Bayesian Time Series Matching and Privacy

Ke Li, Hossein Pishro-Nik, Dennis L. Goeckel
Electrical and Computer Engineering Department, University of Massachusetts, Amherst

***Abstract*—A user's privacy can be compromised by matching the statistical characteristics of an anonymized trace of interest to prior behavior of the user. Here, we address this *matching problem* from first principles in the Bayesian case, where user parameters are drawn from a known distribution, to understand the relationship between the length of the observed traces, the characteristics of the distribution defining the differences between user behavior, and user privacy. First, we establish optimal tests (of two hypotheses and extended to multiple hypotheses as well) for the cases with: 1) continuous alphabets, in particular i.i.d. Gaussian observations with a different (unknown) mean for each user, where the means are drawn from a general a priori distribution; 2) binary alphabets where i.i.d. observations are drawn from a Bernoulli distribution, with each user having an (unknown) probability of being in the "0" state drawn from some certain a priori distribution. Next, for the case with Gaussian observations, we provide general (non-asymptotic) bounds to the performance of the tests and also employ these to show the scaling behavior of privacy. Finally, we present simulation results to demonstrate the accuracy of our analytical bounds.**

I. Introduction

Privacy is a major concern in twenty-first century society. The widespread use of electronic devices provides a trace of user activity that can be matched to prior user behavior and identify the user even when their identity has been anonymized to protect their privacy [1]. Examples include RF fingerprinting [2], intrusion detection [3], matching in online social networks [4] or tracking users by matching the current movement trace to prior traces [5]. In this paper, we consider the fundamental limits of a Bayesian approach to this *matching problem,* which has been considered extensively in the non-Bayesian case in recent work of others [1, 6] and indirectly in the asymptotic case with application to location privacy by our group [7].

Consider a population of K users, each of whose traces obeys an unknown probability distribution, where the probability distribution for a given user is itself drawn independently via a known probability law from a collection of potential probability distributions. For example, we might have characterized the ensemble of statistical patterns of visitors to a large city and have a probability for any given user to adopt a given statistical behavior. We will assume that we observe two traces from each user, one that we will term "training data" and one that we will term the "observation sequence", although it will become clear that the problem is symmetric in the two sets of sequences. From the collections of training data traces and observation sequence traces, our goal is to match each sequence in one collection with sequence in the other collection generated from the same (unknown) probability distribution. In other words, we want to find the permutation of the user identities between the two collections [6].

The work of [6] well motivates the problem but approaches it from a different context. In particular, it focuses on discrete alphabets, asymptotic optimality, and the non-Bayesian case. Here, we take a Bayesian approach, as we have considered previously in the asymptotic case in the context of location privacy [7], but now our goals are to consider the general matching problem and: (1) find optimal tests in the non-asymptotic regime; and, (2) characterize the performance of these tests in terms of the lengths of the sequences, underlying probability law from which user characteristics are drawn, and obfuscation (i.e., noise) on the sequences.

We first consider in detail the case of two users: UserA and UserB for whom we have traces of their past behavior, and our goal is to determine from two observed traces Trace1 and Trace2 whether the proper "matching" is Trace1 with UserA and Trace2 with UserB, or vice versa. Then we extend this matching problem to the case with multiple users. After precisely defining the system model and metrics in Section II, we establish optimal tests for the matching problem on a continuous alphabet, and on a binary alphabet in Section III. In Section IV, we present error analysis for the Gaussian case. Section V provides the numerical results, and Section VI contains the conclusion.

II. System Model and Metrics

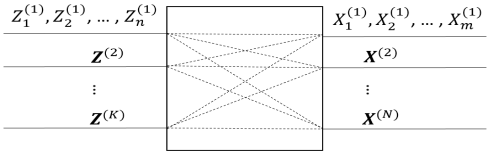


Fig. 1: The matching problem: match each sequence in $[\boldsymbol{X}^{(1)}; \boldsymbol{X}^{(2)}; \cdots; \boldsymbol{X}^{(N)}]$ to that in $[\mathbf{Z}^{(1)}; \mathbf{Z}^{(2)}; \cdots; \mathbf{Z}^{(K)}]$ which obeys the same statistical model.

