De-anonymizing Private Data by Matching Statistics

Jayakrishnan Unnikrishnan and Farid Movahedi Naini
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
Email: {jay.unnikrishnan, farid.movahedinaini}@epfl.ch

***Abstract*— Recent research has illustrated privacy breaches that can be effected on an anonymized dataset by an attacker who has access to auxiliary information about the users. Most of these attack strategies rely on the uniqueness of specific aspects of the users' data - e.g., observing a mobile user at just a few points on the time-location space are sufficient to uniquely identify him/her from an anonymized set of users. In this work, we consider de-anonymization attacks on anonymized summary statistics in the form of histograms. Such summary statistics are useful for many applications that do not need knowledge about exact user behavior. We consider an attacker who has access to an anonymized set of histograms of K users' data and an independent set of data belonging to the same users. Modeling the users' data as i.i.d., we study the composite hypothesis testing problem of identifying the correct matching between the anonymized histograms from the first set and the user data from the second. We propose a Generalized Likelihood Ratio Test as a solution to this problem and show that the solution can be identified using a minimum weight matching algorithm on an $K \times K$ complete bipartite weighted graph. We show that a variant of this solution is asymptotically optimal as the data lengths are increased. We apply the algorithm on mobility traces of over 1000 users on EPFL campus collected during two weeks and show that up to 70% of the users can be correctly matched. These results show that anonymized summary statistics of mobility traces themselves contain a significant amount of information that can be used to uniquely identify users by an attacker who has access to auxiliary information about the statistics.**

I. Introduction

In recent years, many datasets containing information about individuals have been released into public domain in order to provide open access to statistics or to facilitate data mining research. Often these databases are *anonymized* by suppressing identifiers that reveal the identities of the users, like names or social security numbers. Nevertheless, recent research has revealed that the privacy offered by such anonymized databases may be compromised, if an adversary correlates the revealed information with publicly available databases. For instance, in [1] it was shown that anonymous movie ratings released during the Netflix Prize context could be de-anonymized using public user reviews from the Internet Movie Database (IMDB), and more recently, in [2] it was shown that users can be uniquely identified from a database of mobility traces collected at coarse spatio-temporal resolutions. In most works of this kind, the vulnerability to privacy breaches often arises due to the sparsity of the temporal evolution of the user's data. For instance, the fact that a user watched a movie during a particular time-period or the fact that a user was at a specific location during a particular time can be used to easily identify the user's data from the anonymized dataset.

A potential approach to counter such attacks is to reveal only statistics of the data belonging to each user in the anonymized database. For instance, in the case of mobility traces of users, the summary statistics could be the average time spent by each user in the different locations during a day (or during a different time duration). Similarly, for web-browsing histories, the summary statistics would be the average time spent by each user on different websites. Such summary statistics are sufficient for some applications such as estimating 'popularity' of different locations or websites. In this work, we study de-anonymization attacks on such summary statistics, by an adversary who has access to independent auxiliary information about the users, in the form of datasets or statistics.

Since temporal information is not available, we adopt an i.i.d. model for the temporal evolution of each user's data. We assume that the empirical frequency (or histogram) of the data of each user is released in an anonymized fashion. We consider an adversary who has access to a non-anonymized version of the data of the users collected in an independent experiment. We then study the problem of matching the sets of anonymized and non-anonymized data under the i.i.d. model. This problem is closely related to a classification problem studied by Gutman in [3]. In Gutman's problem, labeled training strings are available from K i.i.d. sources having unknown underlying probability distributions, and the objective is to use this information to classify an unlabeled independently drawn test string to the correct source. The current problem is very similar, except that we now have K unlabeled test strings, one from each source, and the objective is to match all K of them to the correct training string. The current problem and Gutman's problem can be considered to be extreme cases of a more general problem in which some $L \leq K$ unlabeled test strings are available in addition to the K labeled training strings. In this paper we stick to the case where $L = K$. We show that an asymptotically optimal procedure for correctly matching the K sources can be derived by following steps similar to that in [3]. The solution is given by a minimum weight matching problem on a bipartite graph and hence can be efficiently implemented.

The privacy literature contains various approaches for using auxiliary information to de-anonymize datasets. For example in [1] the Netflix dataset was de-anonymized using

user reviews from IMDB and in [4] medical records were de-anonymized with the help of external auxiliary information, namely, ZIP code, birth date, and gender. The de-anonymization of mobility traces was investigated in the works of [5]–[10]. These techniques take into account the temporal information available in the traces. For example, in [8], [9], a Markov model is constructed based on the mobility behaviors of the users, and then *similarity measures* based on heuristics were used for de-anonymization. In [11] the authors build a *contact graph* of the users using the spatial and temporal information available in the traces, and then de-anonymize the users by correlating this graph with a social network. Our work differs from these related works in the fact that we assume that only anonymized statistics, e.g., anonymized histograms, of the users's data are available. For instance, in the case of location data, we assume that we know only the average time spent by the users in various locations, i.e., the histograms, and not the exact temporal information, as required by most existing methods. In addition, we assume that the information available in the dataset to be de-anonymized is *independent* from the auxiliary information. Our notion of independence will be clear in the next section where we present the problem statement. One example of independent information is the case where the dataset and the auxiliary information comprise of users' mobility traces that belong to two non-overlapping time periods. We remark that in the related works of [2], [6], the auxiliary information is a subset of the information in the dataset to be de-anonymized and hence the two are not independent. For example, the auxiliary information is some portions of anonymized users' trajectories where the identities of the users are known.

