# Database Alignment with Gaussian Features

**Osman Emre Dai**
Georgia Institute of Technology

**Daniel Cullina**
Princeton University

**Negar Kiyavash**
Georgia Institute of Technology

## Abstract

We consider the problem of aligning a pair of databases with jointly Gaussian features. We consider two algorithms, complete database alignment via MAP estimation among all possible database alignments, and partial alignment via a thresholding approach of log likelihood ratios. We derive conditions on mutual information between feature pairs, identifying the regimes where the algorithms are guaranteed to perform reliably and those where they cannot be expected to succeed.

## 1 Introduction

Consider the following setting: There are a large set of entities (e,g, users) with some measurable characteristics. Let the measures of these characteristics be jointly Gaussian, with known statistics. We refer to these measures as features. Consider two different sources, each providing a database with lists of features for these entities. Furthermore, let one these sources lack proper labeling for features that would allow for the identification of feature pairs from the two sources that correspond to the same entity. This might be due to privacy concerns, if the mentioned features provided by the source contain sensitive information that ought to remain anonymous, or it might simply be that a reliable labeling is not available.

If the correlation between features pairs is sufficiently strong, then it is possible to exploit this correlation to identify correspondences between the two databases and in fact generate a perfect alignment between the feature lists. Such a capability might be a valuable tool to recuperate missing information by labeling unlabeled features or by allowing the junction of measurements coming from distinct sources. However it also has serious implications in privacy as it makes anonymous data vulnerable to deanonymization attacks [1].

It then becomes critical to understand the limitations of database alignment and to identify the conditions that characterize these limitations. This allows us to assess the feasibility and reliability of alignment procedures as well as the vulnerability of deanonymization schemes. In this study we investigate the conditions that guarantee either the achievability of alignment or its infeasibility. We analyze these conditions for both partial alignments and as well as for complete alignments. Cullina et al. have recently analyzed this problem for the case of discrete random variables, introducing a new correlation measure characterizing the feasibility of alignment [2]. Takbiri et al. have investigated a related problem where the feature of each user is Gaussian with characteristic statistics and has correlation with other user features [3]. In this setting an adversary with perfect knowledge of system statistics attempts to match features with the characteristic user statistics. This follows the authors' previous studies of the same setting for discrete valued features and with data obfuscation [4],[5].

The database alignment problem is connected to the widely studied graph alignment problem. In that setting, each feature is associated with a pair of anonymized users. In the simplest case, the feature is a Bernoulli random variable indicating the presence or absence of an edge between the users. A recent line of work has characterized the information theoretic limits of the graph alignment problem [6, 7, 8]. The problem of aligning correlated Wigner matrices, in which each feature is a Gaussian random variable, has served as a proxy for understanding the effectiveness of graph alignment algorithms [9].

## 2 Model

**Notation** We denote random variables by capital letters and fixed values by lowercase letters. For a set $\mathcal{S}$ and finite sets $\mathcal{T}$,$\mathcal{U}$, we denote by $\mathcal{S}^{\mathcal{T}\times\mathcal{U}}$ the set of matrices with entries in $\mathcal{S}$, rows indexed by $\mathcal{T}$ and columns indexed by $\mathcal{U}$. We mark vectors with arrows

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. 

and write matrices in boldface. Given some matrix $\mathbf{z}$, we denote its $i$-th row by $\mathbf{z}_{i*}$, $j$-th column by $\mathbf{z}_{*j}$ and its $(i,j)$-th entry by $\mathbf{z}_{ij}$. We denote the identity matrix with rows and columns indexed by $\mathcal{T}$ by $\mathbf{I}^{\mathcal{T}}$. When the indexing set is clear from context, we will drop the superscript. We denote the set of integers from 1 to $n$ by $[n]$.

### 2.1 General problem formulation

In this model, a database is just a function from a set of users to some space. The value of the function for a user $u$ is the database entry for that user. Cullina, Mittal, and Kiyavash considered database entries in finite alphabets[2]. In this paper, we consider database entries that are finite dimensional real vectors sampled from a gaussian distribution.

We are given two sets of user identifiers, $\mathcal{U}$ and $\mathcal{V}$, with $|\mathcal{U}| = |\mathcal{V}| = n$. We express the content of databases by matrices $\mathbf{A} \in \mathbb{R}^{\mathcal{U}\times[d_a]}$ and $\mathbf{B} \in \mathbb{R}^{\mathcal{V}\times[d_b]}$, so $d_a$ and $d_b$ are the lengths of feature vectors.

There exists a natural bijective correspondence between the identifier sets, i.e. each identifier in one set is related to exactly one identifier in the other set. We express this correspondence by the bijective matching $M \subseteq \mathcal{U} \times \mathcal{V}$.

Let $p_{\vec{X}\vec{Y}}$ be the density of jointly gaussian random variables $\vec{X} \in \mathbb{R}^{d_a}$ and $\vec{Y} \in \mathbb{R}^{d_b}$ such that $(\vec{X}, \vec{Y}) \sim \mathcal{N}(\vec{\mu}, \mathbf{\Sigma})$.

The density $p_{\mathbf{AB}|M}$ is defined as follows. For each $(u,v) \in M$, $(\mathbf{A}_{u*}, \mathbf{B}_{v*}) \sim p_{\vec{X}\vec{Y}}$ and these $n$ random variables are independent:

$$p_{\mathbf{AB}|M}(\mathbf{a}, \mathbf{b}|m) = \prod_{(u,v)\in m} p_{\vec{X}\vec{Y}}(\mathbf{a}_{u*}, \mathbf{b}_{v*}).$$

The matching $M$ is uniformly distributed over the $n!$ bijective matchings between $\mathcal{U}$ and $\mathcal{V}$.

The database alignment problem is to recover $M$ from $\mathbf{AB}$, given knowledge of $p_{\vec{X}\vec{Y}}$.

Observe that the rows of $\mathbf{A}$ are i.i.d. and that $\mathbf{A}$ is independent of $M$. The same is true for $\mathbf{B}$. In other words, by examining one database, an observer learns nothing about $M$.

A pair of databases are illustrated in Figure 1.

**Canonical form of covariance** We write $\vec{\mu} = \begin{bmatrix} \vec{\mu}_a \\ \vec{\mu}_b \end{bmatrix}$ and $\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{\mathrm{a}} & \mathbf{\Sigma}_{\mathrm{ab}} \\ \mathbf{\Sigma}_{\mathrm{ab}}^\top & \mathbf{\Sigma}_{\mathrm{b}} \end{bmatrix}$, so $\mathbf{A}_{u*} \sim \mathcal{N}(\vec{\mu}_a, \mathbf{\Sigma}_{\mathrm{a}})$ for each $u \in \mathcal{U}$ and $\mathbf{B}_{u*} \sim \mathcal{N}(\vec{\mu}_b, \mathbf{\Sigma}_{\mathrm{b}})$ for each $v \in \mathcal{V}$.

Let $d'_a$ be the dimension of the support of $\vec{X}$, i.e. the rank of $\mathbf{\Sigma}_{\mathrm{a}}$. Let $\phi : \mathbb{R}^{[d_a]} \to \mathbb{R}^{d'_a}$ be an affine transformation that is injective on the support of $\vec{X}$. If we apply $\phi$ to each row of $\mathbf{A}$, which can be done without knowledge of $M$, we obtain an equivalent database alignment problem. Similarly, the database $\mathbf{B}$ can be transformed to obtain an equivalent problem.

