Statistical Matching in the Presence of Anonymization and Obfuscation: Non-Asymptotic Results in the Discrete Case

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

Ke Li
Electrical and Computer Engineering
UMass-Amherst
kli0@engin.umass.edu

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

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

***Abstract*—Many popular applications use traces of user data to tune services to their users but come with a significant risk to user privacy. In particular, even if user traces are anonymized, statistical matching of these traces to prior user behavior can be used to identify the user and their behavior. Because of this threat, there has been significant recent work exploring the theoretical foundations of this problem in the limit of a large number of users and/or observations, where the asymptotic nature of the approaches allows for clean analytical results. In this paper, we turn attention to exact performance analysis for a finite number of users and observations. We consider the case where a user is distributed over a discrete set of states according to a probability distribution drawn at random, which we assume is known to the adversary based on his/her analysis of past user behavior. The finite-length traces are then anonymized and obfuscated at a cost in user utility. We analyze the ability of the adversary to correctly identify user data samples as a function of the rate of anonymization and degree of obfuscation, and we arrive at complicated yet readily numerically evaluated expressions. These results allow us to investigate interesting questions left open by the asymptotic nature of previous work.**

***Index Terms*—Privacy Protecting Mechanism (PPM), Mobile networks, Internet of Things (IoT), Anonymization technique, Obfuscation technique.**

I. Introduction

SMART cities, connected vehicles, smart homes, and connected healthcare devices are examples of how the Internet of Things (IoT) will be revolutionizing our lives in the decades ahead [1]–[4]. IoT devices will be generating an astounding amount of data every second in the near future. This data will inherently contain significant amounts of private information about IoT device users. Due to the importance of privacy, there have been many works on inventing new methods or improving existing methods to protect user privacy [5]–[14]. Even if privacy-preserving mechanisms (PPMs) such as anonymization of user identities and obfuscation of submitted data are employed, significant privacy leaks can occur due to the sheer amount of the data generated and powerful statistical inference techniques available to potential adversaries [15]–[19].

Although, the implementation details might vary, we can generally divide PPMs into two broad categories. The first is identity perturbation or anonymization [11], [20]–[24]. In this technique, user or device IDs are replaced with pseudonyms in order to prevent data leakage. Thus the mapping between the users/devices and their data is hidden to the adversary. The second category is data perturbation or obfuscation [13], [25], [26]. In these techniques, the data generated by IoT devices is perturbed such that the adversary is not able to infer private information from the noisy version of the IoT data. In this research we study both techniques and their combination.

One of the most effective ways to break the privacy of users is to statistically match prior user behavior with the user traces of interest [27]–[30]. Unnikrishnan [24], [31] provides a comprehensive analysis of asymptotically optimal matching of time series to source distributions in the non-Bayesian case. Here we consider a stronger adversary whom has used observations of past user behavior to obtain an accurate statistical model of each user's behavior.

However, utility and privacy are conflicting goals. When employing anonymization, changing the pseudonym of users frequently results in achieving higher privacy but decreases usability and functionality by concealing the temporal relation between a user's data samples. When we employ obfuscation techniques, adding noise to the reported values of user data will decrease the level of utility. As a result, understanding how privacy is preserved while utility is maximized is an important issue. Hence, we seek to obtain the minimize level of anonymization and obfuscation to achieve theoretically guaranteed privacy.

While there has been enormous interest in IoT privacy, there is currently a noticeable gap: a unifying theory of IoT privacy does not exist. The work in [11]–[13], [23] pursued quantitative approaches to privacy, yet all of these works lack a solid theoretical framework. In this paper, we obtain the exact expression for the identification probability for the binary case, where there are two possible states $\{0, 1\}$ for each sample

This work was supported by the National Science Foundation under grants CCF-1421957 and CNS-1739462.

of each user's data, and a finite number of users and data samples, while employing both anonymization and obfuscation techniques.