In this work, we formalize the notion of optimal de-anonymization strategies for such data and identify the correct similarity metric between independent instances of the users' data that leads to an asymptotically optimal solution to the de-anonymization task. We apply our solution to Wi-Fi traces obtained from a university campus and demonstrate that using only temporal statistics of users' mobility, we can de-anonymize more than half of the users in a dataset containing more than a thousand users. The rest of the paper is organized as follows. After introducing our notation, we state the problem in mathematical form in section II. We propose our solution and its optimality properties in Section III, and experimentally evaluate it in Section IV. We conclude in Section V.

Notation: For a finite alphabet Z, we use $\mathcal{P}(\mathsf{Z})$ to denote the set of all probability distributions defined on Z. For any string $s \in \mathsf{Z}^n$, we use $\Gamma_s \in \mathcal{P}(\mathsf{Z})$ to denote the empirical distribution of the string defined as

$$\Gamma_s(z) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}\{s_i = z\}, z \in \mathsf{Z}.$$

Further we use T_s to denote the *type class* of s, i.e., the set of all strings of length n with the same empirical distribution as s. Throughout the paper we use $\log$ to refer to logarithm to the base 2.

II. Problem Statement

Consider a set $\mathcal{K}$ of K sources each producing i.i.d. data according to distinct but unknown distributions on Z. Consider a set $S_1 = \{x_1, x_2, \ldots, x_K\}$ of unlabeled strings of length n each generated by a distinct source in $\mathcal{K}$, and an independent set $S_2 = \{y_1, y_2, \ldots, y_L\}$ of labeled strings of length n each generated by a distinct source within a subset of $\mathcal{K}$ of size L. Here S_1 represents the user data whose unlabeled (i.e., anonymized) statistics are released in public and S_2 represents auxiliary information about the users that is obtained by an adversary. The information S_2 is assumed to be independent of S_1. In the case of data such as mobility patterns or web-browsing history, the information in S_2 could be collected, for instance, by tracking the users. Alternatively, it may be the case that the adversary is some network service provider (e.g., location based service provider or internet service provider) who has access to the user's locations or web-browsing history which contain S_2. Let k denote the source that generated string $x_{\pi(k)} \in \mathsf{Z}^n$ and $y_k \in \mathsf{Z}^n$ where $\pi : \{1, 2, \ldots, L\} \mapsto \{1, 2, \ldots, K\}$ is some unknown injective (one-to-one) function. When $L = K$, the function π is just some unknown permutation. Let p_k denote a probability measure on Z that captures the probability law followed by data from source k. The de-anonymization problem that the adversary needs to solve, is to match each string from set S_2 to the string from S_1 produced by the corresponding source. Equivalently, the adversary seeks to estimate π. The special case of this estimation problem when $L = 1$ was studied by Gutman [3]. In the present paper, we study the other extreme case of $L = K$. Since the observations from each source are assumed to be i.i.d., we will show later (see Lemma 3.2) that the optimal testing procedure requires only the *types*, or empirical distributions, of the strings $\{x_i\}$ and $\{y_j\}$. Thus only the types of the strings are used while performing the matching, as required in the de-anonymization problem.

We view this as a hypothesis testing problem with $M = K!$ composite hypotheses. Each hypothesis corresponds to a unique permutation of $\{1, 2, \ldots, K\}$. Let $\pi_1, \pi_2, \ldots, \pi_M$ denote the M possible permutations of $\{1, 2, \ldots, K\}$. The hypothesis H_i corresponds to a particular permutation π_i. The hypotheses are all composite because the probability distributions of each user's data could lie anywhere in $\mathcal{P}(\mathsf{Z})$. We seek a decision rule for this problem that admits exponential decay of error probability as a function of n under each hypothesis. For this purpose, we allow a no-match decision, i.e., rejection of all M hypotheses. Following an approach similar to that in [3] we denote a decision rule for the M-hypotheses problem by a partition $\Omega = (\Omega_1, \Omega_2, \ldots, \Omega_M, \Omega_R)$ of $\mathbf{Z} = (\mathsf{Z}^n)^K \times (\mathsf{Z}^n)^K$ the space of vectors of the form $x_1, x_2, \ldots, x_K, y_1, y_2, \ldots, y_K$, into $(M+1)$ disjoint cells $\Omega_1, \Omega_2, \ldots, \Omega_M, \Omega_R$, where Ω_i is the acceptance region for hypothesis H_i, and $\Omega_R = \mathbf{Z} - \cup_{i=1}^{M} \Omega_i$ is the rejection zone. We consider an error event e under hypothesis H_i to denote a decision in favor of a wrong hypothesis H_j where $j \neq i$. Note that a decision in favor

of rejection does not correspond to an error event under any hypothesis. Thus the probability of error of the decision rule Ω under hypothesis H_i is

$$P_\Omega(e/H_i) = \mathsf{P}_{H_i}\left\{(\underline{x},\underline{y}) \in \bigcup_{\substack{j=1\\ j\neq i}}^{M} \Omega_j\right\} \tag{1}$$

where $\underline{x} = (x_1, x_2, \ldots, x_K)$, and $\underline{y} = (y_1, y_2, \ldots, y_K)$. We consider a generalized Neyman-Pearson criterion wherein we seek to ensure that all error probabilities decay exponentially in n with some predetermined slope λ, and simultaneously minimize the rejection probability subject to these constraints. Specifically, we seek optimal decisions rules Ω such that $\forall p_1, p_2, \ldots p_K \in \mathcal{P}(\mathsf{Z})$

$$\lim_{n\to\infty} \frac{1}{n} \log P_\Omega(e/H_i) \leq -\lambda, i = 1, \ldots, M, \tag{2}$$

and Ω_R is minimal.

