Fundamental Limits of Database Alignment

Daniel Cullina
Dept. of Electrical Engineering
Princeton University
dcullina@princeton.edu

Prateek Mittal
Dept. of Electrical Engineering
Princeton University
pmittal@princeton.edu

Negar Kiyavash
Dept. of Electrical and Computer Engineering
Dept. of Industrial and Enterprise Systems Engineering
Coordinated Science Lab
University of Illinois at Urbana-Champaign
kiyavash@illinois.edu

***Abstract*—We consider the problem of aligning a pair of databases with correlated entries. We introduce a new measure of correlation in a joint distribution that we call cycle mutual information. This measure has operational significance: it determines whether exact recovery of the correspondence between database entries is possible for any algorithm. Additionally, there is an efficient algorithm for database alignment that achieves this information theoretic threshold.**

I. The database deanonymization problem

Suppose that we have two databases. Each item in the databases contain information about a single individual. Some individuals appear in both databases. When a entry in the first database and an entry in the second database concern the same individual, their contents are correlated. The entries may be two noisy observations of the same signal, they may be two completely different types of data that have some correlation through population statistics, or they may even be correlated though the sampling process used to determine which individuals appear in the database.

We consider the following question: If the databases are published with user identities removed from each entry, is it possible to learn the association between database entries that correspond to the same individual by exploiting the correlation between them?

Clearly, when there is enough correlation between entries about the same individual and the databases are small enough, it is possible to learn the true alignment between the database entries. Our goal is to find the precise conditions under which it is possible to learn the complete correspondence between entries with high probability. In particular, we would like to determine the measure of correlation that characterizes feasibility of perfect deanonymization in this setting.

This framework for database alignment is related to several practical deanonymization attacks. Narayanan and Shmatikov linked an anonymized dataset of film ratings to a publicly available dataset using correlations between the ratings [1]. Differential privacy has been widely used to quantifying privacy issues related to databases [2]. More recently, generative adversarial privacy has been proposed [3]. In both cases, if users are present in multiple databases, knowledge of alignment is required to fully apply these frameworks.

Takbiri, Houmansadr, Goeckel, and Pishro-Nik have recently investigated a closely related user privacy problem [4].

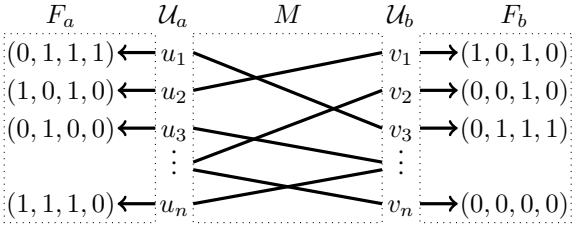

Fig. 1. Two databases, F_a and F_b, with alphabets $\mathcal{X}_a = \mathcal{X}_b = \{0,1\}^4$ and a matching M between their user identifier sets.

A. Notation

For finite sets $\mathcal{X}$ and $\mathcal{Y}$, let $\mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ be the set of real-valued matrices with rows indexed by $\mathcal{X}$ and columns indexed by $\mathcal{Y}$. For $x \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$, let $x^{\odot k} \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ be the entry-wise power of x, i.e. the matrix such that $(x^{\odot k})_{i,j} = (x_{i,j})^k$. Let $x^{\otimes k} \in \mathbb{R}^{\mathcal{X}^k\times\mathcal{Y}^k}$ be the tensor power of x, i.e. the matrix such that for $a \in \mathcal{X}^k$ and $b \in \mathcal{Y}^k$, $(x^{\otimes k})_{a,b} = \prod_{i=0}^{k-1} x_{a_i,b_i}$.

Let $\mathbb{P}(\mathcal{X})$ be the set of probability distributions on $\mathcal{X}$.

B. Formal description

We have the following sets related to the user identifiers:

$\mathcal{U}_a$	Set of user identifiers in the first database
$\mathcal{U}_b$	Set of user identifiers in the second database
$M \subseteq \mathcal{U}_a \times \mathcal{U}_b$	Bijective matching between the two types of user identifierss

A bijection between $\mathcal{U}_a$ and $\mathcal{U}_b$ is a subset of $\mathcal{U}_a \times \mathcal{U}_b$ in which each element of $\mathcal{U}_a$ and $\mathcal{U}_b$ appears exactly once. The matching M contains the pairs of ids that correspond to the same user. The fact that M is a bijection implies that $|M| = |\mathcal{U}_a| = |\mathcal{U}_b|$. Throughout, we let $n = |M|$.