Consider the matching problem shown in Fig. 1. There are K users in the system. We have observations of prior user behavior, which we call the training sequences, $\mathbf{Z} = [\mathbf{Z}^{(1)}; \mathbf{Z}^{(2)}; \cdots; \mathbf{Z}^{(K)}]$, where $\mathbf{Z}^{(i)}, i = 1, 2, \cdots, K$, is a length-$n$ i.i.d. sequence with each element drawn from probability density function $f_1, f_2, \dots, f_K$, respectively, in the case of a continuous alphabet, and from probability distribution function $p_1, p_2, \dots, p_K$, respectively, for the discrete alphabet case. We adopt a Bayesian framework, which means $f_i, i =$

This work was supported by National Science Foundation under grants CCF– 0844725 and CCF–1421957.

$1, 2, \ldots, K$, and $p_i, i = 1, 2, \cdots, K$, while unknown, are each drawn independently according to a known distribution from the set $\mathcal{F}$ of density functions for the continuous alphabet case, and from the set $\mathcal{P}$ of distribution functions for the case of a discrete alphabet. In our paper, we consider the case when $N = K$, and the sequences in the two collections have a one-to-one matching, reserving for future work the case when $N < K$ or when some of the observation sequences do not have a match. After $\boldsymbol{Z}$ is generated, the user indices are randomly permuted, with each of the $K!$ possible permutations equally likely. Then the sequences $\boldsymbol{X} = [\boldsymbol{X}^{(1)}; \boldsymbol{X}^{(2)}; \cdots; \boldsymbol{X}^{(K)}]$ are observed, where $\boldsymbol{X}^{(i)}$ is a length-m i.i.d. sequence for the i^{th} user after the permutation. Our goal is to match the observation sequences $\boldsymbol{X}$ to the training sequences $\boldsymbol{Z}$ given the knowledge of the distribution on $\mathcal{F}$ (or $\mathcal{P}$), but not knowledge of $f_1, f_2, \ldots, f_K$ (or $p_1, p_2, \ldots, p_K$).

Our matching problem between the training sequences and the observation sequences can be cast as a hypothesis testing problem. For clarity of exposition, consider the 2-user problem, for which there are two hypotheses (the K-user problem will be discussed later), and define:

- H_0: $\boldsymbol{Z}^{(1)}$ has the same distribution as $\boldsymbol{X}^{(1)}$; $\boldsymbol{Z}^{(2)}$ has the same distribution as $\boldsymbol{X}^{(2)}$
- H_1: $\boldsymbol{Z}^{(1)}$ has the same distribution as $\boldsymbol{X}^{(2)}$; $\boldsymbol{Z}^{(2)}$ has the same distribution as $\boldsymbol{X}^{(1)}$.

The goal is to decide between H_0 and H_1. Our metric is the probability of error, which because each of the permutations is equally likely, is given by (for the 2-user case):

$$P_e = 1/2(P(e|H_0) + P(e|H_1)) \tag{1}$$

where $P(e|H_i)$ is the error probability when H_i is true.

We hasten to note that, despite the equally likely assumption on the permutations, the problem is Bayesian due to the knowledge of the probability law by which the user distributions are drawn from $\mathcal{F}$ (or $\mathcal{P}$).

III. HYPOTHESIS TEST

A. Two-Hypothesis Test on a Continuous Alphabet

We consider the 2-user problem in the case of i.i.d. Gaussian observations with a known variance σ^2 for each user and an unknown mean μ_1 for one user and μ_2 for the second user drawn from a known distribution.

Theorem 1. *For the Gaussian scenario given above, with μ_1 and μ_2 drawn from a general distribution, the optimal two-hypothesis test between H_0 and H_1 is given by:*

$$(\sum_{i=1}^{n} Z_i^{(1)} - \sum_{i=1}^{n} Z_i^{(2)})(\sum_{j=1}^{m} X_j^{(1)} - \sum_{j=1}^{m} X_j^{(2)}) \underset{H_1}{\overset{H_0}{\gtrless}} 0 \tag{2}$$

Proof. The derivation is technical, starting with the Likelyhood Ratio Test (LRT):

$$\Omega(\boldsymbol{Z}, \boldsymbol{X}) = \frac{P(\boldsymbol{Z}, \boldsymbol{X}; H_0)}{P(\boldsymbol{Z}, \boldsymbol{X}; H_1)}. \tag{3}$$

Due to space reason, it will be included in a supplemental document. □

Note the simple and intuitive form of the optimal test: the training sequence with the larger sum is matched with the observation sequence with the larger sum.

B. Two-Hypothesis Test on a Binary Alphabet

Now consider the case when the observations are discrete i.i.d. binary sequences. Each element of the sequences follows a Bernoulli distribution. Let the probability of of $Z_i^{(1)} = 0$ be P_1 and the probability of $Z_i^{(2)} = 0$ be P_2. The parameters P_1 and P_2 are randomly drawn from a known distribution.