The work in [32]–[35] is the most related work to that in this paper. In [32]–[35] the concept of perfect privacy is defined and the limits of privacy are characterized. However, these papers limit their consideration to the asymptotic case. Here, we obtain the exact expressions for the discrete and finite case where user data samples are independent and identically distributed (*i.i.d.*) and independent of other users' data sets. Our companion paper [36] studies a similar problem in the case with Gaussian observations. Our results, while exact, are unwieldy. Similar to [36], the expression for the error probability could be used in asymptotic analyses to approach the problem from a different perspective from the information theoretic approach used in [33]. In addition to its potential in asymptotic analyses, we demonstrate here how the results can be used to answer meaningful questions in the application. In particular, [33] indicates that, given enough obfuscation, the length m of the observed traces does not matter. Likewise, given a large enough m, obfuscation is not needed. And, conversely, if both are beneath their thresholds, a user does not have privacy. This gives the idea that the two methods work independently, and never need be employed in unison. Here, our expression for the finite case allow us to investigate whether this is true for smaller (practical) values of the number of users n and sequence length m.

Notation: In this paper $P\left(X = x \middle| Y = y\right)$ is used for the conditional probability of $X = x$ given $Y = y$. When we write $P\left(X \middle| Y\right)$, we are referring to a random variable that is defined as a function of Y.

II. System Model and Metric

Consider a system with n users and denote $X_u(k)$ as a sample of the data of user u at time k, which we desire to protect from the adversary. In the discrete case, there is a countable number of possible states for each sample of each user's data. However, here we consider the binary case, where the data is restricted to $\{0, 1\}$. User u is distinguished by P_u the probability that $X_u(k) = 1$ for any k. Per Section I, we assume the adversary knows P_u, $u = 1, 2, \cdots, n$, based on prior observations of the users, and it is this statistical knowledge that he/she will employ to identity users by the characteristics of their data traces. Finally, P_u, $u = 1, 2, \cdots, n$, are drawn independently from a distribution f_P.

As shown in Figure 1, we employ both anonymization and obfuscation techniques to protect the users' identities. In Figure 1, $Z_u(k)$ is the reported sample of the data of user u at time k, where $Z_u(k)$ has a Bernoulli distribution with the obfuscated probability of being in state 1 denoted as $\tilde{P}_u$. $Y_u(k)$ is the data of user u at time k after applying both obfuscation and anonymization; $Y_u(k)$ has a Bernoulli distribution with the estimated probability of being in state 1 denoted as $\hat{P}_u$.

Obfuscation: The obfuscation is characterized by random variables, R_u, $u = 1, 2, \ldots, n$, which are drawn independently

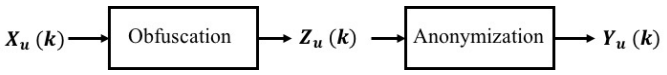

Fig. 1: The sequence $Z_u(k)$, $k = 1, 2, \ldots, m$, is the obfuscated version of $X_u(k)$, $k = 1, 2, \ldots, m$, and the sequence $Y_u(k)$, $k = 1, 2, \ldots, m$, is observed by the adversary after $X_u(k)$, $k = 1, 2, \ldots, m$, is obfuscated and anonymized.

from a distribution f_R. The value of R_u is the probability that a sample of the data of user u is intentionally reported with error. Hence, the effect of the obfuscation is to alter the probability P_u, $u = 1, 2, \ldots, n$ of each user in a way that is unknown to the adversary, since the obfuscation is independent of all past activity of the user. For the binary case, where there are two states (state 0 and state 1) for a user's data pattern, we can write

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

Anonymization: Anonymization is modeled by a random permutation Π such that for user u, the pseudonym of $\Pi(u)$ is assigned. The users' identities are permuted after each m samples, i.e., the observation sequences which the adversary uses to perform statistically matching are of length m. We can write

$$Y_u(k) = Z_{\Pi^{-1}(u)} \text{ and } Z_u(k) = Y_{\Pi(u)}.$$

The adversary attempts to identify the users based on the observations. Per above, we assume a powerful adversary who has complete statistical knowledge of the users' behavior, which means that he/she knows P_u and their distribution f_P, for $u = 1, 2, \ldots, n$. The adversary does not know the instantiation of R_u, $u = 1, 2, \ldots, n$, or the permutation Π for each time period of length m.

The goal of the adversary is to correctly identify the users (i.e., figure out the instantiation of the permutation Π) based on his/her observation of $Y_{\Pi(u)}(k)$, $k = 1, 2, \ldots, m$, $u = 1, 2, \ldots, n$. We illustrate this in Figure 2, where the adversary tries to statistically match each P_u, $u = 1, 2, \ldots, n$, to their corresponding observation sequences $Y_{\Pi(u)}$, $u = 1, 2, \ldots, n$ in order to identify them.