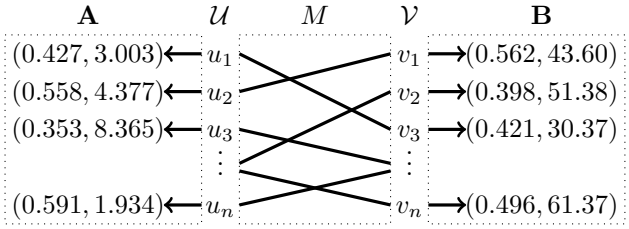


Figure 1: Databases $\mathbf{A}$ and $\mathbf{B}$ with $d_a = d_b = 2$ and a matching $M$ between their user identifier sets.

For any gaussian database alignment problem, there is an equivalent problem with $\vec{\mu} = \vec{0}$ and

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{I}^{[d]} & \mathrm{diag}(\vec{\rho}) \\ \mathrm{diag}(\vec{\rho}) & \mathbf{I}^{[d]} \end{bmatrix} = \bigoplus_{i\in[d]} \begin{bmatrix} 1 & \rho_i \\ \rho_i & 1 \end{bmatrix}$$

where $d = \min(d_a, d_b)$. Thus the correlation structure of $(\vec{X}, \vec{Y})$ is completely summarized by the vector $\vec{\rho} \in \mathbb{R}^d$. The explicit transformations that put $\mathbf{\Sigma}$ into this form are described in our supplementary material.

### 2.2 Correlation measures

Let $I_{\vec{X}\vec{Y}} \triangleq I(\mathbf{A}_{u*}, \mathbf{B}_{v*}, |(u,v) \in M)$ denote the mutual information between any pair of related identifiers coming from $(u,v) \in M$. Then

$$\begin{aligned} I_{\vec{X}\vec{Y}} &= -\frac{1}{2} \log \frac{\det\left(\mathbf{\Sigma}\right)}{\det\left(\mathbf{\Sigma}_{\mathrm{a}}\right) \cdot \det\left(\mathbf{\Sigma}_{\mathrm{b}}\right)} \\ &= -\frac{1}{2} \sum_{i\in[d]} \log\left(1 - \rho_i^2\right). \end{aligned}$$

Under the canonical formulation where $\mathbf{\Sigma}_{\mathrm{a}} = \mathbf{\Sigma}_{\mathrm{b}} = \mathbf{I}^d$ and $\mathbf{\Sigma}_{\mathrm{ab}} = \mathrm{diag}(\vec{\rho})$ this becomes

Given any $(u,v) \in M$ and $(\vec{X}, \vec{Y}) = (\mathbf{A}_{u*}^\top, \mathbf{B}_{v*}^\top)$,

$$\sigma^2_{\vec{X}\vec{Y}} \triangleq \mathrm{Var}\left(\log \frac{p_{\vec{X}\vec{Y}}(\vec{X}, \vec{Y})}{p_{\vec{X}}(\vec{X}) p_{\vec{Y}}(\vec{Y})}\right).$$

Then $\sigma^2_{\vec{X}\vec{Y}} = \mathrm{tr}\left(\mathbf{\Sigma}_{\mathrm{a}}^{-1}\mathbf{\Sigma}_{\mathrm{ab}}\mathbf{\Sigma}_{\mathrm{b}}^{-1}\mathbf{\Sigma}_{\mathrm{ab}}^\top\right)$. Furthermore under the canonical formulation where $\mathbf{\Sigma}_{\mathrm{a}} = \mathbf{\Sigma}_{\mathrm{b}} = \mathbf{I}^d$ and $\mathbf{\Sigma}_{\mathrm{ab}} = \mathrm{diag}(\vec{\rho})$ this simplifies to $\sigma^2_{\vec{X}\vec{Y}} = \sum \rho_i^2$. These calculations are made explicit in our supplementary material.

Note that $\sigma_{\vec{X}\vec{Y}}$ is upper bounded by $\sqrt{2I_{\vec{X}\vec{Y}}}$. This can easily be seen in the canonical formulation, where $\sigma^2_{\vec{X}\vec{Y}} = \sum \rho_i^2 \leq -\sum \log(1-\rho_i^2) = 2I_{\vec{X}\vec{Y}}$.

## 3 Results

Our results identify conditions on $I_{\vec{X}\vec{Y}}$ and $\sigma_{\vec{X}\vec{Y}}$, as defined in Section 2.2.

**MAP estimation** The algorithm considers all possible alignments between the two sets and chooses the most likely one. The log likelihood of an alignment is, by the independence of correlated feature pairs, equal to the sum of the log likelihood of each aligned feature pair. MAP estimation can then be implemented by computing the joint likelihood for each feature pair in $\mathcal{O}(n^2 d)$-time and computing the maximum weight matching between databases in $\mathcal{O}(n^3)$-time using the Hungarian algorithm.

**Theorem 1. (Achievability)** *If mutual information between feature pairs $I_{\vec{X}\vec{Y}} \geq 2\log n + \omega(1)$, then the MAP estimator returns the proper alignment with probability $1 - o(1)$.*

**Theorem 2. (Converse)** *Let $d \in \mathbb{N}$ such that $d \geq \omega(1)$. Furthermore let $\mathbf{\Sigma}_a = \mathbf{\Sigma}_b = \mathbf{I}^d$ and $\mathbf{\Sigma}_{ab} = \rho\mathbf{I}$. If $I_{\vec{X}\vec{Y}} \leq 2\log n(1-\Omega(1))$, then any for algorithm, the probability of returning the proper alignment is $o(1)$.*

**Binary hypothesis testing** The algorithm checks every possible pair of identifiers and uses a threshold-based method to decide whether to match the pair or not. This can be done in $\mathcal{O}(n^2 d)$-time, which is the complexity of computing joint likelihoods for each feature pair.

**Theorem 3. (Achievability)** *If*

$$I_{\vec{X}\vec{Y}} \geq \sigma_{\vec{X}\vec{Y}} \cdot \sqrt{\frac{n}{\varepsilon_{FN}}} + \log \frac{n^2}{\varepsilon_{FP}},$$

*then, choosing the threshold such that $\log(n^2/\varepsilon_{FP}) \leq \tau \leq I_{\vec{X}\vec{Y}} - \sigma_{\vec{X}\vec{Y}}\sqrt{n/\varepsilon_{FN}}$, the binary hypothesis test gives no more than $\varepsilon_{FN}$ false negatives and $\varepsilon_{FP}$ false positives in expectation.*

It follows that the following regimes are achievable:

- $I_{\vec{X}\vec{Y}} \geq \log(n) + \omega(1) \quad \varepsilon_{FN} \leq o(n) \quad \varepsilon_{FP} \leq o(n)$
- $I_{\vec{X}\vec{Y}} \geq 2\log(n) + \omega(1) \quad \varepsilon_{FN} \leq o(n) \quad \varepsilon_{FP} \leq o(1)$

The next theorem holds for databases with any distribution of feature pairs, i.e. not only Gaussians.

**Theorem 4. (Converse)** *For any binary hypothesis test, the expected number of false negatives $\varepsilon_{FN}$ and false positives $\varepsilon_{FP}$ is lower bounded as*

$$\varepsilon_{FN} + \varepsilon_{FP} \geq \frac{n}{2}\left(1 - \frac{I_{\vec{X}\vec{Y}}}{\log n}\right)\left(1 - \mathcal{O}\left(\frac{1}{\log n}\right)\right).$$

It follows that, if $I_{\vec{X}\vec{Y}} \leq \log n\big(1 - \Omega(1)\big)$, then any binary hypothesis test has expected number of errors $\varepsilon_{FN} + \varepsilon_{FP} \geq \Omega(n)$.