Theorem 2. *For the binary scenario given above, with P_1 and P_2 drawn from a uniform distribution over $[0, 1]$, the optimal two-hypothesis test between H_0 and H_1 is given by:*

$$(t_1 - t_2)(s_1 - s_2) \underset{H_1}{\overset{H_0}{\gtrless}} 1 \tag{4}$$

where t_1 and t_2 denote the number of 0's in $X^{(1)}$ and $X^{(2)}$, respectively, and s_1 and s_2 denote the number of 0's in $Z^{(1)}$ and $Z^{(2)}$, respectively.

Proof. The proof employs the LRT and will be given in detail in a supplemental document. □

We believe that the optimal detector in (4) is also valid when P_1 and P_2 are drawn from more general distributions. To this point, we have extended it to the case when the distribution can be expressed as a polynomial with positive coefficients.

C. M-ary Hypothesis Test

We can extend the two-hypothesis test to an m-ary hypothesis test with (2) or (4) by employing a (large) collection of binary hypothesis tests and finding the hypothesis which is pairwise optimal versus all others. Assume that we have a K-to-K matching problem and hence $K!$ hypotheses. We can proceed by the following steps:

- Step 1: Pair all hypotheses where one hypothesis can be obtained from the other by permuting two entries in the matching (see example below).
- Step 2: For each pair, drop their common matchings and conduct the test in (2) or (4).
- Step 3: Drop all failed hypotheses and pick the one left as the winner;

For a simple example in the binary scenario: consider a 3-to-3 matching and assume that $s_1 = 1, s_2 = 5, s_3 = 9, t_1 = 2, t_2 = 5, t_3 = 8$. We have six hypotheses in this case:

$$H_0 : s_1 — t_1, s_2 — t_2, s_3 — t_3; \quad H_1 : s_1 — t_1, s_2 — t_3, s_3 — t_2;$$
$$H_2 : s_1 — t_2, s_2 — t_1, s_3 — t_3; \quad H_3 : s_1 — t_2, s_2 — t_3, s_3 — t_1;$$
$$H_4 : s_1 — t_3, s_2 — t_1, s_3 — t_2; \quad H_5 : s_1 — t_3, s_2 — t_2, s_3 — t_1.$$

We should arrive at H_0 as the correct hypothesis (but we do not know this a priori, of course). We first pair (H_0, H_1), (H_0, H_2), (H_0, H_5), (H_1, H_3), (H_1, H_4), (H_2, H_3), (H_2, H_4), (H_3, H_5) and (H_4, H_5). Next, each pair is tested with (4) after dropping their common matchings (e.g., H_0 beats H_1 with t_1 matching s_1 dropped). We will then drop

H_1, H_2, H_3, H_4 and H_5 as they fail a pairwise test. Then, only H_0 is left as the final winner.

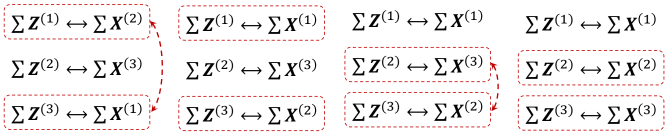
However, the optimal m-ary hypothesis test has a simpler form that can be proven using the pairwise approach above.

Theorem 3. *The optimal m-ary hypothesis test is given by: First, order (either in descending or ascending order) the sums of each observation sequence (i.e., $\sum \boldsymbol{X}^{(i)}$) and the sums of each training sequence (i.e., $\sum \boldsymbol{Z}^{(i)}$) in the Gaussian scenario, or the number of zeros of each observation sequence (i.e., t_i) and the number of zeros of each training sequence (i.e., s_i) in the binary scenario. Then, match the element in the j^{th} position in the first ordered sequence to the element in the j^{th} position in the second ordered sequence, $j = 1, \ldots, K$.*

Proof. We prove it in the Gaussian case; the binary case follows similarly. Algorithm 1 demonstrates that any matching that does not follow that in the Theorem statement results in a hypothesis selection with lower a posterior probability.