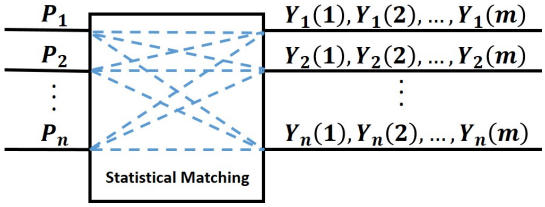

Fig. 2: The goal of the adversary: match each P_u of user u for $u = 1, 2, \ldots, n$ to each observed sequences $Y_{\Pi(u)}(1), Y_{\Pi(u)}(2), \ldots, Y_{\Pi(u)}(m)$ for $u = 1, 2, \ldots, n$.

Our metric is the adversary's probability of being correct, which is the probability that the adversary identifies the data of user u successfully.

Here we assume the distribution f_P and the distribution f_R to be uniform [32]. Note that the problem is still Bayesian because the adversary knows P_u and their distribution f_P, for $u = 1, 2, \cdots, n$.

III. Analytical and Numerical Results

A. Privacy with Anonymization

In this section, we consider the case where only anonymization is employed to provide user privacy. The identification problem can be formulated as a hypothesis testing problem, with the optimal test a straightforward adaptation of the work in [37]. This paper provides an optimal hypothesis test in the case where the adversary has training sequences from the same group of users. Here, the optimal test can be obtained by replacing the empirical number of ones in [37] with the true (ensemble) values of $P_u, u = 1, 2, \cdots, n$. Thus, the optimal test is given by:

Theorem 1. The optimal hypothesis test in the case with binary observations and n users is given by: 1) Order (either descending or ascending) the data sequences by the number of ones they contain, and order $\{P_u, u = 1, 2, \cdots, n\}$; 2) match each data sequence to the P_u (and hence, the user) at the same position in these orders.

Let A_s, $s = 0, 1, \ldots, \lceil \frac{n}{2} \rceil - 1$ be the event that: 1) exactly s of the users have $P_u \leq P_1$ but sum of observation sequence $\sum_{k=1}^{m} Y_{\Pi(u)}(k) \geq \sum_{k=1}^{m} Y_{\Pi(u)}(k)$ (we term this as "user moves from left to right"), and 2) exactly s users have $P_u \geq P_1$ but sum of observation sequence $\sum_{k=1}^{m} Y_{\Pi(u)}(k) \leq \sum_{k=1}^{m} Y_{\Pi(u)}(k)$, (we term this as "user moves from right to left") for $u = 1, 2, \ldots, n$. Given that $A_0, A_1, \cdots, A_{\lceil \frac{n}{2} \rceil - 1}$ are disjoint, the probability P_s that the adversary detects user 1 correctly is given by

$$P_s = P\left(\bigcup_{s=1}^{\lceil \frac{n}{2} \rceil - 1} A_s \right) = \sum_{s=0}^{\lceil \frac{n}{2} \rceil - 1} P(A_s).$$

We denote $\hat{P}_u = \sum_{k=1}^{m} Y_{\Pi(u)}(k)$, $u = 1, 2, \cdots, n$, as the estimation of P_u based on the observed sequence. Thus, in order to obtain $P(A_s|P_1, \hat{P}_1)$, we first consider the probability that a user moves from left to right, which we denote as $P_{L\to R}(P_1 = p_1, \hat{P}_1 = \hat{p}_1)$, and the probability that a user moves from right to left, which we denote as $P_{R\to L}(P_1 = p_1, \hat{P}_1 = \hat{p}_1)$. So we have,

$$\begin{aligned}
&P_{L\to R}\left(P_1 = p_1, \hat{P}_1 = \hat{p}_1\right) = \\
&= E_{P_u}\left[P\left(\left\{ \text{User } u \text{ moves to right} \right\} \middle| \left\{ \text{User } u \text{ starts on left} \right\} \right) \right. \\
&\quad \left. \cdot\, P\left(\left\{ \text{User } u \text{ starts on left} \right\} \middle| P_u, P_1 = p_1, \hat{P}_1 = \hat{p}_1 \right) \right] \\
&= E_{P_u}\left[\sum_{l=\lceil \hat{p}_1 \cdot m \rceil}^{m} \binom{m}{l} P_u^l \left(1 - P_u\right)^{m-l} I_{\{P_u \leq p_1\}} \right] \\
&= \int_0^{p_1} \sum_{l=\lceil \hat{p}_1 \cdot m \rceil}^{m} \binom{m}{l} p_u^l \left(1 - p_u\right)^{m-l} dp_u.
\end{aligned}$$