We have the following sets, functions, and distributions associated with the databases:

$\mathcal{X}_a$	Alphabet of entries in first database
$\mathcal{X}_b$	Alphabet of entries in second database
$F_a : \mathcal{U}_a \to \mathcal{X}_a$	First database
$F_b : \mathcal{U}_b \to \mathcal{X}_b$	Second database
$F = (F_a, F_b)$	
$p \in \mathbb{P}(\mathcal{X}_a \times \mathcal{X}_b)$	Joint distribution between related entries
$p_a \in \mathbb{P}(\mathcal{X}_a)$	Marginal distribution on first alphabet
$p_b \in \mathbb{P}(\mathcal{X}_b)$	Marginal distribution on second alphabet

Figure 1 illustrates a pair of databases.

978-1-5386-4780-6/18/$31.00©2018 IEEE

C. Generative model

For each user $u \in \mathcal{U}_a$, there is a database entry $F_a(u) \in \mathcal{X}_a$. For a pair $(u, v) \in M$, the entries $F_a(u)$ and $F_b(v)$ are correlated via the joint distribution p:

$$\Pr[F_a(u) = i, F_b(v) = j | M] = p(i, j).$$

For distinct $u, v \in \mathcal{U}_a$, $F_a(u)$ and $F_a(v)$ are independent. The same is true for distinct $u, v \in \mathcal{U}_b$. Thus we define

$$r(f_a, f_b; m) = \prod_{(u,v)\in m} p(f_a(u), f_b(v))$$

so the joint distribution of the databases is

$$\Pr[F_a = f_a, F_b = f_b | M = m] = r(f_a, f_b; m). \quad (1)$$

D. Relationship to graph alignment

The methods used in this paper are related to those used to analyze information theoretic thresholds for exact graph alignment [5]–[7]. An undirected graph G can be represented by its edge indicator function: $\binom{V(G)}{2} \to \{0, 1\}$, so we have a very simple type of information about each user pair. The analogue to the generative model (1) is the correlated Erdős-Rényi distribution on graph pairs, where corresponding edge indicator r.v.s are sampled i.i.d. from some joint distribution on $\{0, 1\}^2$. Once the marginal distributions are fixed, the one remaining degree of freedom specifies the level of correlation.

In the database problem, we instead have larger blocks of information about individual users. This allows for more complicated forms of correlations. In this paper, we identify the relevant one-dimensional summary of that correlation.

A further connection is that graph alignment falls into the database alignment framework when seed vertices are used [8], [9]: the list of adjacent seeds is essentially a database entry.

II. Results

Both our achievability and converse bounds use the following measure of correlation in a joint distribution. We propose to call this quantity *cycle mutual information*.

Definition 1. *For $p \in \mathbb{P}(\mathcal{X}_a \times \mathcal{X}_b)$, let $z \in \mathbb{R}^{\mathcal{X}_a \times \mathcal{X}_b}$ be the matrix such that $z_{i,j} = \sqrt{p(i, j)}$ for $i \in \mathcal{X}_a$ and $j \in \mathcal{X}_b$. For an integer $\ell \geq 2$, define the order-ℓ cycle mutual information*

$$I^\circ_\ell(p) = \frac{1}{1-\ell} \log \operatorname{tr}((zz^T)^\ell).$$

Then z has a singular value decomposition $z = U\Sigma V^T$ where $\Sigma = \operatorname{diag}(\sigma)$. Observe that

$$\operatorname{tr}(\Sigma^2) = \operatorname{tr}(U\Sigma V^T V \Sigma U^T) = \operatorname{tr}(zz^T) = \sum_{i,j} z^2_{i,j} = 1, \quad (2)$$

so $\sigma^{\odot 2}$, the vector of squared singular values, constitutes a probability distribution. Thus we have another expression for cycle mutual information of order ℓ: $I^\circ_\ell(p) = H_\ell(\sigma^{\odot 2})$, where H_ℓ is the Rényi entropy of order ℓ. This expression allows us to extend the definition of $I^\circ_\ell(p)$ to all nonnegative real ℓ.

Our achievability theorem allows for arbitrary structure in the joint distribution of database entries.

Theorem 1. *Let $M \subseteq \mathcal{U}_a \times \mathcal{U}_b$ be a uniformly random bijection. Let the alphabets $\mathcal{X}_a$ and $\mathcal{X}_a$ and the joint distribution $p \in \mathbb{P}(\mathcal{X}_a \times \mathcal{X}_b)$ depend on n. If*

$$I^\circ_2(p) \geq 2 \log n + \omega(1),$$

there is an estimator for M given F that is correct with probability $1 - o(1)$.

When the database entries are vectors of independent identically distributed components, we have a converse bound with a leading term that matches the achievability.