III. Proposed solution

The problem of matching strings across two sets can be best visualized as a matching problem on a bipartite graph. Let $G = (V, E)$ denote a complete bipartite graph where each vertex in the set V of vertices denotes a unique element in $S_1 \cup S_2$. There exists an edge from each element i in S_1 to each element j in S_2 and no edges between elements in S_1 or S_2. Thus we have a complete bipartite graph where S_1 and S_2 form the two parts. Corresponding to the $M = K!$ different hypothesis, there are M possible *maximal matchings* on G. A matching is a subset S of edges E of G such that no two edges in S share a vertex. A maximal matching is a matching S such that no edge from G can be added to S while preserving the matching property. The matching corresponding to H_k is the maximal matching in which node i from S_2 is mapped to node $\pi_k(i)$ in S_1. The hypothesis testing task thus is equivalent to identifying the correct maximal matching.

A commonly used solution for universal and composite hypothesis testing problems is the generalized likelihood ratio test (GLRT). The first step for obtaining a GLRT solution is to maximize the log-likelihood function of the observations from all distributions under each hypothesis. For hypothesis H_i this is given by

$$\begin{aligned} L(H_i) &= \sup_{p_1,p_2,\ldots,p_K} \sum_{k=1}^{K} \left[\log p_k(x_{\pi_i(k)}) + \log p_k(y_k)\right] \\ &= -2n \sum_{k=1}^{K} \Big[\mathcal{H}(\Gamma_{x_{\pi_i(k)}}) + \mathcal{H}(\Gamma_{y_k}) \\ &\quad D(\Gamma_{x_{\pi_i(k)}} \| \tfrac{1}{2}(\Gamma_{x_{\pi_i(k)}} + \Gamma_{y_k})) \\ &\quad + D(\Gamma_{y_k} \| \tfrac{1}{2}(\Gamma_{x_{\pi_i(k)}} + \Gamma_{y_k}))\Big] \end{aligned} \tag{3}$$

where the second relation follows by noting that the original expression is maximized by choosing $p_k = \frac{1}{2}(\Gamma_{x_{\pi_i(k)}} + \Gamma_{y_k})$. Here $\mathcal{H}(p)$ denotes the entropy of distribution p. A good solution to the multiple hypothesis testing problem in practice is to decide in favor of the maximum-likelihood (ML) solution[1] given by

$$\widehat{H} = \arg\max_{H_i} L(H_i) \tag{4}$$

or equivalently,

$$\widehat{H} = \arg\min_{H_i} D(H_i) \tag{5}$$

where

$$\begin{aligned} D(H_i) &= \sum_{k=1}^{K} D(\Gamma_{x_{\pi_i(k)}} \| \tfrac{1}{2}(\Gamma_{x_{\pi_i(k)}} + \Gamma_{y_k})) \\ &\quad + D(\Gamma_{y_k} \| \tfrac{1}{2}(\Gamma_{x_{\pi_i(k)}} + \Gamma_{y_k})). \end{aligned} \tag{6}$$

This test can be interpreted as a minimum weight matching [12] on the complete bipartite graph G with appropriate weights assigned to the edges in E. For $i \in S_1$ and $j \in S_2$ let the weight w_{ij} of edge e_{ij} between them be given by

$$w_{ij} = D(\Gamma_{x_i} \| \tfrac{1}{2}(\Gamma_{x_i} + \Gamma_{y_j})) + D(\Gamma_{y_j} \| \tfrac{1}{2}(\Gamma_{x_i} + \Gamma_{y_j})). \tag{7}$$

Weight w_{ij} can be interpreted as a *distance measure* between strings x_i and y_j. The following proposition summarizes this result.

Proposition 3.1: The solution to (4) is given by the hypothesis corresponding to the permutation defined by the minimum weight matching on the bipartite graph G described above with weights given by (7) (refer to Figure 1). □

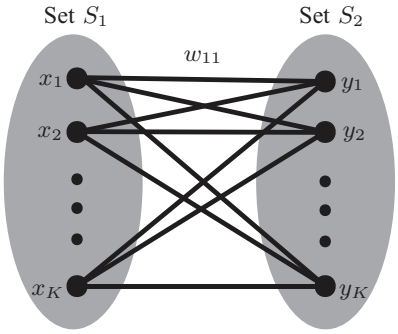

Fig. 1. The solution to the mutiple hypothesis testing problem given in (4) can be obtained by performing a minimum weight bipartite matching with weights given in (7).

The solution of (4) can be justified by the asymptotic optimality properties of a threshold test that uses this statistic. For proving asymptotic optimality we restrict ourselves to tests that are based only on the empirical distributions of the observations. For this purpose, we use Γ_{XY} to denote the collection of empirical distributions:

$$\Gamma_{XY} := (\Gamma_{x_1}, \Gamma_{x_2}, \ldots, \Gamma_{x_K}, \Gamma_{y_1}, \Gamma_{y_2}, \ldots, \Gamma_{y_K}).$$

This is justified in the asymptotic setting because of the following lemma.