Likewise,

$$\begin{aligned}
&P_{R\to L}\left(P_1 = p_1, \hat{P}_1 = \hat{p}_1\right) = \\
&= E_{P_u}\left[P\left(\left\{ \text{User } u \text{ moves to left} \right\} \middle| \left\{ \text{User } u \text{ starts on right} \right\} \right) \right. \\
&\quad \left. \cdot\, P\left(\left\{ \text{User } u \text{ starts on right} \right\} \middle| P_u, P_1 = p_1, \hat{P}_1 = \hat{p}_1 \right) \right] \\
&= E_{P_u}\left[\sum_{l=0}^{\lfloor \hat{p}_1 \cdot m \rfloor} \binom{m}{l} P_u^l \left(1 - P_u\right)^{m-1} I_{\{P_u \geq p_1\}} \right] \\
&= \int_{p_1}^{1} \sum_{l=0}^{\lfloor \hat{p}_1 \cdot m \rfloor} \binom{m}{l} p_u^l \left(1 - p_u\right)^{m-1} dp_u.
\end{aligned}$$

Because a user's movement left-to-right or right-to-left is independent of other users when conditioned on P_1 and $\hat{P}_1$, we obtain $P\left(A_s|P_1, \hat{P}_1\right)$ by employing a multinomial distribution with three categories. We denote N_1 as the number of users that move from left to right, N_2 as the number of users that move from right to left, and N_3 as the number of remaining users. Then,

$$\begin{aligned}
&P\left(A_s|P_1, \hat{P}_1\right) = P\left(N_1 = s, N_2 = s, N_3 = n - 2s - 1\right) \\
&= \frac{(n-1)!}{k!k!(n-2k-1)!} P_I(P_1, \hat{P}_1)^s P_{II}(P_1, \hat{P}_1)^s P_{III}(P_1, \hat{P}_1)^{n-2s-1},
\end{aligned}$$

where $P_I = P_{L\to R}$, $P_{II} = P_{R\to L}$, and $P_{III} = 1 - P_{L\to R} - P_{R\to L}$.

Thus, the probability that the adversary successfully identifies user 1 is given by:

$$\begin{aligned}
P_s &= E_{P_1, \hat{P}_1}\left[\sum_{s=0}^{\lceil \frac{n}{2} \rceil - 1} P\left(A_s|P_1, \hat{P}_1\right) \right] \\
&= \int_0^1 \sum_{h=0}^{m} \sum_{s=0}^{\lceil \frac{n}{2} \rceil - 1} P\left(A_s|P_1, \hat{P}_1\right) f_{P_1 \hat{P}_1}(p_1, h) dp_1. \qquad (1)
\end{aligned}$$

where, noting p_1 is uniformly distributed on $[0, 1]$,

$$\begin{aligned}
f_{P_1 \hat{P}_1}(p_1, h) &= f_{\hat{P}_1|P_1}(h|p_1) \cdot f_{P_1}(p_1) \\
&= \binom{m}{h} p_1^h \left(1 - p_1\right)^{m-h}.
\end{aligned}$$

To get some insight into the effect of anonymization on privacy, we show the probability of correct (P_s) in Figure 3, and compare the theoretical results in (1) with simulation results. As expected, the theoretical results match the simulation results. We can also notice that if we decrease the number m of observations per user, or increase the number n of users, the probability of correct decreases. This shows more users and a higher level of anonymization achieve more privacy, as expected.

B. Privacy with Anonymization and Obfuscation

In this section, we employ both obfuscation and anonymization techniques to achieve privacy, and consider how these two techniques combine with each other to affect user privacy.

