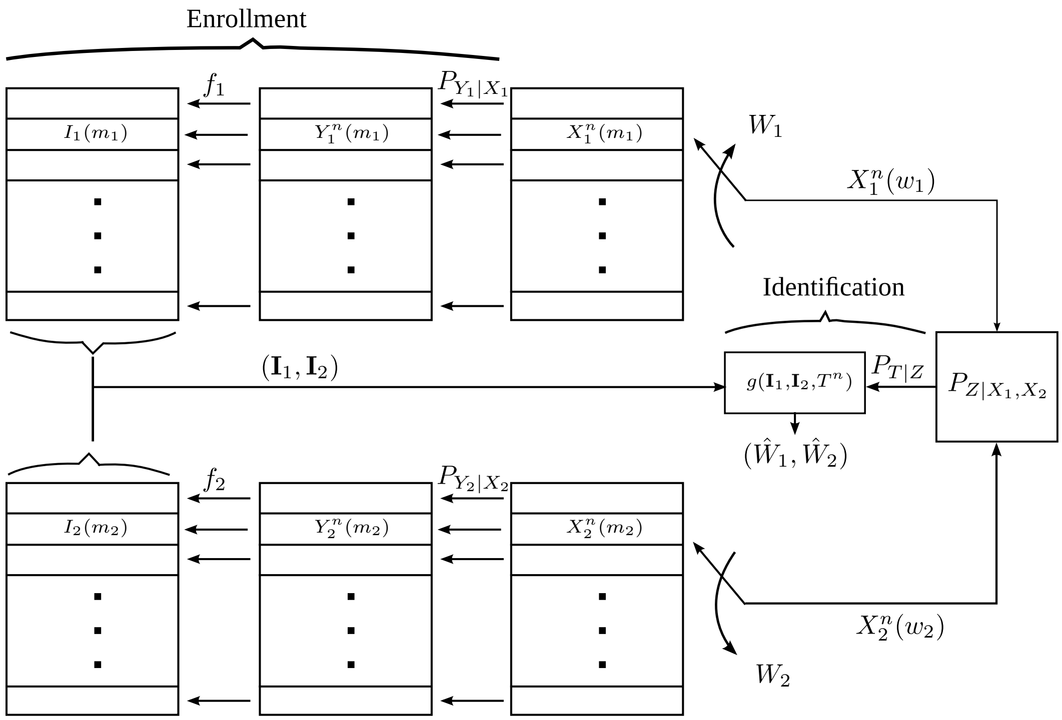

Fig. 1. The illustration of the enrollment and identification processes over two databases.

respectively. See Fig. 1 for an illustration of the system model. The underlying assumption is that the feature vectors are i.i.d. according to a probability distribution specific to each group. We have, for $j = 1, 2$ and $1 \leq m_j \leq M_j$, $P[X_j^n(m_j) = x_j^n] = \prod_{i=1}^n P_{X_j}(x_{ji})$, over the finite feature alphabets $\mathcal{X}_j$.

We assume that we have separate databases for the males and the females in the population. The databases are formed by an enrollment phase, in which the noisy version of the feature vector of an individual is observed and recorded to the corresponding database. We denote the observed noisy feature vectors by $Y_1^n(m_1)$ for males and $Y_2^n(m_2)$ for females, which are assumed to be the outputs of the corresponding discrete memoryless channel (DMC), which might be different for males and females. The DMCs are characterized by $P_{Y_1|X_1}$ and $P_{Y_2|X_2}$ for males and females, respectively, where $\mathcal{Y}_i$, $i = 1, 2$, are the finite observation alphabets. We have

$$P[Y_j^n(m_j) = y_j^n | X_j^n(m_j) = x_j^n] = \prod_{i=1}^n P_{Y_j|X_j}(y_{ji}|x_{ji}),$$

for $j = 1, 2$ and $1 \leq m_j \leq M_j$.