[1]To be precise, this should be a maximum *generalized* likelihood solution because the data distributions $p_1, p_2, \ldots, p_K$ are unknown.

Lemma 3.2: Let $\Omega = (\Omega_1, \Omega_2, \ldots, \Omega_M, \Omega_R)$ be a decision rule based only on the sequences $\{x_1, x_2, \ldots, x_K\}$ and $\{y_1, y_2, \ldots, y_K\}$. Then there exists a decision rule $\Lambda = (\Lambda_1, \Lambda_2, \ldots, \Lambda_M, \Lambda_R)$ based on the sufficient statistics Γ_{XY} such that

$$\limsup_{n\to\infty} \frac{1}{n} \log P_\Lambda(e/H_i) \leq \limsup_{n\to\infty} \frac{1}{n} \log P_\Omega(e/H_i), \quad i = 1, 2, \ldots, M, \forall p_1, p_2, \ldots p_K \in \mathcal{P}(\mathsf{Z}) \tag{8}$$

and

$$\limsup_{n\to\infty} \frac{1}{n} \log |\Lambda_R| \leq \limsup_{n\to\infty} \frac{1}{n} \log |\Omega_R|. \tag{9}$$

□

We provide a proof in the appendix. Note that Λ_R and Ω_R are finite sets, thus their cardinality is well-defined.

In order to prove optimality, we allow for a no-match zone, i.e., we allow a decision in favor of rejecting all the M hypotheses. For this purpose, we need to identify the *second most likely hypothesis*. Let

$$\widetilde{H} = \arg\min_{H_i \neq \widehat{H}} D(H_i) \tag{10}$$

where $\widehat{H}$ is defined in (5). The optimal test with rejection is described in the following theorem.

Theorem 3.3: Let $\Omega = (\Omega_1, \Omega_2, \ldots, \Omega_M, \Omega_R)$ be a decision rule based on the collection Γ_{XY} of empirical distributions such that for all collections of distributions $p_1, p_2, \ldots, p_K$ from $\mathcal{P}(\mathsf{Z})$ we have

$$P_\Omega(e/H_i) \leq 2^{-\lambda n}, i = 1, 2, \ldots, M \tag{11}$$

when source k is distributed according to p_k for $k \in \{1, 2, \ldots, K\}$.

Let $\widetilde{\lambda} = \lambda - \frac{2K|\mathsf{Z}|\log(n+1)}{n}$,

$$\Lambda_i = \{\Gamma_{XY} : D(\widetilde{H}) \geq \widetilde{\lambda}, \widehat{H} = H_i\}, i = 1, 2, \ldots, M,$$

and

$$\Lambda_R = \{\Gamma_{XY} : D(\widetilde{H}) < \widetilde{\lambda}\}.$$

Then

$$\lim_{n\to\infty} \frac{1}{n} \log P_\Lambda(e/H_i) \leq -\lambda, \quad i = 1, 2, \ldots, M, \forall p_1, p_2, \ldots p_K \in \mathcal{P}(\mathsf{Z}) \tag{12}$$

and

$$\Lambda_R \subset \Omega_R. \tag{13}$$

□

We provide a proof to the theorem in the appendix.

The difference in the solution given by Theorem 3.3 and that proposed in (5) only arises due to the rejection region. As seen in the statement of the theorem, the no-match decision is made if the second most likely solution also has high likelihood. Due to this rejection region, we can now guarantee exponential decay of the error probabilities with a given exponent under the various hypotheses. Nevertheless, the optimality property of the solution given by Theorem 3.3 suggests that the ML estimate of (5) is a reasonable choice even if we do not permit a no-match decision.

IV. Experimental Evaluation

We applied the ML test proposed in (5) to mobility traces obtained from connections to Wi-Fi access points on the École Polytechnique Fédérale de Lausanne (EPFL) campus. In this experiment the datasets in S_1 and S_2 correspond to the users' mobility traces measured over two different non-overlapping time periods. Using only the frequency of visits of the various users to various access points in each time-period we show that users' mobility traces in S_1 can be matched to those in S_2 with high accuracy.

A. Dataset Description and Preprocessing

The EPFL campus consists of several buildings and hundreds of wireless access points (APs) (refer to Figure 2). The main wireless network on the campus requires authentication, and can thus be accessed only by members of the university (students, faculty, etc.). The history of connections of every device to the network is recorded in the following way: Whenever a device (user) connects to the network, its (anonymized) MAC address, the ID of the AP to which it connects, and the time of start of the connection measured to a precision of one second, are stored in a log file. When the device moves across the campus and gets connected to a new AP, the time of this new connection and the ID of the AP are similarly stored. However, if a device loses its connection or disconnects, it is not recorded in the log file, unless it reconnects to one of the APs. For our experiments we used the information available in the log file of all such accesses for two consecutive weeks during the academic semester. For privacy reasons, all the MAC addresses inside the log file are encrypted, however the encryption key is the same for the two weeks period, and thus it is possible to recognize a MAC address across different days.