Recall that the obfuscation is characterized by a random variable R_u, $u = 1, 2, \cdots, n$, which given the probability that

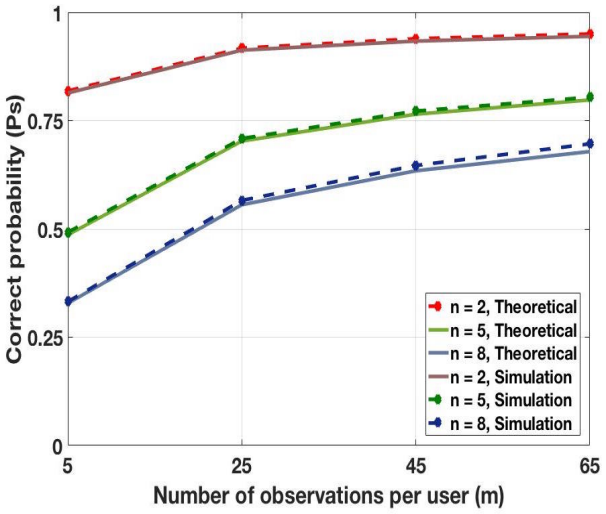

Fig. 3: Comparison of simulation and theoretical results for correct probability (P_s) in identifying a given user when there are 2 users, 5 users, and 8 users in the case that only the anonymization technique is employed.

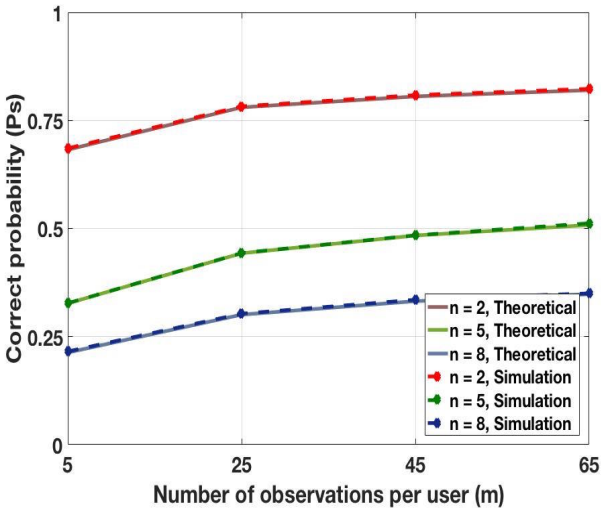

Fig. 4: Comparison of simulation and theoretical results of the correct probability (P_s) in identifying a given user when there are 2 users, 5 users, and 8 users in the case that both obfuscation and anonymization techniques are employed. The noise level is fixed as $a = 0.5$.

any data sample of user u is changed to a different data sample by obfuscation. We assume R_u is distributed uniformly over $[0, a]$, where a is noise level.

Let us define $\tilde{P}_u$ as the probability of $X_u(k) = 1$ after obfuscation; then we have

$$\tilde{P}_u = P_u + R_u(1 - 2P_u)$$

Similar to the previous part, the probability that the adversary correctly identifies the data trace of user 1 is given by:

$$P_s = P\left(\bigcup_{s=1}^{\lceil \frac{n}{2} \rceil - 1} A_s\right) = \sum_{s=0}^{\lceil \frac{n}{2} \rceil - 1} P(A_s).$$

To obtain $P(A_s|P_1, R_1, \hat{P}_1)$, consider the probability a user moves from left to right, which we denote as $P_{L\to R}\left(P_1 = p_1, R_1 = r_1, \hat{P}_1 = \hat{p}_1\right)$ and the probability a user moves from right to left, which we denote as $P_{R\to L}\left(P_1 = p_1, R_1 = r_1, \hat{P}_1 = \hat{p}_1\right)$. Now,

$$\begin{aligned}
&P_{L\to R}\left(P_1 = p_1, R_1 = r_1, \hat{P}_1 = \hat{p}_1\right) = \\
&E_{P_u,R_u}\left[P\left(\left\{\text{User } u \text{ moves to right}\right\}\middle|\left\{\text{User } u \text{ starts on left}\right\}\right)\right. \\
&\left.\cdot\, P\left(\left\{\text{User } u \text{ starts on left}\right\}\middle|P_u, R_u, P_1, R_1, \hat{P}_1\right)\right] \\
&= E_{P_u,R_u}\left[\sum_{l=\lceil \hat{p}_1 \cdot m \rceil}^{m} \binom{m}{l} \tilde{P}_u^l \left(1 - \tilde{P}_u\right)^{m-l} I_{\{P_u \le p_1\}}\right] \\
&= \int_0^{p_1}\int_0^{a} \sum_{l=\lceil \hat{p}_1 \cdot m \rceil}^{m} \binom{m}{l} \tilde{p}_u^l \left(1 - \tilde{p}_u\right)^{m-l} \cdot \left(\frac{1}{a}\right) dr_u dp_u.
\end{aligned}$$

