)\ge\Delta_n\right)\\&\le\sum_{u=1}^{s}\mathbb{P}\left(\left|\overline{X}_u-\overline{W}_u\right|\ge\Delta_n\right)\\&=s\mathbb{P}\left(\left|\overline{X}_u-\overline{W}_u\right|\ge\Delta_n\right)\le 2se^{-\frac{n^{\frac{\alpha''}{2}}}{8\sigma^2}}.\end{aligned}\tag{3}$$

Hence, letting $g=1$, $\mathbb{P}\left(D\left(\overline{\mathbf{W}}^{(1)},\overline{\mathbf{Y}}^{(1)}\right)\le\Delta_n\right)\to 1$, as $n\to\infty$. Next, we want to show that $\mathbb{P}\left(\bigcup_{g=2}^{\frac{n}{s}}D\left(\overline{\mathbf{W}}^{(1)},\overline{\mathbf{Y}}^{(g)}\right)\le\Delta_n\right)\to 0$, as $n\to\infty$. We do this in three steps.

- First, recalling $f_\mu(x)<\delta$ and that the user means are drawn independently, for Group g we obtain:

  $$\mathbb{P}\left(||\mathbf{P}^{(1)}-\mathbf{P}^{(g)}||_\infty\le 4\Delta_n\right)\le(8\Delta_n)^s\delta=8^s n^{-1-\frac{s\alpha''}{4}}\delta.$$

  Similarly, for all $\pi_s\in\Pi_s$, we have

  $$\mathbb{P}\left(||\mathbf{P}^{(1)}-\mathbf{P}^{(g)}{}_{\pi_s}||_\infty\le 4\Delta_n\right)\le 8^s n^{-1-\frac{s\alpha''}{4}}\delta.$$

  Employing union bounds, since $|\Pi_s|=s!$, we have

  $$\begin{aligned}&\mathbb{P}\left(\bigcup_{g=2}^{\frac{n}{s}}\left\{D\left(\mathbf{P}_{\pi_s}^{(g)},\mathbf{P}^{(1)}\right)\le 4\Delta_n\right\}\right)\\&=\mathbb{P}\left(\bigcup_{g=2}^{\frac{n}{s}}\left\{\bigcup_{\pi_s\in\Pi_s}\left\{||\mathbf{P}^{(1)}-\mathbf{P}^{(g)}{}_{\pi_s}||_\infty\le 4\Delta_n\right\}\right\}\right)\\&\le\sum_{g=2}^{\frac{n}{s}}\sum_{\pi_s\in\Pi_s}\mathbb{P}\left(||\mathbf{P}^{(1)}-\mathbf{P}^{(g)}{}_{\pi_s}||_\infty\le 4\Delta_n\right)\\&\le\frac{n}{s}s!8^s n^{-1-\frac{s\alpha''}{4}}\delta=(s-1)!8^s n^{-\frac{s\alpha''}{4}}\delta\to 0,\end{aligned}$$

  as $n\to\infty$. Thus, with high probability, the difference between all $\mathbf{P}^{(g)}$, $g\ge 2$, and $\mathbf{P}^{(1)}$ is bigger than $4\Delta_n$.
- Second, for all $u\in\{2,3,\cdots,n\}$, $\text{erf}(x)\ge 1-e^{-x^2}$ yields

  $$\mathbb{P}\left(\left|\overline{W}_u-\mu_u\right|\ge\Delta_n\right)\le e^{-\frac{l\Delta_n^2}{2\sigma^2}}\le e^{-\frac{n^{\frac{\alpha''}{2}}}{2\sigma^2}}.$$

  Thus, by employing union bounds, we have

  $$\begin{aligned}&\mathbb{P}\left(D\left(\overline{\mathbf{W}}^{(g)},\mathbf{P}^{(g)}\right)\ge\Delta_n\right)\\&\le\mathbb{P}\left(\|\overline{\mathbf{W}}^{(g)}-\mathbf{P}^{(g)}\|_\infty\ge\Delta_n\right)\\&\le\sum_{u\in\text{Group }g}\mathbb{P}\left(\left|\overline{W}_u-\mu_u\right|\ge\Delta_n\right)=se^{-\frac{n^{\frac{\alpha''}{2}}}{2\sigma^2}}.\end{aligned}\tag{4}$$

Now, for $g = 1$, as $n \to \infty$, we have

$$\mathbb{P}\left(D\left(\overline{\mathbf{W}}^{(1)}, \mathbf{P}^{(1)}\right) \geq \Delta_n\right) \leq s e^{-\frac{n^{\frac{\alpha''}{2}}}{2\sigma^2}} \to 0.$$