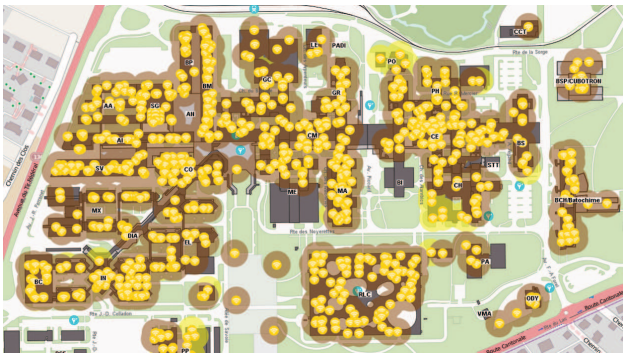

Fig. 2. The position of wireless access points on the EPFL campus. The campus has several buildings and around one thousand wireless access points.

For obtaining the ground truth for our statistics-matching experiments, we estimated the trajectories of all the users on the campus by using the log file. There are two main sources of error: First, all the wireless devices (laptops, smart-phones, etc.), connected at any time to the network, appear in the log file; therefore, if a user does not connect his device to the network, or does not carry the device everywhere he goes, his true trajectory cannot be reconstructed. Second,

whenever a user leaves the campus (disconnects), the time of disconnection is unknown. For reconstructing the trajectories, we assumed that a device remains connected to the same AP until the time when it is connected to a new AP (based on the log file entries). When a device is connected to an AP, it stays in the communication range of the AP (typically 50-100 m), specifically, in the AP's *vicinity*. Thus the reconstructed trajectories of the users have a spatial resolution equal to the coverage region of an AP level of granularity. Further, since the connections to APs are monitored every second, the time resolution is equal to one second. In other words, the user trajectories are reconstructed as a sequence of spatial locations sampled every second. We use these sequence of reconstructed user locations as the data vectors $x_.$'s and $y_.$'s in our experiments. Although the reconstructed trajectories are affected by the above mentioned sources of error, they are reconstructed based on actual wireless connection logs.

B. Experiment One

For our first experiment, we considered all users who are on the campus during the interval 9h00–17h00 on both Mondays during the two weeks period. There are $K = 1154$ such users. For simplicity of exposition let the true matching π be the identity permutation. Then trajectory of a user i in the first Monday and the second Monday can be seen as strings x_i and y_i, respectively, with length $n = (17\text{h}00 - 9\text{h}00) \times 3600 = 28800$. The alphabet Z consists of the set of APs and $|\mathsf{Z}| = 934$. The mean, median, maximum, and standard deviation of the number of visited APs by the users during a Monday are equal to 11, 8, 60, and 10.4, respectively. As in Section II we assumed that for a user i, elements of strings x_i and y_i are drawn in an i.i.d. manner from an unknown underlying distribution, which is specific to user i. Sets S_1 and S_2 consist of the strings of all the users in the first Monday and the second Monday, respectively. The empirical distributions Γ_{x_i} and Γ_{y_i} are equal to the proportion of time that a user i spent in different APs in the first Monday and the second Monday, respectively.

After computing Γ_{x_i} and Γ_{y_i} for every user i, we constructed a complete bipartite graph described in Section III with edge weights given in (7), and computed a minimum weight matching on the graph. The obtained results are shown in Table I under "Matching K users". There are 610 out of 1154 users correctly matched which gives 52.9% accuracy. We observe that although the underlying i.i.d. assumption of users' position at every second of their trajectory is inherently false, the obtained accuracy is high considering the large number of users. This means that given the anonymized proportions of time that 1154 users spend between 9h00–17h00 in different APs across campus on two consecutive Mondays, we are able to correctly match (de-anonymize) more than half of them. Or from a privacy perspective, given anonymized temporal averages of all these users on one Monday, the identities of more than half of these users can be identified by tracking these users on a different Monday.

Days included in dataset	# users (K)	Matching K users in second week		Matching 1 random user in second week
		# correct matches	Accuracy	Fraction of correct matches
Mondays	1154	610	52.9%	44.5%
	2174	934	43.0%	32.9%
Mondays and Tuesdays	1047	738	70.5%	53.5%

TABLE I

RESULTS OF DE-ANONYMIZATION EXPERIMENTS. NOTE THAT (I) INCREASING THE NUMBER OF USERS IN AN EXPERIMENT LEADS TO A REDUCTION IN THE ACCURACY OF THE MATCHING, AND (II) MATCHING INFORMATION ABOUT ALL USERS IN THE SECOND WEEK GIVES STRICTLY BETTER ACCURACY THAN THAT OBTAINED WHEN MATCHING ONLY ONE USER.

We repeated the above experiment by considering all users who were on campus during the interval 10h00–17h00 on both Mondays during the two weeks period. As this interval is smaller, the string length is lower, equal to $n = (17\text{h}00 - 10\text{h}00) \times 3600 = 25200$, and there are more users on campus; namely, $K = 2174$ users. The obtained results are shown in Table I. There are 934 users correctly matched out of a total of 2174 users which gives 43.0% accuracy. As there are many more users and the number of samples n is smaller, there is less information available for finding the correct matching and thus the obtained accuracy is lower than that in the previous experiment.