Likewise,

$$\begin{aligned}
&P_{R\to L}\left(P_1 = p_1, R_1 = r_1, \hat{P}_1 = \hat{p}_1\right) = \\
&E_{P_u,R_u}\left[P\left(\left\{\text{User } u \text{ moves to left}\right\}\middle|\left\{\text{User } u \text{ starts on right}\right\}\right)\right. \\
&\left.\cdot\, P\left(\left\{\text{User } u \text{ starts on right}\right\}\middle|P_u, R_u, P_1, R_1, \hat{P}_1\right)\right] \\
&= E_{P_u,R_u}\left[\sum_{l=0}^{\lfloor \hat{p}_1 \cdot m \rfloor} \binom{m}{l} \tilde{P}_u^l \left(1 - \tilde{P}_u\right)^{m-l} I_{\{P_u \ge p_1\}}\right] \\
&= \int_{p_1}^{1}\int_0^{a} \sum_{l=0}^{\lfloor \hat{p}_1 \cdot m \rfloor} \binom{m}{l} \tilde{p}_u^l \left(1 - \tilde{p}_u\right)^{m-l} \cdot \left(\frac{1}{a}\right) dr_u dp_u,
\end{aligned}$$

As a result, for obtaining $P\left(A_s|P_1, R_1, \hat{P}_1\right)$, we write the multinomial distribution as

$$\begin{aligned}
&P\left(A_s|P_1, R_1, \hat{P}_1\right) = P\left(N_1 = s, N_2 = s, N_3 = n - 2s - 1\right) \\
&= \frac{(n-1)!}{s!s!(n-2s-1)!} P_I(P_1, R_1, \hat{P}_1)^s P_{II}(P_1, R_1, \hat{P}_1)^s \\
&\cdot\, P_{III}(P_1, R_1, \hat{P}_1)^{n-2s-1},
\end{aligned}$$

where $P_I = P_{L\to R}$, $P_{II} = P_{R\to L}$, and $P_{III} = 1 - P_{L\to R} - P_{R\to L}$.

Thus, the probability that the adversary successfully detects user 1 is given by

$$P_s = E_{P_1,R_1,\hat{P}_1}\left[\sum_{s=0}^{\lceil \frac{n}{2} \rceil - 1} P(A_s|P_1, R_1, \hat{P}_1)\right].$$

Now we can conclude

$$P_s = \int_0^1 \int_0^a \sum_{h=0}^{m} \sum_{s=0}^{\lceil \frac{n}{2} \rceil - 1} P\left(A_s | P_1, \hat{P}_1\right) f_{P_1 R_1 \hat{P}_1}(p_1, r_1, h) dr_1 dp_1, \tag{2}$$

where $f_{P_1 R_1 \hat{P}_1}(p_1, r_1, h)$ is given by

$$\begin{aligned} f_{P_1 R_1 \hat{P}_1}(p_1, r_1, h) &= f_{\hat{P}_1 | R_1 P_1}(h | r_1, p_1) f_{R_1 | P_1}(r_1 | p_1) f_{P_1}(p_1) \\ &= \binom{m}{h} p_1^h (1 - p_1)^{m-h} \cdot \left(\frac{1}{a}\right). \end{aligned}$$

Again, to get some insight of how anonymization and obfuscation combine to affect privacy, we provide numerical and simulation results in Figures 4 and 5. We compare the theoretical results in (2) with the simulation results, and, as expected, we see that the theoretical results match the simulation results.

In Figure 4, we show the correct probability (P_s) for different numbers of users (n) and length (m) of observation sequences, with a fixed noise level of $a = 0.5$. The figure implies that, similar to the case with only anonymization, if m decreases or n increases, the correct probability (P_s) decreases. In general, if we compare Figure 3 and Figure 4, we see that anonymization among with obfuscation leads to better results in preserving privacy, as expected from our intuition but in contrast to what is suggested by the asymptotic results of [33]. We investigate this further in Figure 6 below.

In Figure 5, we fix $m = 5$ and show P_s for different n and a. We see that a higher level of noise results in a lower correct probability. It shows the degree to which a high level of obfuscation preserves privacy.