In the enrollment phase, each entry is compressed before it is recorded to the database, and only the compressed descriptions of the observed feature vectors are stored in the database. We consider two distinct deterministic functions for compression of each group: $f_j : \mathcal{Y}_j^n \to \mathcal{L}_j = \{1, \ldots, L_j\}$, for $j = 1, 2$, where $\mathcal{L}_j$ denotes the index set for the compressed observation vectors. We denote the index for entry $m_j \in \{1, \ldots, M_j\}$ as $I_j(m_j) = f_j(Y_j^n(m_j))$. These indices refer to n-length codewords of two separate codebooks of size L_j, $j = 1, 2$.

In the identification phase, W_1 and W_2 are chosen independent of each other and the database entries, and uniformly over the sets $\mathcal{M}_1 = \{1, \ldots, M_1\}$ and $\mathcal{M}_2 = \{1, \ldots, M_2\}$, respectively. The realizations are not known to the user of the database.

Now, consider an individual whose feature vector is derived from those of W_1 and W_2 through a DMC characterized by $P_{Z|X_1,X_2}$ with finite alphabet $\mathcal{Z}$, that is,

$$\begin{aligned} P[Z^n = z^n | X_1^n(W_1) =& x_1^n, X_2^n(W_2) = x_2^n] \\ &= \prod_{i=1}^n P_{Z|X_1,X_2}(z_i | x_{1i}, x_{2i}), \end{aligned} \tag{1}$$

where $Y_j^n - X_j^n - Z^n$ form Markov chains for $j = 1, 2$. The user of the database observes a noisy version of the feature vector Z^n of the individual corrupted by the DMC $P_{T|Z}$ with finite output alphabet $\mathcal{T}$, where $P[T^n = t^n | Z^n = z^n] = \prod_{i=1}^n P_{T|Z}(t_i | z_i)$.

Overall, the joint distribution of the random variables in the system is given by

$$P_{X_1,X_2,Y_1,Y_2,Z,T} = P_{X_1} P_{X_2} P_{Y_1|X_1} P_{Y_2|X_2} P_{Z|X_1,X_2} P_{T|Z}. \tag{2}$$

The user wants to identify W_1 and W_2, the two parents (ancestors) of the observed individual from each database by using the noisy observation vector T^n and the entries of the two databases $\{I_j(m_j)\}_{m_j=1}^{M_j}$, $j = 1, 2$.

The identification function is defined as $g : \mathcal{L}_1^{M_1} \times \mathcal{L}_2^{M_2} \times \mathcal{T}^n \to \mathcal{M}_1 \times \mathcal{M}_2$, and the corresponding estimates are denoted by $(\hat{W}_1, \hat{W}_2) = g(\mathbf{I}_1, \mathbf{I}_2, T^n)$, where we define $\mathbf{I}_1 \triangleq I_1(1), \ldots, I_1(M_1)$, and $\mathbf{I}_2 \triangleq I_2(1), \ldots, I_2(M_2)$.

The average error probability in the identification process is defined as

$$P_e^n \triangleq \frac{1}{M_1 M_2} \sum_{(w_1,w_2)} \Pr[(\hat{W}_1,\hat{W}_2)\neq(W_1,W_2)|(W_1,W_2)=(w_1,w_2)].$$

Definition 1: $(R_1^c, R_2^c, R_1^i, R_2^i)$ is an *achievable* compression/identification rate tuple for a parent identification system if, for any $\epsilon > 0$ and sufficiently large n, there exist deterministic enrollment functions f_1 and f_2 and a deterministic identification function g such that

$$\frac{1}{n}\log L_j \leq R_j^c \text{ and } \frac{1}{n}\log M_j \geq R_j^i, \text{ for } j=1,2, \quad (3)$$

and $P_e^n \leq \epsilon$.

Definition 2: The capacity region $\mathcal{R}$ for a parent identification system is the set of all achievable rate tuples $(R_1^c, R_2^c, R_1^i, R_2^i)$.