C. Experiment Two

In our second experiment, we investigated whether the matching accuracy can be improved by using statistics of users from two different days of the week. We considered all users who are on campus during the interval 10h00–17h00 on both Mondays and both Tuesdays during the two weeks period. There are $K = 1047$ such users. The trajectory of every user in each day can be seen as a string with length $n = (17\text{h}00 - 10\text{h}00) \times 3600 = 25200$. For each user we computed four different empirical distributions associated with the four strings. We observed that combining the traces of each user from Monday and Tuesday to form a single long trace leads to poor matching accuracy - we obtain only 14.0% accuracy. The reason for this is that users have different time tables for different days of the week, and hence by combining different days of the week the users tend to be less distinct. For this reason, in the experiment we assumed that each user has different distributions for different days of the week, and sought a method to match across weeks using the statistics of both Monday and Tuesday. Since we have statistics of two days on two weeks for each user, this problem does not fit the exact structure of that studied in Section II and hence we used a modification of the decision obtained in Section III. We constructed a complete bipartite graph whose edge weights are equal to the summation of weights in (7)

corresponding to the two Mondays and to the two Tuesdays. After computing a minimum weight matching of the graph, 738 out of 1047 users correctly match which gives 70.5% accuracy (refer to the last row of Table I under "Matching K users"). The obtained accuracy is significantly higher that that of the first experiment using 1154 users. This is because we have more information available for matching the users, namely, four days' statistics instead of two days.

D. Experiment Three

We repeated the two previous experiments in the setting in which we are given statistic of all users in the first week and only of one user in the second week. The objective is to match the user's statistics from the second week to the correct statistics from the first week. This corresponds to the case in which set S_2 described in Section II is a singleton, which also corresponds to the setting addressed by Gutman in [3]. We proceed as follows.

We first generalize experiment one to this setting. We let S_1 be the collection of traces of all users on the Monday of the first week as in Section IV-B. For S_2 we use the trace of a randomly selected user on the Monday of the second week. Let Γ_{y_j} denote the empirical distribution of this trace. We match this trace to the trace x_i from the first week that gives the minimum value of the weight computed in (7). If there exist ties (i.e., multiple users in S_1 having minimum weight), we break them randomly. This is exactly the algorithm studied by Gutman [3]. We estimate the average probability of correct matching under this procedure by computing the fraction of choices of S_2 that lead to correct matchings. Following a similar approach, we also generalize experiment two to this setting, i.e., when data from both Mondays and Tuesdays are available in both weeks.

The obtained probabilities are shown in the last column of Table I for the datasets used in experiments one and two. The observed behavior is similar to that when all the K users are matched: matching probability decreases when the number of users increases, and increases when the available information increases. We also observe that the obtained accuracy of matching is lower than the accuracy obtained with the same dataset when all the users are matched. This is expected because in the latter case we have more data (set S_2 is larger) and thus we can do a better matching. This observation has important implications in the perspective of privacy of anonymized statistics. A user's privacy depends not only on how much her trajectory is revealed to the adversary, but also on how much others' trajectories are revealed to the adversary.

V. Conclusion

In this paper we have studied strategies for de-anonymizing anonymized user statistics given auxiliary information about the user's behavior. We obtained an asymptotically optimal strategy for this problem assuming an i.i.d. model for the users' data. We focussed primarily on the setting in which auxiliary information about all the users are available in the form of independent data strings of all users. It may be possible to extend the optimality result to the case where only auxiliary information about a subset of the users are available, a special case of which was studied by Gutman [3], where auxiliary information about only one user is available. Similarly, although in this paper we have assumed that the length of the data-strings in the anonymized statistics and the auxiliary information are all equal to n, the proposed solution and optimality result can easily be generalized to the case where these are distinct, following the same steps as in [3], provided that the length of all data-strings are of equal order. We also saw the performance obtained with the proposed algorithm on real mobility traces recorded on two different days. We saw that de-anonymization can be performed with higher accuracy if traces about all users are available on the second day, as against having information only about a single user. One aspect that we did not consider in the location de-anonymization is that the geometric distance between various locations may be available. In practice, it may be possible to perform better matching by taking this information into account, although obtaining optimality results may be hard.

Acknowledgments

The authors thank Richard Timsit and Yves Despond for providing us with the EPFL Wi-Fi connection logs, and Elio Abi Karam for his help in the experimental evaluation. This research was supported in part by ERC Advanced Investigators Grant: Sparse Sampling: Theory, Algorithms and Applications SPARSAM no 247006.

Appendix

We need the following lemma for the proofs. Recall that for any string $s \in \mathsf{Z}^n$ we use T_s to denote the *type class* of s, i.e., the set of all strings of length n with the same empirical distribution as s. The following lemma is well known (see e.g., [13, Ch. 12]).

Lemma 1.1: For every $p \in \mathcal{P}(\mathsf{Z})$ and every $s \in \mathsf{Z}^n$,

$$\frac{1}{(n+1)^{|\mathsf{Z}|}} 2^{-nD(\Gamma_s \| p)} \leq \mathsf{P}_p(T_s) \leq 2^{-nD(\Gamma_s \| p)}$$

where P_p denotes the probability measure when all observations in s are drawn i.i.d. according to law p. □

A. Proof of Lemma 3.2

Consider an arbitrary tuplet of sequences $(x_1, x_2, \ldots, x_K, y_1, y_2, \ldots, y_K)$. Let $T = (T_{x_1}, \ldots, T_{x_K}, T_{y_1}, \ldots, T_{y_K})$ denote the joint type-class of all the sequences. Any $(x'_1, x'_2, \ldots, x'_K, y'_1, y'_2, \ldots, y'_K) \in T$ belongs to exactly one of the sets $\Omega_1, \Omega_2, \ldots, \Omega_M, \Omega_R$. We modify the decision rule Ω as follows. For any type T we let Λ_i include T if Ω_i contains the most number of the sequences of T, for $i \in \{1, 2, \ldots, M, R\}$. In case of ties we break them arbitrarily and include T in exactly one of the Λ_i's.