Theorem 2. *Let $M \subseteq \mathcal{U}_a \times \mathcal{U}_b$ be a uniformly random bijection. Fix alphabets $\mathcal{Y}_a$ and $\mathcal{Y}_b$ and a joint distribution $q \in \mathbb{P}(\mathcal{Y}_a \times \mathcal{Y}_b)$. Let $\mathcal{X}_a = \mathcal{Y}_a^\ell$, $\mathcal{X}_b = \mathcal{Y}_b^\ell$, and $p = q^{\otimes \ell}$, where ℓ can depend on n. If*

$$I^\circ_2(p) \leq (2 - \Omega(1)) \log n,$$

any estimator for M given F is correct with probability $o(1)$.

III. MAP estimation

The optimal estimator for M given F is the maximum a posteriori estimator:

$$\begin{aligned}\hat{m}(f_a, f_b) &= \operatorname*{argmax}_m \Pr[M = m | F = (f_a, f_b)] \\ &= \operatorname*{argmax}_m \frac{\Pr[F = (f_a, f_b) | M = m] \Pr[M = m]}{\Pr[F = (f_a, f_b)]} \\ &\overset{(a)}{=} \operatorname*{argmax}_m \Pr[F = (f_a, f_b) | M = m].\end{aligned}$$

In (a) we use that fact that M is uniformly distributed.

Define the event

$$\mathcal{E}_{m_2, m_1} = \{(f_a, f_b) : r(f_a, f_b; m_2) \geq r(f_a, f_b; m_1)\}.$$

When m_1 is the true matching, this is the error event in which m_2 is incorrectly preferred to m_1.

A. Algorithm for computing the MAP estimator

Define the matrix $Q(f_a, f_b) \in \mathbb{R}^{\mathcal{X}_a \times \mathcal{X}_b}$,

$$Q(f_a, f_b)_{u,v} = \log p(f_a(u), f_b(v)).$$

The MAP estimator is the max weight matching in $Q(f_a, f_b)$:

$$\hat{m}(f_a, f_b) = \operatorname*{argmax}_m \sum_{(u,v)\in m} Q(f_a, f_b)_{u,v}.$$

Thus $\hat{m}$ can be computed in $\mathcal{O}(n^3)$ time [10].

IV. Generating functions

Let x and y be two matrices of formal variables indexed by $\mathcal{X}_a \times \mathcal{X}_b$, and let x_a and y_a be vectors of formal variables indexed by $\mathcal{X}_a$, and let x_b and y_b be vectors of formal variables indexed by $\mathcal{X}_b$. For a matching $m \in \mathcal{U}_a \times \mathcal{U}_b$ and a pair of databases $f_a : \mathcal{U}_a \to \mathcal{X}_a$ and $f_b : \mathcal{U}_b \to \mathcal{X}_b$, define the generating function of the joint type

$$t(m; f_a, f_b; x) = \prod_{(u,v)\in m} x_{f_a(u), f_b(v)}.$$

Observe that $t(m; f_a, f_b; p) = r(f_a, f_b; m)$.

For a pair of matchings, define the generating function

$$B_{m_1,m_2}(x,y) = \sum_{f_a:\mathcal{U}_a\to\mathcal{X}_a} \sum_{f_b:\mathcal{U}_b\to\mathcal{X}_b} t(m_1; f_a, f_b; x)t(m_2; f_a, f_b; y).$$

By understanding the behavior of this generating function, we can obtains upper bounds on the probability of an estimator making an error.

Throughout this section, let $z \in \mathbb{R}^{\mathcal{X}_a\times\mathcal{X}_b}$ be a matrix and let $z_a \in \mathbb{R}^{\mathcal{X}_a}$ and $z_b \in \mathbb{R}^{\mathcal{X}_b}$ be vectors such that $z_{i,j} = \sqrt{p(i,j)}$, $(z_a)_i = \sqrt{p_a(i)}$, and $(z_b)_j = \sqrt{p_b(j)}$.