## 4 MAP estimation

**Matching algorithm** The maximum a posteriori estimator is the optimal estimator for the exact matching $M$ given $F$. Given some realization $\mathbf{f} = (\mathbf{a}, \mathbf{b})$,

$$\begin{aligned}\hat{m}(\mathbf{f}) &= \operatorname*{argmax}_m \ \Pr\left[M = m | \mathbf{F} = \mathbf{f}\right] \\ &= \operatorname*{argmax}_m \ \frac{p_{\mathbf{F}|M}(\mathbf{f}|m) P_M(m)}{p_{\mathbf{F}}(\mathbf{f})} \\ &\overset{(a)}{=} \operatorname*{argmax}_m \ p_{\mathbf{F}|M}(\mathbf{f}|m)\end{aligned}$$

where $(a)$ follows from the fact that $M$ has a uniform distribution.

### 4.1 Achievability analysis

We establish a sufficient condition on the mutual information $I_{XY}$ between feature pairs to achieve a perfect alignment. The rest of this section assumes the cannonical setting. However, by the equivalence between the general setting and the canonical setting (as shown in our supplementary material), the result directly applies to the general setting.

Our analysis goes as follows: Lemma 4.1 sets an upper bound on the error probability that a given matching is more likely than the actual one. This bound is in the form of a function $R$ whose explicit value remains to be determined. Lemma 4.2 gives an expression of $R$ that has a decomposition with terms corresponding to each cycle of 'mismatchings'. Finally Lemma 4.3 gives the explicit expression for each of these cycle-terms and Lemma 4.4 bounds their product by a function whose value only depends on the number of mismatchings. Joining these results gives us the achievability condition in Theorem 1.

**Definition 4.1.** *Given any pair of bijective matchings $m_1, m_2 \subseteq \mathcal{U} \times \mathcal{V}$, define the event*

$$\mathcal{E}(m_1, m_2) = \left\{\mathbf{f} : p_{\mathbf{F}|M}(\mathbf{f}|m_1) \leq p_{\mathbf{F}|M}(\mathbf{f}|m_2)\right\}.$$

Notice that given matching $m = M$, the MAP estimator fails if and only if there exists some matching $m' \neq m$ such that $\mathbf{F} \in \mathcal{E}(m, m')$.

**Definition 4.2.** *Given any pair of bijective matchings $m_1, m_2 \subseteq \mathcal{U} \times \mathcal{V}$, define the function*

$$R(m_1, m_2) \triangleq \int \sqrt{p_{\mathbf{F}|M}(\mathbf{f}|m_1) p_{\mathbf{F}|M}(\mathbf{f}|m_2)}\, d\mathbf{f}$$

*where the integral is over the whole space $\mathbb{R}^{(\mathcal{U}\sqcup\mathcal{V})\times[d]}$.*

**Lemma 4.1.** *For any pair of bijective matchings $m_1, m_2 \subseteq \mathcal{U} \times \mathcal{V}$*

$$\Pr\left[\mathbf{F} \in \mathcal{E}(m_1, m_2) | M = m_1\right] \leq R(m_1, m_2)$$

*Proof.* For any $\theta \geq 0$

$$\begin{aligned}
&\Pr[\mathbf{F} \in \mathcal{E}(m_1, m_2) | M = m_1] \\
&\quad = \mathbb{E}\left[\mathbb{1}\left\{\frac{p_{\mathbf{F}|M}(\mathbf{f}|m_2)}{p_{\mathbf{F}|M}(\mathbf{f}|m_1)} \geq 1\right\} \middle| M = m_1\right] \\
&\quad \leq \int \left(\frac{p_{\mathbf{F}|M}(\mathbf{f}|m_2)}{p_{\mathbf{F}|M}(\mathbf{f}|m_1)}\right)^{\theta} p_{\mathbf{F}|M}(\mathbf{f}|m_1)\, d\mathbf{f} \\
&\quad = \int (p_{\mathbf{F}|M}(\mathbf{f}|m_2))^{\theta} (p_{\mathbf{F}|M}(\mathbf{f}|m_1))^{1-\theta}\, d\mathbf{f}
\end{aligned}$$

Selecting $\theta = 1/2$ gives the claim. □

**Definition 4.3.** *Define shifted identity matrices $\mathbf{I}^{(k,+)}$ and $\mathbf{I}^{(k,-)}$ of size $k$ as*

$$\begin{aligned}
\mathbf{I}^{(k,+)}_{i,j} &= \mathbb{1}\{j - i = 1 \mod k\} \\
\mathbf{I}^{(k,-)}_{i,j} &= \mathbb{1}\{j - i = -1 \mod k\}.
\end{aligned}$$

*We simply write $\mathbf{I}^{(+)}$ and $\mathbf{I}^{(-)}$ when there is no need to specify the size of the matrix.*

*For any $\ell \in \mathbb{N}^+$,*

$$\mathbf{L}^{\ell}(s,t) \triangleq s\mathbf{I}^{\ell} - \frac{t}{2}\left(\mathbf{I}^{(\ell,+)} + \mathbf{I}^{(\ell,-)}\right),$$

*where $s, t \in \mathbb{R}$.*

**Lemma 4.2.** *Suppose $d_a = d_b = 1$ and $\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. For bijective matchings $m_1, m_2 \subseteq \mathcal{U} \times \mathcal{V}$,*

$$R(m_1, m_2) = \left(1 - \rho^2\right)^{\frac{n}{2}} \prod_{\ell} \left[\det \mathbf{L}^{\ell}\left(1 - \frac{\rho^2}{2}, \frac{\rho^2}{2}\right)\right]^{-\frac{k_\ell}{2}}$$

*where $k_\ell$ is the number of cycles of length $\ell$ of permutation $m_1 \circ m_2^{\top} \subseteq \mathcal{U} \times \mathcal{U}$.*

*Proof.* For a matching $m \subseteq \mathcal{U} \times \mathcal{V}$, let $\mathbf{m} \in \{0,1\}^{\mathcal{U} \times \mathcal{V}}$ be the indicator matrix for $m$.

Because $d_a = d_b = 1$, we will treat the databases as vectors $\vec{A} \in \mathbb{R}^{\mathcal{U}}$ and $\vec{B} \in \mathbb{R}^{\mathcal{V}}$. Let $\vec{F} \in \mathbb{R}^{\mathcal{U} \sqcup \mathcal{V}}$ be the concatenation of $\vec{A}$ and $\vec{B}$. Observe that $\mathbf{\Sigma}^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$. Then we can write

$$\begin{aligned}
&p_{\vec{F}|M}((\vec{a}, \vec{b})|m) = \frac{1}{\left(2\pi\sqrt{1-\rho^2}\right)^n} \cdot \\
&\exp\left(-\frac{1}{2(1-\rho^2)}\begin{bmatrix}\vec{a} \\ \vec{b}\end{bmatrix}^{\top}\begin{bmatrix}\mathbf{I}^{\mathcal{U}} & -\rho\mathbf{m} \\ -\rho\mathbf{m}^{\top} & \mathbf{I}^{\mathcal{V}}\end{bmatrix}\begin{bmatrix}\vec{a} \\ \vec{b}\end{bmatrix}\right). \quad (1)
\end{aligned}$$