By construction we have for any type $T \subset \Lambda_R$

$$|\Omega_R| \geq \frac{1}{M+1}|T|.$$

Moreover, we have

$$\begin{aligned}|\Lambda_R| &= \sum_{T\subset\Lambda_R}|T| \\ &\leq \sum_{T\subset\Lambda_R}(M+1)|\Omega_R| \\ &\leq |\Omega_R|(1+(M+1)\tau_n)\end{aligned}$$

where τ_n represents the number of types of length n. Since $\frac{\log\tau_n}{n}\to 0$ [13] we have (9).

Now for any type $T\subset\Lambda_i$ with $i\in\{1,2,\ldots,M\}$ we have by Lemma 1.1 and definition of Λ_i:

$$\begin{aligned}\mathsf{P}_{H_i}\{\Omega_i\} &\geq \mathsf{P}_{H_i}\{\Omega_i\cap T\}\geq\frac{1}{M+1}\mathsf{P}_{H_i}\{T\} \\ &\geq \frac{2^{-n\sum_{k=1}^{K}\left(D(\Gamma_{x_{\pi_i(k)}}\|p_k)+D(\Gamma_{y_k}\|p_k)+\delta(n)\right)}}{M+1}\end{aligned}$$

where $\delta(n)=\frac{2|\mathsf{Z}|\log(n+1)}{n}$. Combining the above result along with the definition of Λ_i and Lemma 1.1, we have

$$\begin{aligned}\mathsf{P}_{H_i}\{\Lambda_i\} &= \sum_{T\subset\Lambda_i}\mathsf{P}_{H_i}\{T\} \\ &\leq \sum_{T\subset\Lambda_i}2^{-n\sum_{k=1}^{K}\left(D(\Gamma_{x_{\pi_i(k)}}\|p_k)+D(\Gamma_{y_k}\|p_k)\right)} \\ &\leq \sum_{T\subset\Lambda_i}2^{n\delta(n)}(M+1)\mathsf{P}_{H_i}\{\Omega_i\} \\ &\leq \tau_n 2^{n\delta(n)}(M+1)\mathsf{P}_{H_i}\{\Omega_i\}\end{aligned}$$

Since $\frac{\log\tau_n}{n}\to 0$ [13] we have (8).

B. Proof of Theorem 3.3

Define

$$\widetilde{\Lambda}_i=\{\Gamma_{XY}: D(H_i)\geq\widetilde{\lambda}\}, i=1,2,\ldots,M.$$

Clearly,

$$\Lambda_j\subset\widetilde{\Lambda}_i \text{ for all } j\neq i$$

and hence

$$\cup_{j\neq i}\Lambda_j\subset\cup_{j\neq i}\left(\cap_{k\neq j}\widetilde{\Lambda}_k\right)\subset\widetilde{\Lambda}_i.$$

Therefore,

$$\begin{aligned}P_\Lambda(e/H_i) &= \sum_{\cup_{j\neq i}\Lambda_j}\prod_{k=1}^{K}p_k(x_{\pi_i(k)})p_k(y_k) \\ &\leq \sum_{\widetilde{\Lambda}_i}\prod_{k=1}^{K}p_k(x_{\pi_i(k)})p_k(y_k) \\ &\overset{(a)}{\leq} \sum_{\widetilde{\Lambda}_i}\prod_{k=1}^{K}2^{-2n\mathcal{H}(\frac{1}{2}(\Gamma_{x_{\pi_i(k)}}+\Gamma_{y_k}))} \\ &= \sum_{\widetilde{\Lambda}_i}2^{-2n\sum_{k=1}^{K}\mathcal{H}(\frac{1}{2}(\Gamma_{x_{\pi_i(k)}}+\Gamma_{y_k}))} \\ &= \sum_{\widetilde{\Lambda}_i}2^{-n\sum_{k=1}^{K}\Big(\mathcal{H}(\Gamma_{x_{\pi_i(k)}})+\mathcal{H}(\Gamma_{y_k})+} \\ &\qquad {\scriptstyle D(\Gamma_{x_{\pi_i(k)}}\|\frac{1}{2}(\Gamma_{x_{\pi_i(k)}}+\Gamma_{y_k}))+} \\ &\qquad {\scriptstyle D(\Gamma_{y_k}\|\frac{1}{2}(\Gamma_{x_{\pi_i(k)}}+\Gamma_{y_k}))\Big)} \\ &\leq \sum_{\widetilde{\Lambda}_i}2^{-n\sum_{k=1}^{K}\left(\mathcal{H}(\Gamma_{x_{\pi_i(k)}})+\mathcal{H}(\Gamma_{y_k})+\widetilde{\lambda}\right)} \\ &\leq 2^{-n\widetilde{\lambda}}\sum_{\mathbf{Z}}2^{-n\sum_{k=1}^{K}\left(\mathcal{H}(\Gamma_{x_{\pi_i(k)}})+\mathcal{H}(\Gamma_{y_k})\right)} \\ &\leq 2^{-n\widetilde{\lambda}}\end{aligned}$$