Lemma 1. *For any two bijections $m_1, m_2 \subseteq \mathcal{U}_a \times \mathcal{U}_b$,*

$$\Pr[\mathcal{E}_{m_2,m_1}|M = m_1] \le B_{m_1,m_2}(z,z)$$

Proof: For any $\theta \ge 0$, we have

$$\begin{aligned}
&\Pr[\mathcal{E}_{m_2,m_1}|M = m_1] \\
&= \mathbb{E}\left[\mathbf{1}\left(\frac{r(f_a,f_b;m_2)}{r(f_a,f_b;m_1)} \ge 1\right)\middle| M = m_1\right] \\
&\le \mathbb{E}\left[\left(\frac{r(f_a,f_b;m_2)}{r(f_a,f_b;m_1)}\right)^{\theta}\middle| M = m_1\right].
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{r(f_a,f_b;m_2)}{r(f_a,f_b;m_1)}\right)^{\theta}\middle| M = m_1\right] \\
&= \sum_{f_a,f_b}\left(\frac{r(f_a,f_b;m_2)}{r(f_a,f_b;m_1)}\right)^{\theta} r(f_a,f_b;m_1) \\
&= \sum_{f_a,f_b} r(f_a,f_b;m_2)^{\theta} r(f_a,f_b;m_1)^{1-\theta} \\
&= \sum_{f_a,f_b} t(m_1;f_a,f_b;p)^{\theta} t(m_2;f_a,f_b;p)^{1-\theta} \\
&= \sum_{f_a,f_b} t(m_1;f_a,f_b;p^{\odot\theta}) t(m_2;f_a,f_b;p^{\odot(1-\theta)}) \\
&= B_{m_1,m_2}(p^{\odot\theta}, p^{\odot(1-\theta)})
\end{aligned}$$

where the matrix and vector exponents with $\odot$ are applied entrywise. Selecting $\theta = \frac{1}{2}$ gives the claim. ∎

Define the generating function

$$b^{\circ}_{\ell}(x,y) = \mathrm{tr}((xy^T)^{\ell}).$$

Regard m_1 as a function $\mathcal{X}_a \to \mathcal{X}_b$ and regard m_2^T as a function $\mathcal{X}_b \to \mathcal{X}_a$. Then their composition $m_2^T \circ m_1$ is a permutation of $\mathcal{X}_a$

Lemma 2. *Let $m_1, m_2 \subseteq \mathcal{U}_a \times \mathcal{U}_b$ be bijections. Let t°_{ℓ} be the number of cycles of length ℓ in the permutation $m_2^T \circ m_1$. Then $t^{\circ}_1 = |m_1 \cap m_2|$, $\sum_{\ell} \ell t^{\circ}_{\ell} = |\mathcal{X}_a|$, and*

$$B_{m_1,m_2}(x,y) = \prod_{\ell\in\mathbb{N}} (b^{\circ}_{\ell}(x,y))^{t^{\circ}_{\ell}}.$$

Lemma 3. *For $z' \in \mathbb{R}^{\mathcal{X}_a\times\mathcal{X}_b}$ with nonnegative entries and for $\ell \ge 2$, $b^{\circ}_{\ell}(z',z') \le b^{\circ}_2(z',z')^{\ell/2}$.*

Proof: We have $b^{\circ}_{\ell}(z',z') = \sum_k \sigma_k^{2\ell}$ where σ_k are the singular values of z'. By a standard inequality on p-norms, $\sum_k \sigma_k^{2\ell} \le \left(\sum_k \sigma_k^4\right)^{\ell/2}$. ∎

Lemma 4. *Let $m_1, m_2 \subseteq \mathcal{U}_a \times \mathcal{U}_b$ be bijections and let $d = n - |m_1 \cap m_2|$. Then*

$$B_{m_1,m_2}(z,z) \le b^{\circ}_2(z,z)^{d/2}.$$

Proof: From (2), $b^{\circ}_1(z,z) = 1$. Then the claim follows from Lemmas 2 and 3. ∎

V. Achievability

Proof of Theorem 1: We will use a union bound over all possible errors.

$$\begin{aligned}
&\Pr\left[\bigcup_{m_2\ne m_1} \mathcal{E}_{m_2,m_1}\middle| M = m_1\right] \\
&\le \sum_{m_2\ne m_1} \Pr[\mathcal{E}_{m_2,m_1}|M = m_1] \\
&= \sum_{d=2}^{n} \sum_{m_2\in S_{m_1,d}} \Pr[\mathcal{E}_{m_2,m_1}|M = m_1]
\end{aligned}$$

where $S_{m,d}$ is the set of matchings that differ from m is exactly d places. We have

$$|S_{m,d}| \le \binom{n}{d} d! \le n^d.$$

From Lemma 1 and Lemma 4, we have

$$\begin{aligned}
\Pr[\mathcal{E}_{m_2,m_1}|M = m_1] &\le \prod_{\ell} b^{\circ}_{\ell}(z,z)^{t^{\circ}_{\ell}} \\
&\le \prod_{\ell} (b^{\circ}_2(z,z)^{\ell/2})^{t^{\circ}_{\ell}} \\
&= b^{\circ}_2(z,z)^{d/2}.
\end{aligned}$$