For compactness, call the matrix that appears in (1) $\mathbf{\Sigma}(m)$. This gives us

$$\begin{aligned}
&\left(p_{\vec{F}|M}(\vec{f}; m_1) p_{\vec{F}|M}(\vec{f}; m_2)\right)^{\frac{1}{2}} \\
&= \frac{1}{\left(2\pi\sqrt{1-\rho^2}\right)^n} \exp\left(-\frac{\vec{f}^{\top}\left[\mathbf{\Sigma}(m_1) + \mathbf{\Sigma}(m_2)\right]\vec{f}}{4(1-\rho^2)}\right).
\end{aligned}$$

We obtain $R(m_1, m_2)$ by integrating this expression over the whole space:

$$\begin{aligned}
R(m_1, m_2) &= \int \sqrt{p_{\vec{F}|M}(\vec{f}; m_1) p_{\vec{F}|M}(\vec{f}; m_2)}\, df \\
&= \left[\frac{\left(1-\rho^2\right)^n}{\det\left(\frac{1}{2}\mathbf{\Sigma}(m_1) + \frac{1}{2}\mathbf{\Sigma}(m_2)\right)}\right]^{1/2}. \quad (2)
\end{aligned}$$

Observe that $\begin{bmatrix}\mathbf{I} & \mathbf{z} \\ \mathbf{z}^{\top} & \mathbf{I}\end{bmatrix} = \begin{bmatrix}\mathbf{I} & \mathbf{0} \\ \mathbf{z}^{\top} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{z} \\ \mathbf{0} & \mathbf{I} - \mathbf{z}^{\top}\mathbf{z}\end{bmatrix}$ for any matrix $\mathbf{z}$. Then $\det\begin{bmatrix}\mathbf{I} & \mathbf{z} \\ \mathbf{z}^{\top} & \mathbf{I}\end{bmatrix} = \det\left(\mathbf{I} - \mathbf{z}^{\top}\mathbf{z}\right)$. Using this relation we have

$$\begin{aligned}
&\det\left(\frac{1}{2}\mathbf{\Sigma}(m_1) + \frac{1}{2}\mathbf{\Sigma}(m_2)\right) \\
&= \det\begin{bmatrix}\mathbf{I} & -\frac{\rho}{2}\left(\mathbf{m}_1 + \mathbf{m}_2\right) \\ -\frac{\rho}{2}\left(\mathbf{m}_1 + \mathbf{m}_2\right)^{\top} & \mathbf{I}\end{bmatrix} \\
&= \det\left(\mathbf{I} - \frac{\rho^2}{4}\left(\mathbf{m}_1 + \mathbf{m}_2\right)^{\top}\left(\mathbf{m}_1 + \mathbf{m}_2\right)\right) \\
&= \det\left(\left(1 - \frac{\rho^2}{2}\right)\mathbf{I} - \frac{\rho^2}{4}\left(\mathbf{m}_1^{\top}\mathbf{m}_2 + \mathbf{m}_2^{\top}\mathbf{m}_1\right)\right). \quad (3)
\end{aligned}$$

Notice that $\mathbf{m}_1^{\top}\mathbf{m}_2 \in \{0,1\}^{\mathcal{U} \times \mathcal{U}}$ is the permutation matrix corresponding to permutation $\pi = m_1 \circ m_2^{\top}$ described in the statement of the lemma. Let $\mathcal{C}$ be the set of cycles of $\pi$ and $\{\ell_c\}_{c \in \mathcal{C}}$ denote their lengths. Consider the cycle notation of this permutation, i.e. $(u_1, u_2, \cdots, u_{\ell_c})(u'_1, \cdots, u'_{\ell_{c'}}) \cdots$, and specify an ordering of $\mathcal{U}$ based on this expression: $u_1, u_2, \cdots, u_{\ell_c}, u'_1, \cdots, u'_{\ell_{c'}}, \cdots$. Given this ordering of rows and columns, the permutation matrix $\mathbf{m}_1^{\top}\mathbf{m}_2$ has block diagonal matrix form, with one block for each cycle $c \in \mathcal{C}$ and every block having the form of a shifted identity matrix $\mathbf{I}^{(\ell_c,+)}$. Then $\mathbf{m}_2^{\top}\mathbf{m}_1 = \left(\mathbf{m}_1^{\top}\mathbf{m}_2\right)^{\top}$ has the same block diagonal form with the shifted identity matrices $\mathbf{I}^{(\ell_c,-)}$, since $\mathbf{I}^{(\ell_c,-)} = \left(\mathbf{I}^{(\ell_c,+)}\right)^{\top}$.

The determinant of a block diagonal matrix is equal to the product of the determinants of each block. Then we have

$$\begin{aligned}
&\det\left(\left(1 - \frac{\rho^2}{2}\right)\mathbf{I} - \frac{\rho^2}{4}\left(\mathbf{m}_1^{\top}\mathbf{m}_2 + \mathbf{m}_2^{\top}\mathbf{m}_1\right)\right) \\
&= \prod_{c \in \mathcal{C}} \det\left(\left(1 - \frac{\rho^2}{2}\right)\mathbf{I} - \frac{\rho^2}{4}\left(\mathbf{I}^{(\ell_c,+)} + \mathbf{I}^{(\ell_c,-)}\right)\right) \\
&= \prod_{c \in \mathcal{C}} \det\left(\mathbf{L}^{\ell_c}\left(1 - \frac{\rho^2}{2}, \frac{\rho^2}{2}\right)\right) \\
&= \prod_{\ell \in [n]} \left[\det\left(\mathbf{L}^{\ell}\left(1 - \frac{\rho^2}{2}, \frac{\rho^2}{2}\right)\right)\right]^{k_\ell},
\end{aligned}$$

where $k_\ell$ denotes the number of cycles of length $\ell$ in the permutation $\pi$. Combining this with (2) and (3) gives us the claimed result. □

**Lemma 4.3.** *For any $\ell \in \mathbb{N}^+$,*

$$\det\left(\mathbf{L}^\ell(s,t)\right) = \prod_{j\in[\ell]} \left[s - t\cdot\cos\left(j\frac{2\pi}{\ell}\right)\right].$$

*In particular*

$$\det\left(\mathbf{L}^1(s,t)\right) = s - t \quad \textit{and} \quad \det\left(\mathbf{L}^2(s,t)\right) = s^2 - t^2$$

*Proof.* Let $\vec{z}^k \in \mathbb{C}^\ell$ denote a family of vectors such that for any $k \in [\ell]$, $\vec{z}^k_j = e^{2\pi i \frac{jk}{\ell}}$, where $i^2 = -1$. Observe that

$$\mathbf{I}^{(+)}\vec{z}^k = e^{2\pi i\frac{k}{\ell}}\vec{z}^k$$
$$\mathbf{I}^{(-)}\vec{z}^k = e^{-2\pi i\frac{k}{\ell}}\vec{z}^k$$

Vectors $\vec{z}^k$ are the eigenvectors of $\mathbf{L}^k(s,t)$:

$$\begin{aligned}\mathbf{L}^\ell(s,t)\ \vec{z}^k &= \left[s\cdot\mathbf{I} - \frac{t}{2}\left(\mathbf{I}^{(+)} + \mathbf{I}^{(-)}\right)\right]\vec{z}^k \\ &= \left[s - \frac{t}{2}\left(e^{2\pi i\frac{k}{\ell}} + e^{-2\pi i\frac{k}{\ell}}\right)\right]\vec{z}^k \\ &= \left[s - t\cdot\cos\left(2\pi\frac{k}{\ell}\right)\right]\vec{z}^k\end{aligned}$$