where (a) follows from the inequality $p(s)\leq 2^{-n\mathcal{H}(\Gamma_s)}$ for all p. This proves (12). For proving (13) we observe that for any test based on empirical distributions, we have

$$\begin{aligned}2^{-\lambda n} &\geq P_\Omega(e/H_i) \\ &= \sum_{\cup_{j\neq i}\Omega_j}\prod_{k=1}^{K}p_k(x_{\pi_i(k)})p_k(y_k) \\ &\overset{(a)}{\geq} \sum_{T\subset\cup_{j\neq i}\Omega_j}2^{-n\sum_{k=1}^{K}\Big(D(\Gamma_{x_{\pi_i(k)}}\|p_k)+} \\ &\qquad {\scriptstyle D(\Gamma_{y_k}\|p_k)+\delta(n)\Big)} \\ &\geq 2^{-n\sum_{k=1}^{K}\left(D(\Gamma_{x'_{\pi_i(k)}}\|p_k)+D(\Gamma_{y'_k}\|p_k)+\delta(n)\right)}\end{aligned}$$

where (a) follows from Lemma 1.1 with $T=(T_{x_1},\ldots,T_{x_K},T_{y_1},\ldots,T_{y_K})$ and $\delta(n)=\frac{2|\mathsf{Z}|\log(n+1)}{n}$, and $(x'_1,x'_2,\ldots,x'_K,y'_1,y'_2,\ldots,y'_K)\in\cup_{j\neq i}\Omega_j$ and $p_1,p_2,\ldots,p_K\in\mathcal{P}(\mathsf{Z})$ is arbitrary. Now letting $p_k=\frac{1}{2}(\Gamma_{x'_{\pi_i(k)}}+\Gamma_{y'_k})$ we get

$$\begin{aligned}\lambda &\leq \sum_{k=1}^{K}\Big(D(\Gamma_{x'_{\pi_i(k)}}\|\tfrac{1}{2}(\Gamma_{x'_{\pi_i(k)}}+\Gamma_{y'_k}))+ \\ &\qquad D(\Gamma_{y'_k}\|\tfrac{1}{2}(\Gamma_{x'_{\pi_i(k)}}+\Gamma_{y'_k}))+\delta(n)\Big)\end{aligned}$$

which further implies that

$$\cup_{j\neq i}\Omega_j\subset\widetilde{\Lambda}_i. \tag{14}$$

Now let

$$\widehat{\Lambda}_i:=\cap_{j\neq i}\widetilde{\Lambda}_j.$$

Hence,

$$\cup_i \Lambda_i = \{\Gamma_{XY} : D(\widetilde{H}) \geq \widetilde{\lambda}\} = \cup_i \widehat{\Lambda}_i.$$

Combining with (14) we get

$$\widehat{\Lambda}_i = \cap_{j \neq i} \widetilde{\Lambda}_j \supset \cap_{j \neq i} \cup_{k \neq j} \Omega_k \supset \Omega_i$$

and thus

$$\Lambda_R^c = \cup_i \Lambda_i = \cup_i \widehat{\Lambda}_i \supset \cup_i \Omega_i = \Omega_R^c.$$

Hence

$$\Lambda_R \subset \Omega_R.$$

REFERENCES

[1] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. 2008 IEEE Symposium on Security and Privacy*, Washington, DC, USA. [Online]. Available: http://dx.doi.org/10.1109/SP.2008.33

[2] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the Crowd: The privacy bounds of human mobility," *Scientific Reports*, vol. 3, Mar. 2013. [Online]. Available: http://dx.doi.org/10.1038/srep01376

[3] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 401–408, 1989.

[4] L. Sweeney, "Weaving technology and policy together to maintain confidentiality," *The Journal of Law, Medicine & Ethics*, vol. 25, no. 2-3, pp. 98–110, 1997.

[5] X. Xiao, Y. Zheng, Q. Luo, and X. Xie, "Finding similar users using category-based location history," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2010, pp. 442–445.

[6] C. Y. Ma, D. K. Yau, N. K. Yip, and N. S. Rao, "Privacy vulnerability of published anonymous mobility traces," in *Proceedings of the sixteenth annual international conference on Mobile computing and networking*. ACM, 2010, pp. 185–196.

[7] J. Freudiger, R. Shokri, and J.-P. Hubaux, "Evaluating the privacy risk of location-based services," in *Financial Cryptography and Data Security*. Springer, 2012, pp. 31–46.

[8] S. Gambs, M.-O. Killijian, and M. Nunez Del Prado Cortez, "De-anonymization attack on geolocated datasets," in *Proceedings of the The 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (IEEE TrustCom-13)*, Melbourne, Australie, Jul. 2013, p. 9p. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00718763

[9] Y. De Mulder, G. Danezis, L. Batina, and B. Preneel, "Identification via location-profiling in gsm networks," in *Proceedings of the 7th ACM workshop on Privacy in the electronic society*. ACM, 2008, pp. 23–32.

[10] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *Proceedings of the 17th annual international conference on Mobile computing and networking*. ACM, 2011, pp. 145–156.

[11] M. Srivatsa and M. Hicks, "Deanonymizing mobility traces: Using social network as a side-channel," in *Proceedings of the 2012 ACM conference on Computer and communications security*. ACM, 2012, pp. 628–637.

[12] D. B. West *et al.*, *Introduction to graph theory*. Prentice hall Englewood Cliffs, 2001, vol. 2.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006.