Thus the overall probability of error is at most

$$\sum_{d=2}^{n} n^d b^{\circ}_2(z,z)^{d/2}.$$

From the main condition of the theorem, we have

$$\begin{aligned}
I^{\circ}_2(p) &\ge 2\log n + \omega(1) \\
b^{\circ}_2(z,z) &\le \exp(-2\log n - \omega(1)) \\
&= o(n^{-2}),
\end{aligned}$$

so for sufficiently large n, $nb^{\circ}_2(z,z)^{1/2} < 1$ and we have

$$\sum_{d=2}^{n} n^d b^{\circ}_2(z,z)^{d/2} \le \frac{n^2 b^{\circ}_2(z,z)}{1 - nb^{\circ}_2(z,z)^{1/2}} \le o(1)$$

which proves the claim. ∎

VI. Converse

Lemma 5. *For any two bijections $m_1, m_2 \subseteq \mathcal{U}_a \times \mathcal{U}_b$,*

$$B_{m_1,m_2}(x,y) = B_{m_2,m_1}(x,y).$$

Proof: For each ℓ, $b^\circ_\ell(x,y) = b^\circ_\ell(y,x)$. The permutations $m_2^T \circ m_1$ and $m_1^T \circ m_2$ are inverses and thus have the same cycle decomposition. The claim follows from Lemma 2. ∎

Lemma 6. *Fix alphabets $\mathcal{Y}_a$ and $\mathcal{Y}_b$ and a joint distribution $q \in \mathbb{P}(\mathcal{Y}_a \times \mathcal{Y}_b)$. Let ℓ depend on n such that $\ell = \omega(1)$. Let $\mathcal{X}_a = \mathcal{Y}_a^\ell$, $\mathcal{X}_b = \mathcal{Y}_b^\ell$, $p = q^{\otimes \ell}$. For any two bijections $m_1, m_2 \subseteq \mathcal{U}_a \times \mathcal{U}_b$ such that $|m_1 \cap m_2| = n-2$,*

$$\Pr[\mathcal{E}_{m_2,m_1} | M = m_1] \geq b^\circ_2(z,z)^{(1+o(1))}.$$

Proof: The function $c(\theta) = B_{m_1,m_2}(p^{\odot\theta}, p^{\odot(1-\theta)})$ is a conditional moment generating function:

$$c(\theta) = \mathbb{E}\left[\exp\left(\theta \log\left(\frac{r(f_a,f_b;m_2)}{r(f_a,f_b;m_1)}\right)\right)\middle| M = m_1\right].$$

From Lemma 2, we have

$$\begin{aligned} &B_{m_1,m_2}(p^{\odot\theta}, p^{\odot(1-\theta)}) \\ &= b^\circ_1(p^{\odot\theta}, p^{\odot(1-\theta)})^{n-2} b^\circ_2(p^{\odot\theta}, p^{\odot(1-\theta)}) \\ &= b^\circ_2(p^{\odot\theta}, p^{\odot(1-\theta)}). \end{aligned}$$

because

$$b^\circ_1(p^{\odot\theta}, p^{\odot(1-\theta)}) = \mathrm{tr}((p^{\odot\theta})(p^{\odot(1-\theta)})^T) = \sum_{i,j} p_{i,j}^\theta p_{i,j}^{1-\theta} = 1.$$

By Lemma 5

$$c(\theta) = b^\circ_2(p^{\odot\theta}, p^{\odot(1-\theta)}) = b^\circ_2(p^{\odot(1-\theta)}, p^{\odot\theta}) = c(1-\theta).$$

Moment generating functions are log-convex, so $c(\theta)$ is minimized at $\theta = \frac{1}{2}$.

Because $p = q^{\otimes\ell}$, $c(\theta)$ is the product of ℓ identical terms. Let $u = q^{\odot\theta}$ and $v = q^{\odot(1-\theta)}$.

$$\begin{aligned} b^\circ_2(p^{\odot\theta}, p^{\odot(1-\theta)}) &= b^\circ_2(u^{\otimes\ell}, v^{\otimes\ell}) \\ &= \mathrm{tr}((u^{\otimes\ell})(v^{\otimes\ell})^T (u^{\otimes\ell})(v^{\otimes\ell})^T) \\ &= \mathrm{tr}(uv^T uv^T)^\ell \\ &= b^\circ_2(u,v)^\ell \\ &= b^\circ_2(q^{\odot\theta}, q^{\odot(1-\theta)})^\ell \end{aligned}$$