- Thirdly, since we have shown above that with high probability, $D\left(\mathbf{P}^{(g)}, \mathbf{P}^{(1)}\right) \geq 4\Delta_n$ and $D\left(\overline{\mathbf{W}}^{(g)}, \mathbf{P}^{(g)}\right) \leq \Delta_n$, for all $l \in \{2, 3, \cdots, \frac{n}{s}\}$, by the triangle inequality we have

$$\begin{aligned}
&\mathbb{P}\left(\ D\left(\overline{\mathbf{W}}^{(g)}, \overline{\mathbf{W}}^{(1)}\right) \leq 2\Delta_n\right) \\
&\leq \mathbb{P}\left(\ D\left(\overline{\mathbf{W}}^{(g)}, \mathbf{P}^{(g)}\right) \geq \Delta_n\right) \\
&\leq s e^{-\frac{n^{\frac{\alpha''}{2}}}{2\sigma^2}},
\end{aligned}$$

and by applying a union bound, as $n \to \infty$,

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{g=2}^{\frac{n}{s}} \left\{D\left(\overline{\mathbf{W}}^{(g)}, \overline{\mathbf{W}}^{(1)}\right) \leq 2\Delta_n\right\}\right) \\
&\leq \sum_{g=2}^{\frac{n}{s}} \mathbb{P}\left(\ D\left(\overline{\mathbf{W}}^{(g)}, \overline{\mathbf{Y}}^{(1)}\right) \leq 2\Delta_n\right) \\
&= \frac{n}{s} s e^{-\frac{n^{\frac{\alpha''}{2}}}{2\sigma^2}} = n e^{-\frac{n^{\frac{\alpha''}{2}}}{2\sigma^2}} \to 0.
\end{aligned}$$

- Finally, since we have shown that, with high probability, $D\left(\overline{\mathbf{W}}^{(g)}, \overline{\mathbf{Y}}^{(g)}\right) \leq \Delta_n$ and $D\left(\overline{\mathbf{W}}^{(g)}, \overline{\mathbf{W}}^{(1)}\right) \geq 2\Delta_n$, for all $g \in \{2, 3, \cdots, \frac{n}{s}\}$:

$$\begin{aligned}
&\mathbb{P}\left(\ D\left(\overline{\mathbf{W}}^{(1)}, \overline{\mathbf{Y}}^{(g)}\right) \leq \Delta_n\right) \\
&\leq \mathbb{P}\left(\ D\left(\overline{\mathbf{Y}}^{(g)}, \overline{\mathbf{W}}^{(g)}\right) \geq \Delta_n\right) \\
&\leq 2s e^{-\frac{n^{\frac{\alpha''}{2}}}{8\sigma^2}} \to 0, \qquad (5)
\end{aligned}$$

as $n \to \infty$, where the second inequality follows from (3). Employing (5) and a union bound, as $n \to \infty$ we have

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{g=2}^{\frac{n}{s}} \left\{D\left(\overline{\mathbf{W}}^{(1)}, \overline{\mathbf{Y}}^{(g)}\right) \leq \Delta_n\right\}\right) \\
&\leq \sum_{g=2}^{\frac{n}{s}} \mathbb{P}\left(\ D\left(\overline{\mathbf{Y}}^{(g)}, \overline{\mathbf{W}}^{(g)}\right) \geq \Delta_n\right) \\
&\leq \frac{n}{s} 2s e^{-\frac{n^{\frac{\alpha''}{2}}}{8\sigma^2}} = 2n e^{-\frac{n^{\frac{\alpha''}{2}}}{8\sigma^2}} \to 0.
\end{aligned}$$

Hence, we can conclude that if $m = n^{\frac{2}{s}+\alpha}$, $l = n^{\frac{2}{s}+\alpha'}$, and $n \to \infty$, the adversary can identify the data traces in the observed data set belonging to users in Group 1 with small error probability. □

Finally, we show that once the data traces in the observed data set belonging to users in Group 1 are identified, the adversary can identify the data trace in the observed data set for each of the members of Group 1 with arbitrarily small error probability.

Lemma 3: If for any $\alpha, \alpha' > 0$, the adversary obtains learning data sets containing $l = n^{\frac{2}{s}+\alpha'}$ data points of past behavior for each user, and observation data sets containing $m = n^{\frac{2}{s}+\alpha}$ data points for each user, and knows which traces in the observation data set belong to members of Group 1, the adversary can identify the trace in the observation set belonging to user 1 with arbitrarily small error probability.

Proof: We claim that, for $m = n^{\frac{2}{s}+\alpha}$, $l = n^{\frac{2}{s}+\alpha'}$, and as $n \to \infty$,

1) $\mathbb{P}\left(\ \left|\overline{X}_1 - \overline{W}_1\right| \leq \Delta_n\right) \to 1$,
2) $\mathbb{P}\left(\bigcup_{u=2}^{s} \left|\overline{X}_u - \overline{W}_1\right| \leq \Delta_n\right) \to 0$,

where $\Delta_n = n^{-(\frac{1}{s}+\frac{\alpha''}{4})}$, and $\alpha'' = \min\{\alpha, \alpha'\}$.