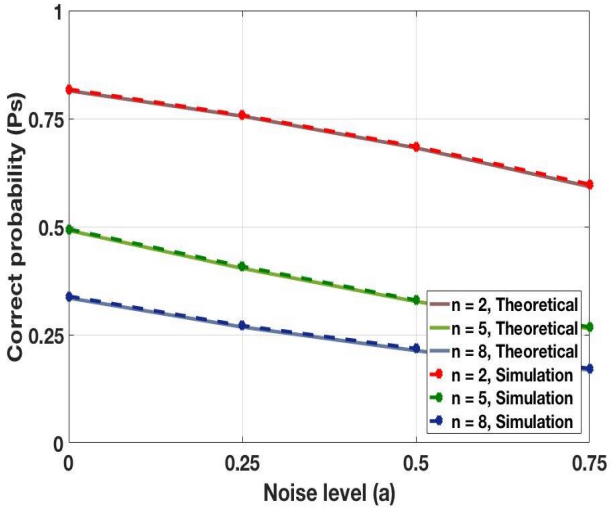

Fig. 5: Comparison of simulation and theoretical results of the correct probability (P_s) in identifying a given user when there are 2 users, 5 users, and 8 users in the case that both obfuscation and anonymization techniques are employed. The length of the observation sequences is fixed as $m = 5$.

Finally, Figure 6 shows for small n and m, anonymization and obfuscation work together for preserving users' privacy. We see that when the anonymization level is not high enough (i.e. m is large) obfuscation helps in protecting user privacy (i.e. P_s decreases when a is large), and when the obfuscation level is not high enough (i.e. a is small), anonymization helps in protecting user privacy (i.e. P_s decreases when m is small). In fact, the sharp corner observed in the asymptotic case, which would suggest the center of the plot in Figure 6, would contain the corner of a box, is not evident. Instead, we see a smooth transition where the techniques can be used in conjunction when neither is sufficient by itself.

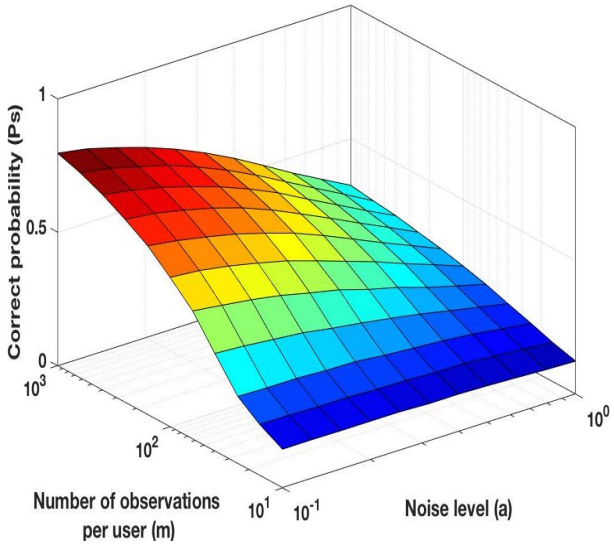

Fig. 6: Simulation results for the correct probability (P_s) in identifying a given user vs. the number of observations per user (m) and noise level (a) for 10 users in the case that both obfuscation and anonymization techniques are employed.

IV. Conclusion

IoT privacy is a major concern in modern society. In this paper, we have explored how anonymization and obfuscation impact user privacy. We consider the discrete case, in particular, when the observation sequences are binary sequences, and we focus on the non-asymptotic case where users' data samples are *i.i.d.* Then we analyzed the ability of a strong adversary, who knows the prior distribution of users' behavior, to correctly identify users' data samples as a function of the rate of anonymization and degree of obfuscation. We obtained the exact expression for two cases: case 1) only the anonymization technique is used to achieve privacy; case 2) both anonymization and obfuscation techniques are used to achieve privacy. We have shown that the level of privacy of the users depends on three factors: Number of users (n), number of observations per user (m), and noise level (a). We also provide numerical and simulation results for the correct probability with different parameter settings to investigate the degree to which privacy is protected for various values of n, m, and a. The results were then used to answer a compelling

question left open in [33]: can the two techniques could be used productively together in the finite case? In contrast to what previous asymptotic results suggest, we find that the two techniques can be used in conjunction to provide privacy when neither is sufficient by itself.

In future research, we will consider the exact expression for the probability of being correct when there are more than two possible states for users' data samples or the case where users' behavior is modeled by Markov chains.

REFERENCES

[1] M. B.-A. Charisma F. Choudhurya and M. Abou-Zeid, "Dynamic latent plan models," *Journal of Choice Modelling*, vol. 3, no. 2, pp. 50–70, 2010.