We compute the determinant by taking the product of the $\ell$ eigenvalues (one for each $k \in [\ell]$). □

**Lemma 4.4.** *For any $\ell \in \mathbb{N}\setminus\{0,1\}$ and $s,t \in \mathbb{R}$ such that $s > |t|$,*

$$\det\left[\mathbf{L}^\ell(s,t)\right] \geq \left(\det\left[\mathbf{L}^2(s,t)\right]\right)^{\ell/2}$$

*Proof.* First note that, by Lemma 4.3,

$$\det\left[\mathbf{L}^2(s,t)\right] = s^2 - t^2.$$

We want to bound the determinant of the matrix $\mathbf{L}^\ell(s,t)$, which is equal to the product of its eigenvalues $(\lambda_j)_{j\in[\ell]}$. The sum of eigenvalues is equal to the trace of the matrix, which is known, since all diagonal elements of $\mathbf{L}^\ell(s,t)$ equal $s$ for any $\ell \geq 2$. So $\sum\lambda_k = \operatorname{tr}\left(\mathbf{L}^\ell(s,t)\right) = s\ell$. Furthermore, observe that all eigenvalues are in the range $[s-t, s+t]$. Consider a sequence formed of two copies of each eigenvalue $\lambda_i$. This sequence has mean $s$ and has all entries within the range $[s-t, s+t]$. Then, as it is proven in our supplementary material as Lemma B.1,

$$\prod_{j\in[2\ell]}\lambda_j^2 \geq (s-t)^\ell\,(s+t)^\ell$$

Taking the square root of both sides results in the claim. □

**Theorem 1. (Achievability)** *If mutual information between feature pairs $I_{\vec{X}\vec{Y}} \geq 2\log n + \omega(1)$, then the MAP estimator returns the proper alignment with probability $1 - o(1)$.*

*Proof.* Recall the canonical setting where $\mathbf{\Sigma}_{\mathrm{a}} = \mathbf{\Sigma}_{\mathrm{b}} = I$ and $\mathbf{\Sigma}_{\mathrm{ab}} = \operatorname{diag}(\vec{\rho})$.

Let $R_i(m,m')$ be the value of $R(m,m')$ when $\mathbf{\Sigma} = \begin{bmatrix}1 & \vec{\rho}_i\\ \vec{\rho}_i & 1\end{bmatrix}$. By the union bound

$$\begin{aligned}&\Pr\left[F \in \bigcup_{m'\neq m}\mathcal{E}(m,m')|M=m\right] \\ &\leq \sum_{m'\neq m}\Pr\left[F\in\mathcal{E}(m,m')|M=m\right] \\ &\overset{(a)}{\leq} \sum_{m'\neq m} R(m,m') = \sum_{m'\neq m}\prod_{i\in[d]} R_i(m,m') \\ &\overset{(b)}{=} \sum_{m'\neq m}\prod_{i\in[d]}\left[\frac{\left(1-\rho_i^2\right)^n}{\prod_\ell\left[\det\left(\mathbf{L}^\ell\left(1-\frac{\rho_i^2}{2},\frac{\rho_i^2}{4}\right)\right)\right]^{k_\ell}}\right]^{\frac{1}{2}} \\ &\overset{(c)}{\leq} \sum_{m'\neq m}\prod_{i\in[d]}\left[\frac{\left(1-\rho_i^2\right)^n}{(s-t)^{|m\cap m'|}\left(s^2-t^2\right)^{\frac{1}{2}(n-|m\cap m'|)}}\right]^{\frac{1}{2}} \\ &= \sum_{m'\neq m}\prod_{i\in[d]}\left(1-\rho_i^2\right)^{\frac{n-|m\cap m'|}{4}},\end{aligned}$$

where (a) follows from Lemma 4.1, (b) follows from Lemma 4.2, with $k_\ell$ denoting the number of cycles of length $\ell$ in the permutation $m'\circ m^\top$, and (c) follows from Lemmas 4.3 and 4.4, with $s = 1-\frac{\rho_i^2}{2}$ and $t = \frac{\rho_i^2}{2}$, which gives us $s - t = s^2 - t^2 = 1 - \rho_i^2$.

Given any $k \in \mathbb{N}$ there are exactly $(!k)\times\binom{n}{k}$ different matchings $m'$ such that $k = n - |m\cap m'|$, where $(!k)$ represents the number of derangements over a set of size $k$. We bound $(!k)\times\binom{n}{k} \leq n^k$. Thus

$$\begin{aligned}&\Pr\left[F\in\bigcup_{m'\neq m}\mathcal{E}(m,m')|M=m\right] \\ &\qquad\leq \sum_{k\in\mathbb{N}} n^k\cdot\prod_{i\in[d]}\left(1-\rho_i^2\right)^{k/4}\end{aligned}$$

If $n\prod_{i\in[d]}\left(1-\rho_i^2\right)^{\frac{1}{4}} \leq o(1)$, then by summing the geometric series, we see that the above expression is $o(1)$. Therefore

$$\exp\left(-I_{XY}\right) = \prod_{i\in[d]}\left(1-\rho_i^2\right)^{\frac{1}{2}} \leq o(1/n^2)$$

is a sufficient condition for exact recovery under the canonical setting. Taking the logarithm of both sides gives us the claimed result. □

## 5 Converse analysis

We establish a necessary condition on the mutual information $I_{XY}$ between feature pairs to achieve a perfect alignment.

**Lemma 5.1.** *Let $d \in \mathbb{N}$ such that $d = \omega(1)$ as well as $\mathbf{\Sigma}_A = \mathbf{\Sigma}_B = \mathbf{I}^d$ and $\mathbf{\Sigma}_{AB} = \rho\mathbf{I}$. Given bijective matchings $m_1, m_2 \subset \mathcal{U}\times\mathcal{V}$ such that $|m_1 \cap m_2| = n-2$,*

$$\Pr[F \in \mathcal{E}(m_1, m_2)|M = m_1] \geq (1-\rho_i^2)^{d(1+o(1))}.$$

*Proof.* Consider the conditional generating function

$$\begin{aligned} c_i(\theta) &= \mathbb{E}\left[\exp\left(\theta \log \frac{p_{\mathbf{F}_{*i}|M}(\mathbf{f}_i|m_2)}{p_{\mathbf{F}_{*i}|M}(\mathbf{f}_i|m_1)}\right)\Big|M = m_1\right] \\ &= \int \left(\frac{p_{\mathbf{F}_{*i}|M}(\mathbf{f}_i|m_2)}{p_{\mathbf{F}_{*i}|M}(\mathbf{f}_i|m_1)}\right)^{\theta} p_{\mathbf{F}_{*i}|M}(\mathbf{f}_i|m_1)d\mathbf{f}_i \\ &= \int (p_{\mathbf{F}_{*i}|M}(\mathbf{f}_i|m_2))^{\theta}(p_{\mathbf{F}_{*i}|M}(\mathbf{f}_i|m_1))^{1-\theta}d\mathbf{f}_i \end{aligned}$$

The generating function is minimized at $\theta = 1/2$ in which case we get $c_i(\theta) = R_i(m_1, m_2)$.

We evaluate the value of this function using Lemmas 4.2 and 4.3 with $s = 1-\rho^2/2$ and $t = \rho^2/2$. By $|m_1 \cap m_2| = n-2$ we get $R_i(m_1, m_2) = \sqrt{1-\rho^2}$.