Algorithm 1 Proof of Optimal Matching for K Users

```
Input: $S_Z = [\sum \boldsymbol{Z}^{(1)}, \sum \boldsymbol{Z}^{(2)}, \cdots, \sum \boldsymbol{Z}^{(K)}]$ and $S_X = [\sum \boldsymbol{X}^{(1)}, \sum \boldsymbol{X}^{(2)}, \cdots, \sum \boldsymbol{X}^{(K)}]$
Rearrange $S_Z$ in ascendant order and $S_X$ in random order
Denote hypothesis $H_0$ as the matching of the $j^{\text{th}}$ element of $S_x$ with the $j^{\text{th}}$ element of $S_Z$, $j = 1, \ldots, K$
$S \leftarrow S_X$
$i = 1$;
for $i < K$ do
    if $S_{X,i} \neq \min S$ then
        Switch the position of $S_{X,i}$ and $\min S$ in $S_X$
        Denote hypothesis $H_i$ as the matching of the $j^{\text{th}}$ element of $S_x$ with the $j^{\text{th}}$ element of $S_Z$, $j = 1, \ldots, K$
        Pairwise compare $H_{i-1}$ and $H_i$ by test (2) ($H_{i-1}$ and $H_i$ share only two different matchings. But unlike $H_i$, $H_{i-1}$ does not follow the rule that the training sequence with the larger sum is matched to the observation sequence with the larger sum. So, $H_i$ wins and $H_{i-1}$ is dropped.)
        Remove $\min S$ from $S$
    $i++$;
```



$\sum Z^{(1)} \leftrightarrow \sum X^{(2)}$
$\sum Z^{(2)} \leftrightarrow \sum X^{(3)}$
$\sum Z^{(3)} \leftrightarrow \sum X^{(1)}$
$\sum Z^{(1)} \leftrightarrow \sum X^{(1)}$
$\sum Z^{(2)} \leftrightarrow \sum X^{(3)}$
$\sum Z^{(3)} \leftrightarrow \sum X^{(2)}$
$\sum Z^{(1)} \leftrightarrow \sum X^{(1)}$
$\sum Z^{(2)} \leftrightarrow \sum X^{(3)}$
$\sum Z^{(3)} \leftrightarrow \sum X^{(2)}$
$\sum Z^{(1)} \leftrightarrow \sum X^{(1)}$
$\sum Z^{(2)} \leftrightarrow \sum X^{(2)}$
$\sum Z^{(3)} \leftrightarrow \sum X^{(3)}$

(a) Hypothesis H_0 (b) Hypothesis H_1 (c) Hypothesis H_1 (d) Hypothesis H_2

Fig. 2: Example of Algorithm 1 with three users.

We further illustrate the algorithm with an example in Fig.2. For three users, assume that $\sum \boldsymbol{Z}^{(1)} < \sum \boldsymbol{Z}^{(2)} < \sum \boldsymbol{Z}^{(3)}$ and $\sum \boldsymbol{X}^{(1)} < \sum \boldsymbol{X}^{(2)} < \sum \boldsymbol{X}^{(3)}$. Consider the optimality of hypothesis H_0 given in Fig. 2(a), and the operation of Algorithm 1 to determine such. First, we switch the position of $\sum \boldsymbol{X}^{(2)}$ and $\sum \boldsymbol{X}^{(1)}$ to arrive at the hypothesis matching in Fig. 2(b). We know the matching in Fig. 2(b) beats that in Fig. 2(a) based on test (2). Next, we switch the position of $\sum \boldsymbol{X}^{(3)}$ and $\sum \boldsymbol{X}^{(2)}$ to arrive at the matching in Fig. 2(d) which (2) shows is a better matching than that in Fig. 2(c). □

IV. Performance Analysis

In this section, we provide a performance analysis of the optimal tests in the case of Gaussian observations.

A. Error Probability for the Two-Hypothesis Case

1) Mean μ_i drawn from a zero-mean Gaussian distribution: Assume that μ_1 and μ_2 are drawn from a known zero-mean Gaussian distribution, i.e., $\mu_1 \sim N(0, \sigma_0^2)$ and $\mu_2 \sim N(0, \sigma_0^2)$, where σ_0^2 is known. Let $Z = \sum_{i=1}^{n} Z_i^{(1)} - \sum_{i=1}^{n} Z_i^{(2)}$ and $X = \sum_{j=1}^{m} X_j^{(1)} - \sum_{j=1}^{m} X_j^{(2)}$. We know that Z and X are conditionally Gaussian random variables given μ_1, μ_2. When H_0 is true, $Z \sim N(n\Delta_\mu, 2n\sigma^2)$ and $X \sim N(m\Delta_\mu, 2m\sigma^2)$, where $\Delta_\mu = \mu_1 - \mu_2$ and $\Delta_\mu \sim N(0, 2\sigma_0^2)$. We find the error probability of the optimal test conditioned on μ_1 and μ_2, and we take its expectation. The probability of error is given by:

$$\begin{aligned} P_e(\Delta_\mu) &= P(X \le 0)P(Z \ge 0) + P(X \ge 0)P(Z \le 0) \\ &= \left(\frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{m\Delta_\mu}{\sqrt{4m\sigma^2}}\right)\right)\left(\frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{n\Delta_\mu}{\sqrt{4n\sigma^2}}\right)\right) \\ &+ \left(\frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{n\Delta_\mu}{\sqrt{4n\sigma^2}}\right)\right)\left(\frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{m\Delta_\mu}{\sqrt{4m\sigma^2}}\right)\right) \end{aligned} \quad (5)$$

where erf$(\cdot)$ is the standard error function.

Taking the expectation yields:

$$E[P_e(\Delta_\mu)] = \frac{1}{2} - \frac{\int\limits_0^\infty \text{erf}\left(\frac{\sqrt{m}\Delta_\mu}{2\sigma}\right)\text{erf}\left(\frac{\sqrt{n}\Delta_\mu}{2\sigma}\right) e^{-\frac{\Delta_\mu^2}{4\sigma_0^2}}\, d\Delta_\mu}{2\sigma_0\sqrt{\pi}}$$

We consider a lower bound for P_e. Using $\text{erf}(x) \le 1 - \alpha e^{-\beta x^2}$, $x \ge 0$ where $\alpha = \sqrt{\frac{2e}{\pi}\frac{\sqrt{\beta-1}}{\beta}}$ and $\beta > 1$ (which can be picked to minimize the approximation error [8]), we write:

$$\begin{aligned} P_e &\ge \max_{\beta>1}\left\{\frac{1}{2} - \frac{\int\limits_0^\infty \text{erf}\left(\frac{\sqrt{m}\Delta_\mu}{2\sigma}\right)\left(1 - \alpha e^{-\frac{\beta n\Delta_\mu^2}{4\sigma^2}}\right) e^{-\frac{\Delta_\mu^2}{4\sigma_0^2}}\, d\Delta_\mu}{2\sigma_0\sqrt{\pi}}\right\} \\ &= \max_{\beta>1}\left\{\frac{1}{\pi}\tan^{-1}\left(\frac{\sigma}{\sqrt{m}\sigma_0}\right) + \frac{\frac{\alpha}{2} - \frac{\alpha}{\pi}\tan^{-1}\left(\sqrt{\frac{\beta n\sigma_0^2+\sigma^2}{m\sigma_0^2}}\right)}{\sqrt{\frac{\beta n\sigma_0^2}{\sigma^2}+1}}\right\} \end{aligned} \quad (6)$$

If we use $\text{erf}(x) \ge 1 - e^{-x^2}$, $x \ge 0$ in the expectation, we will obtain an upper bound for P_e:

$$P_e \le \frac{1}{\pi}\tan^{-1}\left(\frac{\sigma}{\sqrt{m}\sigma_0}\right) + \frac{\frac{1}{2} - \frac{1}{\pi}\tan^{-1}\left(\sqrt{\frac{n\sigma_0^2+\sigma^2}{m\sigma_0^2}}\right)}{\sqrt{\frac{n\sigma_0^2}{\sigma^2}+1}} \quad (7)$$

To get some insight into the scaling behavior of privacy with regards to various parameter settings, consider the case when $m = n$. In this case (or the case when $m \approx n$), we provide an even tighter upper bound for P_e than that in (7).

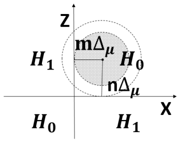


Fig. 3: Region S for the derivation of upper bounds on the error probability P_e for the optimal detector with i.i.d. Gaussian observations and zero-mean Gaussian means (when $m \approx n$).