1) The first claim follows from (2) with $u = 1$ as $n \to \infty$.
2) Next we establish the second claim. Recall the (mild) technical assumption that $f_\mu(x) < \delta$ for some δ. Then, for all $u \in \{2, 3, \cdots, n\}, \mathbb{P}(|\mu_u - \mu_1| \leq 4\Delta_n) \leq 8\Delta_n\delta$. A union bound yields

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{u=2}^{s} \{\ |\mu_u - \mu_1| \leq 4\Delta_n\}\right) \\
&\leq \sum_{u=2}^{s} \mathbb{P}\left(\ |\mu_u - \mu_1| \leq 4\Delta_n\right) \\
&\leq 8s\Delta_n\delta = 8sn^{-1-\frac{\alpha''}{4}}\delta \to 0,
\end{aligned}$$

as $n \to \infty$. This means that, with high probability, all of the μ_u for $u > 1$ fall outside of the range of $\mu_1 \pm 4\Delta_n$. Next, for all $u \in \{2, 3, \cdots, n\}$, $\text{erf}(x) \geq 1 - e^{-x^2}$ yields

$$\mathbb{P}\left(\ \left|\overline{W}_u - \mu_u\right| \geq \Delta_n\right) \leq e^{-\frac{l\Delta_n^2}{2\sigma^2}}.$$

Thus, for $u = 1$, as $n \to \infty$, we have

$$\mathbb{P}\left(\ \left|\overline{W}_1 - \mu_1\right| \geq \Delta_n\right) \leq e^{-\frac{n^{\frac{\alpha''}{2}}}{2\sigma^2}} \to 0,$$

which means $\overline{W}_1$ is inside $\mu_1 \pm \Delta_n$ with high probability. Thus, we have now shown with high probability that $|\mu_u - \mu_1| \geq 4\Delta_n$ and $|\overline{W}_u - \mu_u| \leq \Delta_n$, for all $u \in \{2, 3, \cdots, n\}$; thus, the triangle inequality yields:

$$\begin{aligned}
\mathbb{P}\left(\ \left|\overline{W}_u - \overline{W}_1\right| \leq 2\Delta_n\right) &\leq \mathbb{P}\left(\ \left|\overline{W}_u - \mu_u\right| \geq \Delta_n\right) \\
&\leq e^{-\frac{l\Delta_n^2}{2\sigma^2}} \leq e^{-\frac{n^{\frac{\alpha''}{2}}}{2\sigma^2}}.
\end{aligned}$$

Applying a union bound, as $n \to \infty$, we have

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{u=2}^{s} \{\ \left|\overline{W}_u - \overline{W}_1\right| \leq 2\Delta_n\}\right) \\
&\leq \sum_{u=2}^{s} \mathbb{P}\left(\ \left|\overline{W}_u - \overline{W}_1\right| \leq 2\Delta_n\right) \\
&= s e^{-\frac{n^{\frac{\alpha''}{2}}}{2\sigma^2}} \to 0.
\end{aligned}$$

Finally, since $|\overline{X}_u - \overline{W}_u| \leq \Delta_n$ and $|\overline{W}_u - \overline{W}_1| \geq 2\Delta_n$, for all $u \in \{2, 3, \cdots, s\}$, with high probability, we can employ (2) to obtain:

$$\begin{aligned}
\mathbb{P}\left(\ \left|\overline{X}_u - \overline{W}_1\right| \leq \Delta_n\right) &= \mathbb{P}\left(\ \left|\overline{X}_u - \overline{W}_u\right| \geq \Delta_n\right) \\
&\leq 2e^{-\frac{n^{\frac{\alpha''}{2}}}{8\sigma^2}}.
\end{aligned}$$

TABLE I

CONDITIONS ON THE LENGTH OF OBSERVED DATASET (m) AND LENGTH OF LEARNING DATASET (l) FOR "NO PRIVACY". HERE, s IS THE SIZE OF GROUP OF USERS WHOSE DATA TRACES ARE DEPENDENT, AND THE RESULTS HOLD FOR ANY $\alpha, \alpha' > 0$

Knowledge of the Adversary	Imperfect		Perfect
	l	m	m
Independent Users	$\Omega\left(n^{2+\alpha'}\right)$	$\Omega\left(n^{2+\alpha}\right)$	$\Omega\left(n^{2+\alpha}\right)$
Dependent Users	$\Omega\left(n^{\frac{2}{s}+\alpha'}\right)$	$\Omega\left(n^{\frac{2}{s}+\alpha}\right)$	$\Omega\left(n^{\frac{2}{s}+\alpha}\right)$