By Cramér's Theorem on the asymptotic tightness of the Chernoff bound (see for example [10]), there is some $\epsilon(d) \leq o(1)$ such that

$$\begin{aligned} &\Pr\left[\log \frac{p_{\mathbf{F}|M}(\mathbf{f}|m_2)}{p_{\mathbf{F}|M}(\mathbf{f}|m_1)} \geq 0\right] \\ &= \Pr\left[\sum_{i\in[d]} \log \frac{p_{\mathbf{F}_{*i}|M}(\mathbf{f}_{*i}|m_2)}{p_{\mathbf{F}_{*i}|M}(\mathbf{f}_{*i}|m_1)} \geq 0\right] \\ &\geq \exp\left(-d\left[\epsilon - \inf_{\theta} \log c_i(\theta)\right]\right) \\ &= \exp\left(d(\log R_i(m_1, m_2) - \epsilon\right) \\ &\geq (R_i(m_1, m_2))^{d(1+o(1))} \\ &= \left(1-\rho^2\right)^{d(1+o(1))} \end{aligned}$$

□

**Lemma 5.2.** *If $\mathbf{\Sigma}_A = \mathbf{\Sigma}_B = I$ and $\mathbf{\Sigma}_{AB} = \rho\mathbf{I}$, then given any bijective matchings $m_1, m_2, m_3 \in \mathcal{U}\times\mathcal{V}$*

$$\begin{aligned} &\Pr[F \in \mathcal{E}(m_1, m_2) \cap \mathcal{E}(m_1, m_3)|M = m_1] \\ &\qquad \leq \left(1-\rho_i^2\right)^{\frac{d}{2}(n-|m_2\cap m_3|)}. \end{aligned}$$

*Proof.* We will abbreviate $p_{\mathbf{F}|M}(\cdot|\cdot)$ as $p(\cdot|\cdot)$.

For any $\theta, \theta' > 0$ we have

$$\begin{aligned} &\Pr[F \in \mathcal{E}(m_1, m_2) \cap \mathcal{E}(m_1, m_3)|M = m_1] \\ &= \mathbb{E}\left[\mathbb{1}\left\{\frac{p(\mathbf{f}|m_2)}{p(\mathbf{f}|m_1)} \geq 1, \frac{p(\mathbf{f}|m_3)}{p(\mathbf{f}|m_1)} \geq 1\right\}\Big|M = m_1\right] \\ &\leq \int \left(\frac{p(\mathbf{f}|m_2)}{p(\mathbf{f}|m_1)}\right)^{\theta}\left(\frac{p(\mathbf{f}|m_3)}{p(\mathbf{f}|m_1)}\right)^{\theta'} p(\mathbf{f}|m_1)d\mathbf{f} \\ &= \int (p(\mathbf{f}|m_2))^{\theta}(p(\mathbf{f}|m_3))^{\theta'}(p(\mathbf{f}|m_1))^{1-\theta-\theta'}d\mathbf{f}. \end{aligned}$$

The choice of $\theta = \theta' = 1/2$ gives the upper bound as $R(m_2, m_3)$. We evaluate this function using Lemmas 4.2 and 4.3, which give us the claimed result. □

**Theorem 2. (Converse)** *Let $d \in \mathbb{N}$ such that $d \geq \omega(1)$. Furthermore let $\mathbf{\Sigma}_a = \mathbf{\Sigma}_b = \mathbf{I}^d$ and $\mathbf{\Sigma}_{ab} = \rho\mathbf{I}$. If $I_{\vec{X}\vec{Y}} \leq 2\log n(1-\Omega(1))$, then any for algorithm, the probability of returning the proper alignment is $o(1)$.*

*Proof.* Let $\mathcal{M}^{\mathcal{E}}(f, m) \triangleq \{m' | f \in \mathcal{E}(m, m'), m' \neq m\}$ denote the set of matches that are at least as likely as $m$ under the database instance $f$. The MAP algorithm succeeds if and only if $\mathcal{M}^{\mathcal{E}}(F, M) = \emptyset$.

Also define $\mathcal{M}_2(m) \triangleq \left\{m' \middle| |m \cap m'| = n-2\right\}$. For compactness, let $X \triangleq |\mathcal{M}^{\mathcal{E}}(f, m) \cap \mathcal{M}_2(m)|$. Clearly $0 \leq X \leq |\mathcal{M}^{\mathcal{E}}(f, m)|$.

We apply Chebyshev's inequality:

$$\begin{aligned} \Pr[|\mathcal{M}^{\mathcal{E}}(F, M)| = 0] &\leq \Pr[X = 0] \\ &\leq \Pr\left[(X - \mathbb{E}X)^2 \geq \mathbb{E}^2 X\right] \leq \frac{\operatorname{Var} X}{\mathbb{E}^2 X} \end{aligned}$$

All matchings are equally likely. Therefore, given any bijective matching $m \in \mathcal{U}\times\mathcal{V}$,

$$\mathbb{E}\left|\mathcal{M}_2^{\mathcal{E}}(F, M)\right| = \sum_{m'\in\mathcal{M}_2(m)} \Pr[F \in \mathcal{E}(m, m')|M = m]$$

Let $\varepsilon_1 \triangleq \Pr[F \in \mathcal{E}(m, m')|M = m]$ given $|m \cap m'| = n-2$. Notice that this probability does not depend on the choice of $m' \in \mathcal{M}_2(m)$. Then $\mathbb{E}\left|\mathcal{M}_2^{\mathcal{E}}(F, M)\right| = |\mathcal{M}_2(m)| \cdot \varepsilon_1 = \binom{n}{2} \cdot \varepsilon_1$.

$$\begin{aligned} &|\mathcal{M}_2^{\mathcal{E}}(f, m)|^2 \\ &= \Big(\sum_{m'\in\mathcal{M}_2(m)} \mathbb{1}\left\{f \in \mathcal{E}(m, m')\right\}\Big)^2 \\ &= \sum_{m'\in\mathcal{M}_2(m)} \mathbb{1}\left\{\mathcal{E}(m, m')\right\} \\ &\quad +2 \sum_{\{m', m''\}\subset\mathcal{M}_2(m)} \mathbb{1}\left\{\mathcal{E}(m, m'), \mathcal{E}(m, m'')\right\} \end{aligned}$$

There are $3\binom{n}{4}$ different ways to choose to matchings $\{m', m''\} \subset \mathcal{M}_2(m)$ such that $|m' \cap m''| = n-4$, and

$3\binom{n}{3}$ ways to choose them such that $|m' \cap m''| = n-3$. Notice that $3\binom{n}{4}+3\binom{n}{3} = \binom{|\mathcal{M}_2(m)|}{2}$ and these partition are all the choices $\{m', m''\} \subset \mathcal{M}_2(m)$.

When $|m' \cap m''| = n-4$, the error events become independent and we get

$$\Pr[F \in \mathcal{E}(m,m') \cap \mathcal{E}(m,m'')|M=m] = \varepsilon_1^2.$$

Let $\varepsilon_2 \triangleq \Pr[F \in \mathcal{E}(m,m') \cap \mathcal{E}(m,m'')|M=m]$ given $|m' \cap m''| = n-3$.