Given μ_1 and μ_2, we know that $(X - m\Delta_\mu) \sim N(0, 2m\sigma^2)$ and $(Z - n\Delta_\mu) \sim N(0, 2n\sigma^2)$. Let S denote the shaded region in Fig. 3. By the symmetry of the problem (and detector) in the two hypotheses, the probability of error is given by

$$P_e(\Delta_\mu) = 1 - P(H_0|H_0, \Delta_\mu) \leq \quad 1 - P(S|H_0, \Delta_\mu > 0)$$

Suppose $m \leq n$, and replace the smaller variance $2m\sigma^2$ with the larger $2n\sigma^2$, i.e., assume that $(X - m\Delta_\mu) \sim N(0, 2n\sigma^2)$. Then, let $Y = \sqrt{(X - m\Delta_\mu)^2 + (Z - n\Delta_\mu)^2}$, and we know that $Y \sim Rayleigh(\sqrt{2n\sigma^2})$. Since increasing the variance of X and Z makes S less likely given H_0, i.e., $P(S|H_0, \Delta_\mu > 0) \geq P(Y \leq m\Delta_\mu)$, the error probability is then bounded as:

$$P_e(\Delta_\mu) \leq 1 - P(Y \leq m\Delta_\mu) = e^{-\frac{m^2\Delta_\mu^2}{4n\sigma^2}}, \; \Delta_\mu > 0 \qquad (8)$$

Taking its expectation over μ_1 and μ_2 yields:

$$E[P_e(\Delta_\mu)] \leq \frac{1}{\sqrt{4\pi\sigma_0^2}} \int_0^\infty e^{-\frac{m^2\Delta_\mu^2}{4n\sigma^2} - \frac{\Delta_\mu^2}{4\sigma_0^2}}\, d\Delta_\mu = \frac{1}{2\sqrt{\frac{\sigma_0^2 m^2}{\sigma^2 n} + 1}}$$

Generalizing to any m and n, we can bound P_e as:

$$P_e \leq \frac{1}{2\sqrt{\frac{\sigma_0^2 \min(m,n)^2}{\sigma^2 \max(m,n)} + 1}} \qquad (9)$$

If we set $m = n$ in (6) and (9), we have P_e bounded as:

$$\frac{1}{\pi}\tan^{-1}\left(\sqrt{\frac{\sigma^2}{m\sigma_0^2}}\right) + \frac{\frac{\alpha}{2} - \frac{\alpha}{\pi}\tan^{-1}\left(\sqrt{\beta + \frac{\sigma^2}{m\sigma_0^2}}\right)}{\sqrt{\frac{\beta m\sigma_0^2}{\sigma^2} + 1}}$$
$$\leq P_e \leq \frac{1}{2\sqrt{\frac{m\sigma_0^2}{\sigma^2} + 1}} \qquad (10)$$

If m is large, $\frac{\sigma^2}{m\sigma_0^2} < 1$. Also, expand $\tan^{-1}(\cdot)$. Then, we see that a square root law emerges in both the lower and the upper bound: the error probability behaves as $\mathcal{O}(\sqrt{\frac{\sigma^2}{m\sigma_0^2}})$, the square root of the ratio of the variance of the "noise" on the observations to the product of the variance of the difference in the means and the number of samples in the sequences.

2) Mean μ_i drawn from general distribution: Consider the case that Δ_μ follows a distribution f_{Δ_μ} (note that f_{Δ_μ} is even). The distribution $f_{\Delta_\mu^2}$ of Δ_μ^2 can then be obtained. We take the expectation in (5) over f_{Δ_μ} and write the lower bound of P_e:

$$P_e \geq \frac{1}{2} - \frac{1}{2}\int_0^\infty \left(1 - \alpha e^{-\frac{\beta m \Delta_\mu^2}{4\sigma^2}}\right)\left(1 - \alpha e^{-\frac{\beta n \Delta_\mu^2}{4\sigma^2}}\right) f_{\Delta_\mu}\, d\Delta_\mu$$

Then, replacing f_{Δ_μ} with $f_{\Delta_\mu^2}$, we have:

$$P_e \geq \frac{1}{4} + \frac{\alpha}{4} M_{\Delta_\mu^2}\left(-\frac{\beta n}{4\sigma^2}\right) + \frac{\alpha}{4} M_{\Delta_\mu^2}\left(-\frac{\beta m}{4\sigma^2}\right)$$
$$- \frac{\alpha^2}{4} M_{\Delta_\mu^2}\left(-\frac{\beta(m+n)}{4\sigma^2}\right) \qquad (11)$$

where $M_{\Delta_\mu^2}(\cdot)$ is the moment generating function of Δ_μ^2.