By Cramér's Theorem on the asymptotic tightness of the Chernoff bound [11]

$$\begin{aligned} &\Pr\left[\log\left(\frac{r(f_a,f_b;m_2)}{r(f_a,f_b;m_1)}\right) \geq 0 \middle| M = m_1\right] \\ &\geq b^\circ_2(q^{\odot\frac{1}{2}}, q^{\odot\frac{1}{2}})^{\ell(1-o_\ell(1))} \\ &= b^\circ_2(p^{\odot\frac{1}{2}}, p^{\odot\frac{1}{2}})^{1-o(1)}. \end{aligned}$$

Because $\ell = \omega(1)$, $o_\ell(1)$ and $o(1)$ are equivalent. ∎

Lemma 7. *For any three bijections $m_1, m_2, m_3 \subseteq \mathcal{U}_a \times \mathcal{U}_b$,*

$$\Pr[\mathcal{E}_{m_2,m_1} \cap \mathcal{E}_{m_3,m_1} | M = m_1] \leq b^\circ_2(z,z)^{d/2}$$

where $d = n - |m_2 \cap m_3|$.

Proof: For $\theta \geq 0$ and $\theta' \geq 0$,

$$\begin{aligned} &\Pr\left[\frac{r(f_a,f_b;m_2)}{r(f_a,f_b;m_1)} \geq 1 \wedge \frac{r(f_a,f_b;m_3)}{r(f_a,f_b;m_1)} \geq 1 \middle| M = m_1\right] \\ &= \mathbb{E}[\mathbf{1}(\mathcal{E}_{m_3,m_1})\mathbf{1}(\mathcal{E}_{m_2,m_1}) | M = m_1] \\ &\leq \mathbb{E}\left[\left(\frac{r(f_a,f_b;m_2)}{r(f_a,f_b;m_1)}\right)^\theta \left(\frac{r(f_a,f_b;m_3)}{r(f_a,f_b;m_1)}\right)^{\theta'} \middle| M = m_1\right] \\ &= \sum_{f_a,f_b} \left(\frac{r(f_a,f_b;m_2)}{r(f_a,f_b;m_1)}\right)^\theta \left(\frac{r(f_a,f_b;m_3)}{r(f_a,f_b;m_1)}\right)^{\theta'} r(f_a,f_b;m_1) \\ &= \sum_{f_a,f_b} r(f_a,f_b;m_2)^\theta r(f_a,f_b;m_3)^{\theta'} r(f_a,f_b;m_1)^{1-\theta-\theta'} \end{aligned}$$

Choosing $\theta = \theta' = \frac{1}{2}$, we obtain

$$\begin{aligned} &\mathbb{E}[\mathbf{1}(\mathcal{E}_{m_3,m_1})\mathbf{1}(\mathcal{E}_{m_2,m_1}) | M = m_1] \\ &\leq \sum_{f_a,f_b} r(f_a,f_b;m_2)^{\frac{1}{2}} r(f_a,f_b;m_3)^{\frac{1}{2}} \\ &= B_{m_2,m_3}(z,z) \\ &\overset{(a)}{\leq} b^\circ_2(z,z)^{d/2} \end{aligned}$$

where (a) follows from Lemma 4. ∎

Proof of Theorem 2: Let m_1 be the matching used to generate the databases and let $S = S_{m_1,2}$ be the set of matchings of size n that differ from m_1 in exactly two places. That is, for all $m \in S$, $|m_1 \cap m| = n-2$. Observe that $|S| = \binom{n}{2}$, because each element of S can be specified by the two users in $\mathcal{U}_a$ that it matches differently than m_1 does. Let X be the number of error events that occur:

$$X = \sum_{m \in S} \mathbf{1}(\mathcal{E}_{m,m_1}).$$

Let $\epsilon_1 = \Pr[\mathcal{E}_{m,m_1} | M = m_1]$, i.e. the probability that a specific transposition error occurs.

We need a lower bound on the probability that $X > 0$. From Chebyshev's inequality, we have

$$\Pr\left[(X - \mathbb{E}[X])^2 \geq \mathbb{E}[X]^2\right] \leq \mathbb{E}\left[\frac{(X-\mathbb{E}[X])^2}{\mathbb{E}[X]^2}\right] = \frac{\mathbb{E}[X^2]}{\mathbb{E}[X]^2} - 1$$

and we need to find conditions that make this $o(1)$. We have

$$\begin{aligned} X^2 &= \sum_{(m_2,m_3) \in S^2} \mathbf{1}(\mathcal{E}_{m_2,m_1})\mathbf{1}(\mathcal{E}_{m_3,m_1}) \\ &= \sum_{m_2 \in S} \mathbf{1}(\mathcal{E}_{m_2,m_1}) + 2 \sum_{\{m_2,m_3\} \in \binom{S}{2}} \mathbf{1}(\mathcal{E}_{m_2,m_1})\mathbf{1}(\mathcal{E}_{m_3,m_1}) \end{aligned}$$