By the relation $z + 2\binom{z}{2} = z^2$. For $z = |\mathcal{M}_2(m)| = \binom{n}{2}$ and $\binom{z}{2} = 3\binom{n}{3} + 3\binom{n}{4}$ we can write:

$$\begin{aligned}
\mathbb{E}^2\left|\mathcal{M}_2^{\mathcal{E}}(F,M)\right| &= |\mathcal{M}_2(m)|^2\varepsilon_1^2 \\
&= \binom{n}{2}\varepsilon_1^2 + \left[6\binom{n}{3} + 6\binom{n}{4}\right]\varepsilon_1^2 \\
\mathbb{E}\left[|\mathcal{M}_2^{\mathcal{E}}(F,M)|^2\right] &= \binom{n}{2}\varepsilon_1 + 6\binom{n}{3}\varepsilon_2 + 6\binom{n}{4}\varepsilon_1^2 \\
\operatorname{Var}\left|\mathcal{M}_2^{\mathcal{E}}(F,M)\right| &= \binom{n}{2}(\varepsilon_1 - \varepsilon_1^2) + 6\binom{n}{3}(\varepsilon_2 - \varepsilon_1^2) \\
&\le \binom{n}{2}\varepsilon_1 + 6\binom{n}{3}\varepsilon_2
\end{aligned}$$

Plugging these values into the Chernoff bound we get

$$\begin{aligned}
\Pr[|\mathcal{M}^{\mathcal{E}}(F,M)| = 0] &\le \frac{\binom{n}{2}\varepsilon_1 + 6\binom{n}{3}\varepsilon_2}{\binom{n}{2}^2\varepsilon_1^2} \\
&\le \mathcal{O}\left(\frac{1}{n^2\varepsilon_1} + \frac{\varepsilon_2}{n\varepsilon_1^2}\right)
\end{aligned}$$

By lemma 5.1 and 5.2 we have $\varepsilon_1 \ge (1-\rho_i^2)^{d(1+o(1))}$ and $\varepsilon_2 \le (1-\rho_i^2)^{3d/2}$. Thus $\varepsilon_1^2/\varepsilon_2 \ge (1-\rho^2)^{\frac{d}{2}(1+o(1))}$. If $(1-\rho_i^2)^d \ge n^{-2+\Omega(1)}$, then

$$n^2(1-\rho_i^2)^{d(1+o(1))} \ge n^{2+(1+o(1))(-2+\Omega(1))} \ge n^{\Omega(1)}$$

and $\Pr[|\mathcal{M}^{\mathcal{E}}(F,M)| = 0] \le \mathcal{O}(n^{-\Omega(1)}) \le o(1)$. □

# 6 Binary hypothesis testing

**Matching algorithm** We consider an algorithm that does gives us a 'matching' $\hat{m} \subseteq \mathcal{U} \times \mathcal{V}$ that is not necessarily bijective, i.e. any entry can have multiple matches in the other dataset.

Recall that we denote the $j$-th row of a matrix $\mathbf{z}$ by $\vec{z}_{j*}$.

Given some $\mathbf{a} \in \mathbb{R}^{\mathcal{U}\times[d_1]}$ and $\mathbf{b} \in \mathbb{R}^{\mathcal{V}\times[d_2]}$ and $f = (\mathbf{a}, \mathbf{b})$ the estimated 'matching' is given by

$$\hat{m}(f) = \left\{(u,v) \in \mathcal{U}\times\mathcal{V} \middle| (\vec{a}_{u*}^\top, \vec{b}_{v*}^\top) \in H_\tau\right\}.$$

$H_\tau$ is the log ratio test given by

$$H_\tau = \left\{(\vec{x},\vec{y}) \in \mathbb{R}^d \times \mathbb{R}^d \middle| \log\frac{p_{\vec{X}\vec{Y}}(\vec{x},\vec{y})}{p_{\vec{X}}(\vec{x})p_{\vec{Y}}(\vec{y})} \ge \tau\right\}$$

where $p_{\vec{X}}$ and $p_{\vec{Y}}$ denote the probability density functions of feature vectors associated with identifiers in $\mathcal{U}$ and $\mathcal{V}$ respectively, and $\tau \in \mathbb{R}$ is some constant to be determined.

## 6.1 Achievability analysis

In our analysis we establish upper and lower bounds on the threshold $\tau$ that allow given probability bounds on false negatives and false positives. The mean and variance of the log ratio random variable were computed in Section 2.2. Using these values we get an upper bound on the probability of false negatives in Lemma 6.1 by the Chebyshev inequality. Lemma 6.2 gives an upper bound on the number of false positives. Finally, taking the intersection of the conditions on $\tau$ allows us to derive the achievability result given in Theorem 3.

**Lemma 6.1.** *If $\tau \le I_{\vec{X}\vec{Y}} - \sigma_{\vec{X}\vec{Y}}/\sqrt{\varepsilon}$ then*

$$\Pr\left[(\mathbf{A}_{u*}^\top, \mathbf{B}_{v*}^\top) \notin H_\tau | (u,v) \in M\right] \le \varepsilon.$$

*Proof.* Let $(u,v) \in M$ and $(\vec{X},\vec{Y}) = (\mathbf{A}_{u*}^\top, \mathbf{B}_{v*}^\top)$. Given $\mu = I_{\vec{X}\vec{Y}} = \mathbb{E}\left[\log\frac{p_{\vec{X}\vec{Y}}(\vec{X},\vec{Y})}{p_{\vec{X}}(\vec{X})p_{\vec{Y}}(\vec{Y})}\right] = I_{\vec{X}\vec{Y}}$ and $\sigma^2 = \sigma_{\vec{X}\vec{Y}}^2 = \operatorname{Var}\left(\log\frac{p_{\vec{X}\vec{Y}}(\vec{X},\vec{Y})}{p_{\vec{X}}(\vec{X})p_{\vec{Y}}(\vec{Y})}\right)$, by Chebyshev's inequality we get

$$\Pr\left[\left|\mu - \log\frac{p_{\vec{X}\vec{Y}}(\vec{X},\vec{Y})}{p_{\vec{X}}(\vec{X})p_{\vec{Y}}(\vec{Y})}\right| \ge \frac{\sigma}{\sqrt{\varepsilon}}\right] \le \varepsilon$$

This probability is lower bounded by $\Pr\left[\mu - \log\frac{p_{\vec{X}\vec{Y}}(\vec{X},\vec{Y})}{p_{\vec{X}}(\vec{X})p_{\vec{Y}}(\vec{Y})} \ge \frac{\sigma}{\sqrt{\varepsilon}}\right]$ which is equal to $\Pr\left[(\mathbf{A}_{u*}^\top, \mathbf{B}_{v*}^\top) \notin H_\tau | (u,v) \in M\right]$ for $\tau = \mu - \sigma/\sqrt{\varepsilon}$. Then this choice of $\tau$, or any smaller value, is a sufficient condition to bound the error probability by $\varepsilon$. □

**Lemma 6.2.** *Given any $\tau \in \mathbb{R}$,*

$$\Pr\left[(\mathbf{A}_{u*}^\top, \mathbf{B}_{v*}^\top) \in H_\tau | (u,v) \notin M\right] \le e^{-\tau}$$

*Proof.* Let $(u,v) \in M$ and $(\vec{X},\vec{Y}) = (\mathbf{A}_{u*}^\top, \mathbf{B}_{v*}^\top)$. By Markov's inequality we get

$$\Pr\left[\log\frac{p_{\vec{X}\vec{Y}}(\vec{X},\vec{Y})}{p_{\vec{X}}(\vec{X})p_{\vec{Y}}(\vec{Y})} \ge \tau\right] \le e^{-\tau} \cdot \mathbb{E}\left[\frac{p_{\vec{X}\vec{Y}}(\vec{X},\vec{Y})}{p_{\vec{X}}(\vec{X})p_{\vec{Y}}(\vec{Y})}\right]$$