Similarly, using (5) we can derive the upper bound for P_e:

$$P_e \leq \frac{1}{4} + \frac{1}{4} M_{\Delta_\mu^2}\left(-\frac{n}{4\sigma^2}\right) + \frac{1}{4} M_{\Delta_\mu^2}\left(-\frac{m}{4\sigma^2}\right)$$
$$- \frac{1}{4} M_{\Delta_\mu^2}\left(-\frac{m+n}{4\sigma^2}\right) \qquad (12)$$

When m is close to n, we use (8) with general m and n and provide a tighter upper bound for P_e:

$$P_e \leq \int_0^\infty e^{-\frac{\min(m,n)^2\Delta_\mu^2}{4\max(m,n)\sigma^2}} f_{\Delta_\mu} d\Delta_\mu = \frac{1}{2} M_{\Delta_\mu^2}\left(-\frac{\min(m,n)^2}{4\max(m,n)\sigma^2}\right)$$

B. Error Probability for the M-ary Hypothesis Case

Now we consider the error probability of the m-ary hypothesis test with Gaussian priors. Assume that there are K observation sequences, and hence, C_2^k pairwise tests. We denote the event that the i^{th} test is correct as A_i. Then, the probability of a correct decision is $P(\cap_{i=1}^{C_2^K} A_i)$ since we need all tests to be correct. The error probability P_e for the m-ary hypothesis test is $P(\cup_{i=1}^{C_2^K} \overline{A_i}) \leq \sum_{i=1}^{C_2^K} P(\overline{A_i})$ based on the union bound. Then, taking the expectation we have $P_e \leq \sum_{i=1}^{C_2^K} E[P(\overline{A_i})]$. Using (7), we then write the upper bound of P_e as:

$$P_e \leq C_2^K \left(\frac{1}{\pi}\tan^{-1}\left(\frac{\sigma}{\sqrt{m}\sigma_0}\right) + \frac{\frac{1}{2} - \frac{1}{\pi}\tan^{-1}\left(\sqrt{\frac{n\sigma_0^2 + \sigma^2}{m\sigma_0^2}}\right)}{\sqrt{\frac{n\sigma_0^2}{\sigma^2} + 1}}\right) \qquad (13)$$

When m is close to n, using (9) we write:

$$P_e \leq \frac{C_2^K}{2\sqrt{\frac{\sigma_0^2 \min(m,n)^2}{\sigma^2 \max(m,n)} + 1}} \qquad (14)$$

V. Numerical Results

In this section, we present numerical results (with 10^6 iterations) for the error probability of the optimal tests with Gaussian observations and means drawn from a Gaussian prior for various users. Our goal is to: (1) consider the tightness of the bounds derived in Section IV; (2) show how the length of the sequences and other parameters impact the performance of the tests.

Fig. 4 shows the case of two users when the length of the training and the observation sequences are the same (i.e., $n = m$) and when $n = 2m$. We compare the error probability P_e with the lower and upper bound, in (10) for $m = n$, and in (6) and (7) for $n = 2m$, with various σ and σ_0. Fig. 5 illustrates

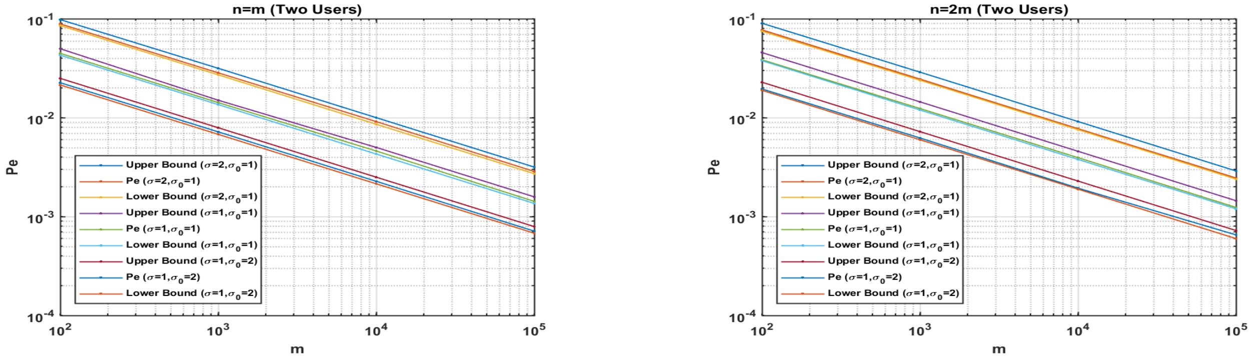


Fig. 4: Error probability with Gaussian observations and means drawn from a Gaussian prior for two users when the length of the training and the observation sequences are the same (i.e., $m = n$) and when $n = 2m$ for variance σ^2 of the observations and variance σ_0^2 of the means of the observations. Parameter $\beta = 1.5$ for the lower bounds.