For a set $\{m_2, m_3\} \in \binom{S}{2}$, either $|m_2 \cap m_3| = n-3$ or $|m_2 \cap m_3| = n-4$. There are $3\binom{n}{3}$ pairs of the former type and $3\binom{n}{4}$ pairs of the latter type. In the latter case, the indicator variables $\mathcal{E}_{m_2,m_1}$ and $\mathcal{E}_{m_3,m_1}$ are independent. In the former case, let $\epsilon_2 = \Pr[\mathcal{E}_{m_2,m_1} \cap \mathcal{E}_{m_3,m_1} | M = m_1]$.

Now we compute

$$\mathbb{E}[X]^2 = \binom{n}{2}^2 \epsilon_1^2 = \left(\binom{n}{2} + 6\binom{n}{3} + 6\binom{n}{4}\right)\epsilon_1^2$$

and

$$\mathbb{E}[X^2] = \binom{n}{2}\epsilon_1 + 6\binom{n}{3}\epsilon_2 + 6\binom{n}{4}\epsilon_1^2$$
$$\frac{\mathbb{E}[X^2] - \mathbb{E}[X]^2}{\mathbb{E}[X]^2} = \frac{\binom{n}{2}(\epsilon_1 - \epsilon_1^2) + 6\binom{n}{3}(\epsilon_2 - \epsilon_1^2)}{\binom{n}{2}^2\epsilon_1^2}$$
$$\leq \mathcal{O}\left(\frac{1}{n^2\epsilon_1} + \frac{\epsilon_2}{n\epsilon_1^2}\right).$$

From Lemma 7 we have $\epsilon_2 \leq (b_2^\circ(z,z))^{\frac{3}{2}}$ and from Lemma 6 we have $\epsilon_1 \geq (b_2^\circ(z,z))^{1+o(1)}$, so

$$\Pr[X=0] \leq \mathcal{O}\left(\frac{1}{n^2(b_2^\circ(z,z))^{1+o(1)}} + \frac{1}{n(b_2^\circ(z,z))^{\frac{1}{2}+o(1)}}\right).$$

If $b_2^\circ(z,z) \geq n^{-2+\Omega(1)}$, then

$$n^2 b_2^\circ(z,z)^{1+o(1)} \geq n^{2+(1+o(1))(-2+\Omega(1))} \geq n^{\Omega(1)} \geq \omega(1)$$

and $\Pr[X=0] \leq o(1)$. ∎

VII. Properties of cycle mutual information

Consider a joint distribution $p \in \mathbb{P}(\mathcal{X}_a \times \mathcal{X}_b)$ and recall the definitions of z and σ from Section II. The properties of $\sigma^{\odot 2}$ reflect the correlation in the distribution p. The following three conditions are equivalent: $\sigma^{\odot 2}$ is supported on one point, the rank of the matrix z is one, and the p is the product of distributions on $\mathcal{X}_a$ and $\mathcal{X}_b$.

$I_\ell^\circ(p)$ shares several properties with mutual information. It is symmetric: $I_\ell^\circ(p) = I_\ell^\circ(p^T)$. It tensorizes: $I_\ell^\circ(p^{\otimes k}) = kI_\ell^\circ(p)$. It reduces to entropy in the case of identical random variables: if $\mathcal{X}_a = \mathcal{X}_b$ and $p = \operatorname{diag}(p')$, then

$$I_\ell^\circ(\operatorname{diag}(p')) = H_\ell(p').$$

because $\sigma^{\odot 2}$ is a rearrangement of p'. In general, we have

$$I_\ell^\circ(p) \leq \min(H_\ell(p_a), H_\ell(p_b)).$$

Something stronger is true: the distribution $\sigma^{\odot 2}$ majorizes p_a and p_b. The diagonal of zz^T is the marginal distributions p_a:

$$(zz^T)_{i,i} = \sum_j z_{i,j}^2 = \sum_j p_{i,j}.$$

Furthermore,

$$(zz^T)_{i,i} = (U\Sigma V^T V \Sigma U^T)_{i,i} = \sum_k U_{i,k}^2 \sigma_k^2.$$

Because U is an orthogonal matrix, the Hadamard product $U \odot U$ is doubly stochastic. Thus $\sigma^{\odot 2}$ majorizes p_a. The diagonal of $z^T z$ contains p_b, which is also majorized by $\sigma^{\odot 2}$.