We calculate the mean:

$$\begin{aligned}&\mathbb{E}\left[\frac{p_{\vec{X}\vec{Y}}(\vec{X},\vec{Y})}{p_{\vec{X}}(\vec{X})p_{\vec{Y}}(\vec{Y})}\right]\\&=\int_{\vec{X},\vec{Y}} p_{\vec{X}}(\vec{X})p_{\vec{Y}}(\vec{Y})\cdot\frac{p_{\vec{X}\vec{Y}}(\vec{X},\vec{Y})}{p_{\vec{X}}(\vec{X})p_{\vec{Y}}(\vec{Y})}d(\vec{X},\vec{Y})\\&=\int_{\vec{X},\vec{Y}} p_{\vec{X}\vec{Y}}(\vec{X},\vec{Y})d(\vec{X},\vec{Y})\end{aligned}$$

which equals to 1 since $p_{\vec{X}\vec{Y}}(\vec{X},\vec{Y})$ is as a probability density function. □

**Proof of Theorem 3** By Lemma 6.1, if $\tau \le I_{\vec{X}\vec{Y}} - \sigma_{\vec{X}\vec{Y}}/\sqrt{\bar{\varepsilon}}$, then the probability that any correct match is not included in $H_\tau$ is upper bounded by $\varepsilon_{FN}/n$. There are $n$ correct matches in $\mathcal{U}\times\mathcal{V}$. Then the expected number of correct matches not included in $H_\tau$, i.e. the expected number of false negatives, is upper bounded by $\varepsilon_{FN}$.

By Lemma 6.2, if $\tau \ge \log\left(n^2/\varepsilon_{FP}\right)$, then the probability that any incorrect match is included in $H_\tau$ is upper bounded by $\varepsilon_{FP}/n^2$. There are $\binom{n}{2} < n^2$ incorrect matches in $\mathcal{U}\times\mathcal{V}$. Then the expected number of incorrect matches included in $H_\tau$, i.e. the expected number of false positives, is upper bounded by $\varepsilon_{FP}$.

A choice for $\tau\in\mathbb{R}$ satisfying both conditions exists if and only if the condition in the theorem statement holds. □

### 6.2 Converse analysis

We present a converse on the performance of the binary hypothesis testing algorithm based on Fano's inequality.

**Lemma 6.3.** *For $u\in\mathcal{U}$ and $v\in\mathcal{V}$, $H(\mathbf{M}_{u,v}|\mathbf{A}_u,\mathbf{B}_v)\ge\frac{\log n - I_{\vec{X}\vec{Y}}}{n}$.*

*Proof.* We have

$$\begin{aligned}&H(\mathbf{M}_{u,v}|\mathbf{A}_u,\mathbf{B}_v)=H(\mathbf{M}_{u,v})+I(\mathbf{A}_u;\mathbf{B}_v)\\&-I(\mathbf{M}_{u,v};\mathbf{A}_u)-I(\mathbf{M}_{u,v};\mathbf{B}_v)-I(\mathbf{A}_u;\mathbf{B}_v|\mathbf{M}_{u,v}).\end{aligned}$$

Then $I(\mathbf{M}_{u,v};\mathbf{A}_u)=I(\mathbf{M}_{u,v};\mathbf{B}_v)=0$ and

$$\begin{aligned}&I(\mathbf{A}_u;\mathbf{B}_v|\mathbf{M}_{u,v})\\&=\frac{n-1}{n}I(\mathbf{A}_u|(\mathbf{M}_{u,v}=0);\mathbf{B}_v|(\mathbf{M}_{u,v}=0))\\&\quad+\frac{1}{n}I(\mathbf{A}_u|(\mathbf{M}_{u,v}=1);\mathbf{B}_v|(\mathbf{M}_{u,v}=1))\\&=\frac{n-1}{n}\cdot 0+\frac{1}{n}I_{\vec{X}\vec{Y}}.\end{aligned}$$

Finally $I(\mathbf{A}_u;\mathbf{B}_v)\ge 0$ and $H(\mathbf{M}_{u,v})=\frac{1}{n}\log n+\frac{n-1}{n}\log\frac{n}{n-1}\ge\frac{\log n}{n}$. □

*Proof of Theorem 4.* Let $\hat{\mathbf{M}}_{u,v}\triangleq\mathbb{1}\left\{\left(\mathbf{A}_u,\mathbf{B}_v\right)\in H_\tau\right\}$ denote the estimation on the relation between identifiers $u$ and $v$. We have a correct estimation if $\hat{\mathbf{M}}_{u,v}=\mathbf{M}_{u,v}$. Define $E\triangleq\mathbb{1}\left\{\hat{\mathbf{M}}_{u,v}\ne\mathbf{M}_{u,v}\right\}$. Then by Fano's inequality,

$$H(\mathbf{M}_{u,v}|\mathbf{A}_u,\mathbf{B}_v)\le H(E)+\Pr[E=1],$$

which gives the upper bound as $H(E)$.

Let $\epsilon\triangleq\Pr[E=1]$. This value can also be expressed as the expected frequency of false matches, i.e. given $\varepsilon_{FN}$ and $\varepsilon_{FP}$ the expected number of false negatives and false positives, $\epsilon=\frac{\varepsilon_{FN}+\varepsilon_{FP}}{|\mathcal{U}\times\mathcal{V}|}=\frac{\varepsilon_{FN}+\varepsilon_{FP}}{n^2}$.

Let $H_b$ denote the binary entropy function. By Fano's inequality, using Lemma 6.3, we have

$$H_b(\epsilon)\ge H(\mathbf{M}_{u,v}|\mathbf{A}_u,\mathbf{B}_v)\ge\frac{\log n - I_{\vec{X}\vec{Y}}}{n}\qquad(4)$$

We have

$$\begin{aligned}H_b(\epsilon)&\le-\epsilon\log\epsilon+\epsilon\\&=\frac{\varepsilon_{FN}+\varepsilon_{FP}}{n^2}\left(2\log n-\log\left(\varepsilon_{FN}+\varepsilon_{FP}\right)+1\right).\end{aligned}$$

Combining this with (4) gives us

$$\varepsilon_{FN}+\varepsilon_{FP}\ge\frac{n}{2}\left(\frac{\log n - I_{\vec{X}\vec{Y}}}{2\log n+1}\right)$$

and the claim follows. □

### References


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

[2] D. Cullina, P. Mittal, and N. Kiyavash, "Fundamental limits of database alignment," in *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 651–655, IEEE, 2018.

[3] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Matching anonymized and obfuscated time series to users' profiles," *CoRR*, vol. abs/1710.00197, 2017.

[4] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Matching anonymized and obfuscated time series to users' profiles," *CoRR*, vol. abs/1710.00197, 2017.

[5] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Privacy against statistical matching: Inter-user correlation," *CoRR*, vol. abs/1805.01296, 2018.

[6] 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*, pp. 1235–1243, ACM, 2011.

[7] D. Cullina and N. Kiyavash, "Exact alignment recovery for correlated Erdős Rényi graphs," *arXiv:1711.06783 [cs, math]*, Nov. 2017. arXiv: 1711.06783.

[8] D. Cullina and N. Kiyavash, "Improved achievability and converse bounds for Erdős-Rényi graph matching," in *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, pp. 63–72, ACM, 2016.

[9] J. Ding, Z. Ma, Y. Wu, and J. Xu, "Efficient random graph matching via degree profiles," *arXiv preprint arXiv:1811.07821*, 2018.

[10] B. Hajek, *Random Processes for Engineers*. Cambridge University Press, 2015.