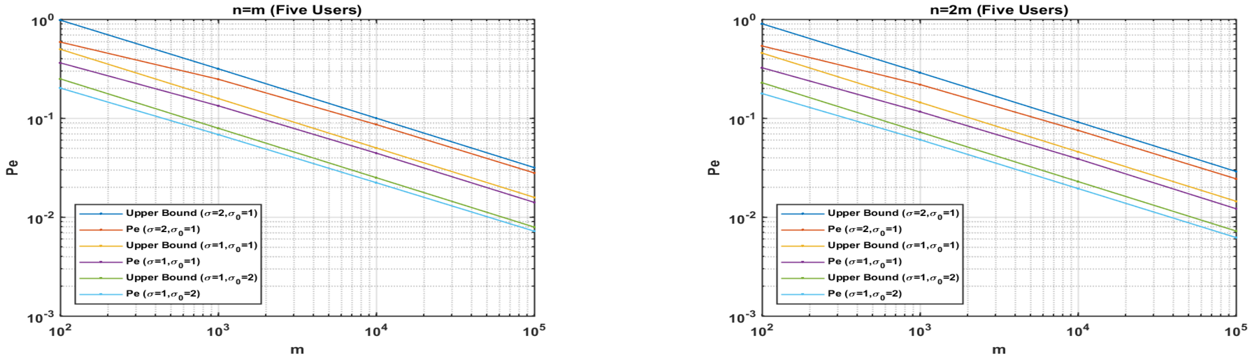


Fig. 5: Error probability with Gaussian observations and means drawn from a Gaussian prior for five users when the length of the training and the observation sequences are the same (i.e., $m = n$) and when $n = 2m$ for variance σ^2 of the observations and variance σ_0^2 of the means of the observations.

the case of five users. We compare P_e with the upper bound in (13) when $m = n$ and in (14) when $n = 2m$. In both figures, the bounds are accurate and capture the scaling behavior.

VI. Conclusion

Statistical matching of a trace of user behavior to prior traces can compromise a user's privacy. In this paper, we have considered this *matching problem* to find optimal tests and understand how privacy depends on the length of the observed sequences and the parameter variation between various users. For discrete alphabets, we have derived the optimal detector when a user's trace is an i.i.d. binary sequence, with the probability of "0" drawn from a distribution that can be expressed as a polynomial with positive coefficients. For continuous alphabets, we have derived the optimal detector and analyzed its performance when a user's trace is an i.i.d. Gaussian sequence with a mean drawn from a general distribution. When the mean is drawn from a Gaussian distribution with variance σ_0^2, upper and lower bounds to the performance suggest how user privacy scales with the length m of the observed traces, variance σ^2 of the observations, and the parameter σ_0^2. We notice a square root law: the upper and lower bound to the error probability behave as $\mathcal{O}(\sqrt{\frac{\sigma^2}{m\sigma_0^2}})$.

References

[1] 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. 358372, Feb. 2016

[2] A. Polak and D. L. Goeckel, "Identification of Wireless Devices of Users Who Actively Fake Their RF Fingerprints With Artificial Data Distortion," *IEEE Trans. Signal Process*, vol. 14, no. 11, pp. 5889-5899, 2015.

[3] A. Korba, M. Nafaa and Y. Ghamri-Doudane, "Anomaly-Based Intrusion Detection System for Ad hoc Networks," *IEEE 7th Int. Conf. Netw. Future*, 2016.

[4] O. Peled, M. Fire, L. Rokach, and Y. Elovici, "Entity Matching in Online Social Networks," *IEEE Social Computing*, pp. 339-344. 2013.

[5] Y. De Mulder, G. Danezis, L. Batina and B. Preneel, "Identification via Location-Profiling in GSM Networks," *Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society*, 2008.

[6] J. Unnikrishnan, "Asymptotically Optimal Matching of Multiple Sequences to Source Distributions and Training Sequences," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 452-468, 2015.

[7] N. Takbiri, A. Houmansadr, D. L. Goeckel and H. Pishro-Nik, "Limits of Location Privacy Under Anonymization and Obfuscation," *2017 IEEE Int. Symp. Inf. Theory*, to appear

[8] S. H. Chang, P. C. Cosman and L. B. Milstein, "Chernoff-type bounds for the Gaussian error function," *IEEE Trans. Commun.*, vol.59, pp.2939-2944, Nov. 2011.