As $n \to \infty$, a union bound yields:

$$\mathbb{P}\left(\bigcup_{u=2}^{s}\left\{\ |\overline{X}_u - \overline{W}_1| \leq \Delta_n\right\}\right)$$
$$\leq \sum_{u=2}^{s}\mathbb{P}\left(\ |\overline{X}_u - \overline{W}_u| \geq \Delta_n\right)$$
$$\leq 2se^{-\frac{n\frac{\alpha''}{2}}{8\sigma^2}} \to 0.$$

□

From Lemmas 1, 2, and 3, a user will have no privacy if m and l are both significantly larger than $n^{\frac{2}{s}}$ as the number of users (n) goes to infinity. Note that m is the number of data points per user in the observation data set, l is the number of data points per user in the the learning data set, and s is the size of the group of the user of interest.

Theorem 1: For the system model with Gaussian data points of Section II, where **Y** is the anonymized version of **X**, and **W** is the behavioral history of users, user 1 has no privacy at time k if:

- The adversary knows the structure of the association graph;
- The adversary has access to a $l-$length behavioral history for each of the users, where $l = \Omega\left(n^{\frac{2}{s}+\alpha'}\right)$ for any $\alpha > 0$;
- The adversary has access to a $m-$length observation for each of the users, where $m = \Omega\left(n^{\frac{2}{s}+\alpha}\right)$ for any $\alpha > 0$;

When the adversary has perfect prior knowledge about users' past behavior in the Gaussian case, which is not coveblack by our prior work, the result follows from arguments similar to those leading to Theorem 1 and [15, Theorem 1].

Theorem 2: For the system model with Gaussian data points of Section II where **Y** is the anonymized version of **X**, user 1 has no privacy at time k if:

- The adversary knows the structure of the association graph;
- The adversary has access to perfect prior knowledge about users' behavior;
- The adversary has access to a $m-$length observations for each of the users, where $m = \Omega\left(n^{\frac{2}{s}+\alpha}\right)$ for any $\alpha' > 0$;

IV. CONCLUSION

Given that anonymization is employed to ensure IoT users' privacy, we consider a broad set of assumptions compablack to previous work: (i) the adversary only has access to limited data sets for users' past behavior; (ii) data traces of different users are dependent; (iii) users' data sequences are governed by an i.i.d. Gaussian model. We established sufficient conditions for an adversary to break user privacy. If the length (l) of the learning data set and the length (m) of the observed data set are each significantly larger than $n^{\frac{2}{s}}$, users have no privacy. The summary of the results is shown in Table I.

REFERENCES

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

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

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

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

[5] Z. Montazeri, A. Houmansadr, and H. Pishro-Nik, "Achieving perfect location privacy in wireless devices using anonymization," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 11, pp. 2683–2698, Nov. 2017.

[6] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Matching anonymized and obfuscated time series to Users' profiles," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 724–741, Feb. 2019.

[7] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Privacy of dependent users against statistical matching," *IEEE Trans. Inf. Theory*, vol. 66, no. 9, pp. 5842–5865, Sep. 2020.

[8] N. Takbiri, D. L. Goeckel, A. Houmansadr, and H. Pishro-Nik, "Asymptotic limits of privacy in Bayesian time series matching," in *Proc. 53rd Annu. Conf. Inf. Sci. Syst. (CISS)*. Baltimore, MD, USA: IEEE, Mar. 2019.

[9] K. Li, H. Pishro-Nik, and D. L. Goeckel, "Bayesian time series matching and privacy," in *Proc. 51st Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Oct. 2017, pp. 1677–1681.

[10] L. Xiao, S. Boyd, and S. Lall, "A scheme for robust distributed sensor fusion based on average consensus," in *Proc. 4th Int. Symp. Inf. Process. Sensor Netw. (IPSN)*, 2005, pp. 63–70.

[11] D. Wagner, "Resilient aggregation in sensor networks," in *Proc. 2nd ACM Workshop Secur. Ad Hoc Sensor Netw. (SASN)*, 2004, pp. 78–87.

[12] C. Cheng, H. Yang, I. King, and M. R. Lyu, "Fused matrix factorization with geographical and social influence in location-based social networks," in *Proc. AAAI*, 2012, pp. 17–23.

[13] O. E. Dai, D. Cullina, and N. Kiyavash, "Achievability of nearly-exact alignment for correlated Gaussian databases," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 1230–1235.

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

[15] N. Takbiri, R. Soltani, D. L. Goeckel, A. Houmansadr, and H. Pishro-Nik, "Asymptotic loss in privacy due to dependency in Gaussian traces," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*. Marrakech, Morocco: IEEE Press, Apr. 2019.