A. Data processing inequality

Lemma 8. *Let $p \in \mathbb{P}(\mathcal{X})$, let $q \in \mathcal{X} \to \mathbb{P}(\mathcal{Y})$, and let $r \in \mathcal{Y} \to \mathbb{P}(\mathcal{Z})$, so $\operatorname{diag}(p) \in \mathbb{P}(\mathcal{X} \times \mathcal{X})$, $\operatorname{diag}(p)q \in \mathbb{P}(\mathcal{X} \times \mathcal{Y})$, and $\operatorname{diag}(p)qr \in \mathbb{P}(\mathcal{X} \times \mathcal{Z})$. Then for integer $\ell \geq 2$, $I_\ell^\circ(\operatorname{diag}(p)q) \geq I_\ell^\circ(\operatorname{diag}(p)qr)$.*

Proof: Define the matrices $z_{i,k} = \sqrt{(\operatorname{diag}(p)q)_{i,k}}$ and $w_{i,l} = \sqrt{(\operatorname{diag}(p)qr)_{i,l}}$. Then

$$(zz^T)_{i,i} = (ww^T)_{i,i} = p_i$$

We have

$$(zz^T)_{i,j} = \sqrt{p_i p_j} \sum_{k \in \mathcal{Y}} \sqrt{q_{i,k} q_{j,k}}.$$

The sum is the Bhattacharyya coefficient of the distributions $q_{i,\cdot}$ and $q_{j,\cdot}$, which can be written in terms of the Bhattacharyya divergence as follows: $\exp\left(-\frac{1}{2}D_{\frac{1}{2}}(q_{i,\cdot}||q_{j,\cdot})\right)$. Similarly

$$(ww^T)_{i,j} = \sqrt{p_i p_j} \sum_{l \in \mathcal{Z}} \sqrt{(qr)_{i,l}(qr)_{j,l}}.$$

By the data processing inequality for Rényi divergences [12], we have

$$D_{\frac{1}{2}}(q_{i,\cdot}||q_{j,\cdot}) \geq D_{\frac{1}{2}}((qr)_{i,\cdot}||(qr)_{j,\cdot}).$$

Thus

$$(zz^T)_{i,j} \leq (ww^T)_{i,j}$$
$$\operatorname{tr}((zz^T)^\ell) \leq \operatorname{tr}((ww^T)^\ell)$$
$$I_\ell^\circ(\operatorname{diag}(p)q) \geq I_\ell^\circ(\operatorname{diag}(p)qr)$$

for all integer $\ell \geq 2$. ∎

Acknowledgement

This work was supported in part by NSF grants CCF 16-19216, CCF 16-17286, and CNS 15-53437.

References

[1] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *IEEE Symposium on Security and Privacy*, 2008, pp. 111–125.

[2] C. Dwork, "Differential privacy: A survey of results," in *International Conference on Theory and Applications of Models of Computation*. Springer, 2008, pp. 1–19.

[3] C. Huang, P. Kairouz, X. Chen, L. Sankar, and R. Rajagopal, "Context-aware generative adversarial privacy," *Entropy*, vol. 19, no. 12, p. 656, 2017.

[4] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Matching Anonymized and Obfuscated Time Series to Users' Profiles," *arXiv:1710.00197 [cs, math]*, Sep. 2017.

[5] D. Cullina and N. Kiyavash, "Exact alignment recovery for correlated erdos-rényi graphs," 2017.

[6] ——, "Improved achievability and converse bounds for Erdos-Rnyi graph matching," in *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*. ACM, 2016, pp. 63–72.

[7] P. Pedarsani and M. Grossglauser, "On the privacy of anonymized networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1235–1243.

[8] S. Ji, W. Li, N. Z. Gong, P. Mittal, and R. A. Beyah, "On your social network de-anonymizablity: Quantification and large scale evaluation with seed knowledge." in *The Network and Distributed System Security Symposium (NDSS)*, 2015.

[9] O. E. Dai, D. Cullina, N. Kiyavash, and M. Grossglauser, "On the Performance of a Canonical Labeling for Matching Correlated Erd\H{o}s-R\'enyi Graphs," *arXiv:1804.09758 [cs, stat]*, Apr. 2018, arXiv: 1804.09758.

[10] J. Edmonds and R. M. Karp, "Theoretical Improvements in Algorithmic Efficiency for Network Flow Problems," *J. ACM*, vol. 19, no. 2, pp. 248–264, Apr. 1972. [Online]. Available: http://doi.acm.org/10.1145/321694.321699

[11] B. Hajek, *Random processes for engineers*. Cambridge university press, 2015.

[12] T. Van Erven and P. Harremos, "Rényi divergence and Kullback-Leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.