[2] Z. R. D. A. M. Noble, Shane B. McLaughlin and T. A. Dingus, "Crowdsourced connected-vehicle warning algorithm using naturalistic driving data," Downloaded from http://hdl.handle.net/10919/53978, 2014.

[3] C. F. Choudhury, "Modeling driving decisions with latent plans," Ph.D. dissertation, Massachusetts Institute of Technology, 2007.

[4] J. M. C. S. T. Chrysler and D. C. Marshall, "Cost of warning of unseen threats:unintended consequences of connected vehicle alerts," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2518, pp. 79–85, 2015.

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

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

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

[8] Y. Cai and G. Xu, "Cloaking with footprints to provide location privacy protection in location-based services," Jan. 1 2015, uS Patent App. 14/472,462. [Online]. Available: https://www.google.com/patents/US20150007341

[9] H. Kido, Y. Yanagisawa, and T. Satoh, "An anonymous communication technique using dummies for location-based services," in *Pervasive Services, 2005. ICPS'05. Proceedings. International Conference on*. IEEE, 2005, pp. 88–97.

[10] H. Lu, C. S. Jensen, and M. L. Yiu, "Pad: privacy-area aware, dummy-based location privacy in mobile services," in *Proceedings of the Seventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*. ACM, 2008, pp. 16–23.

[11] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, "Quantifying location privacy," in *Security and Privacy (SP), 2011 IEEE Symposium on*. IEEE, 2011, pp. 247–262.

[12] R. Shokri, G. Theodorakopoulos, G. Danezis, J.-P. Hubaux, and J.-Y. Le Boudec, "Quantifying location privacy: the case of sporadic location exposure," in *Privacy Enhancing Technologies*. Springer, 2011, pp. 57–76.

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

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

[15] F. S. Report, "Internet of things: Privacy and security in a connected world," 2015.

[16] A. Mohsenia and N. K. Jha, "The quest for privacy in the internet of things," *IEEE Cloud Computing*, 2016.

[17] A. Mohsen-Nia and N. K. Jha, "A comprehensive study of security of internet-of-things," *IEEE Transactions on Emerging Topics in Computing*, 2016.

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

[19] S. Hosseinzadeh, S. Rauti, S. Hyrynsalmi, A. Leppänen, Ville Ukil, S. Bandyopadhyay, and A. Pal, "Security in the internet of things through obfuscation and diversification," in *Computing, Communication and Security (ICCCS), IEEE Conference on*. IEEE, 2015, pp. 123–124.

[20] 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.

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

[22] J. Freudiger, M. Raya, M. Félegyházi, P. Papadimitratos, and J.-P. Hubaux, "Mix-zones for location privacy in vehicular networks," in *CM Workshop on Wireless Networking for Intelligent Transportation Systems (WiN-ITS)*, 2007.

[23] Z. Ma, F. Kargl, and M. Weber, "A location privacy metric for v2x communication systems," in *Sarnoff Symposium, 2009. SARNOFF'09. IEEE*. IEEE, 2009, pp. 1–6.

[24] F. 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.

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

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

[27] A.Polak and D. Goeckel, "Identification of wireless devices of users who actively fake their rf fingerprints with artificial data distortion," *IEEE Transactions on Signal Process*, vol. 14, no. 11, pp. 5889–5899, 2015.

[28] M. A.Korba and Y.Ghamri-Doudane, "Anomaly-based intrusion detection system for ad hoc networks," in *IEEE 7th International Conference on the Network of the Future (NOF)*. IEEE, 2016.

[29] L. O.Peled, M.Fire and Y.Elovici, "Entity matching in online social networks," in *2013 International Conference on Social Computing (SocialCom)*. Alexandria, VA, USA: IEEE, 2013.

[30] L. B. Y. De Mulder, G. Danezis and B. Preneel, "Identification via location-profiling in gsm networks," in *Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society*. Alexandria, VA, USA: ACM, 2008.

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

[32] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Matching Anonymized and Obfuscated Time Series to Users' Profiles," *Submitted to IEEE Transaction on Information Theory*, 2017.

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

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

[35] 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.

[36] K. Li, H. Pishro-Nik, and D. L. Goeckel, "Privacy under anonymization and obfuscation with gaussian series," in *Conference on Information Sciences and Systems (CISS)*. Princeton, NJ,USA: IEEE, 2017.

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