III. Main Result

In this section, we provide single-letter inner and outer bounds on the achievable rate region. These two bounds do not match in general, and the capacity region for the parent identification problem is still open. As is common for many multi-user information theory problems, the bounds are expressed in terms of auxiliary random variables U_1 and U_2, which are defined over finite alphabet sets $\mathcal{U}_1$ and $\mathcal{U}_2$, respectively. For a given joint distribution $P_{X_1,X_2,Y_1,Y_2,Z,T}$ that is in the form of (2), we define two different sets:

$$\begin{aligned}\mathcal{P}_{in} \triangleq \{(U_1,U_2) \quad : \quad & P_{U_1,U_2X_1,X_2,Y_1,Y_2,Z,T} = \\ & P_{U_1|Y_1}P_{U_2|Y_2}P_{X_1,X_2,Y_1,Y_2,Z,T}\},\end{aligned}$$

and

$$\begin{aligned}\mathcal{P}_{out} \triangleq \{(U_1,U_2) \quad : \quad & U_1 - Y_1 - X_1 - Z - T, \\ & U_2 - Y_2 - X_2 - Z - T\}.\end{aligned}$$

The set $\mathcal{P}_{in}$, in addition to the two Markov chain constraints in $\mathcal{P}_{out}$, has the additional constraint that (U_1, Y_1, X_1) is independent of (U_2, Y_2, X_2).

For a given pair of auxiliary random variables (U_1, U_2) jointly distributed with X_1, X_2, Y_1, Y_2, Z, T, we define the following rate region:

$$\begin{aligned}\mathcal{R}_{U_1,U_2} = \{(R_1^c,R_2^c,R_1^i,R_2^i) \quad : \quad & R_1^c \geq I(U_1;Y_1), \\ & R_2^c \geq I(U_2;Y_2), \\ & R_1^i \leq I(U_1;T|U_2), \\ & R_2^i \leq I(U_2;T|U_1) \text{ and} \\ & R_1^i + R_2^i \leq I(U_1,U_2;T)\}.\end{aligned}$$

Single letter bounds on $\mathcal{R}$ are given in the following theorem, whose proof can be found in Section IV.

Theorem 1: $\bar{\mathcal{R}}_{in} \subseteq \mathcal{R} \subseteq \bar{\mathcal{R}}_{out}$, where we define

$$\begin{aligned}\mathcal{R}_{in} \triangleq &\{(R_1^c,R_2^c,R_1^i,R_2^i) : (R_1^c,R_2^c,R_1^i,R_2^i) \in \mathcal{R}_{U_1,U_2} \\ & \text{for } (U_1,U_2) \in \mathcal{P}_{in}\}, \text{ and} \quad (4)\\ \mathcal{R}_{out} \triangleq &\{(R_1^c,R_2^c,R_1^i,R_2^i) : (R_1^c,R_2^c,R_1^i,R_2^i) \in \mathcal{R}_{U_1,U_2} \\ & \text{for } (U_1,U_2) \in \mathcal{P}_{out}\}, \quad (5)\end{aligned}$$

and $\bar{A}$ denotes the convex hull of the set A.

Corollary 1: If $R_j^c = H(Y_j)$ for $j = 1, 2$, then the two bounds match and the identification rate region is characterized as the rate pairs (R_1^i, R_2^i) satisfying

$$\begin{aligned}R_1^i &\leq I(Y_1;T|Y_2), \\ R_2^i &\leq I(Y_2;T|Y_1) \text{ and} \\ R_1^i + R_2^i &\leq I(Y_1,Y_2;T)\}.\end{aligned}$$

If we are interested in identifying a single parent, e.g., the case of asexual reproduction, the problem reduces to finding the rate region for compression and identification for a single database, i.e., $\mathcal{X}_2 = \mathcal{Y}_2 = \emptyset$. In this case, we have a single Markov chain and the upper and lower bounds match, and we recover the rate region obtained in [3], [4]:

Corollary 2: The compression/identification rate region for a single database system is the union of all rate pairs (R_1^c, R_1^i) satisfying

$$R_1^c \geq I(U_1;Y_1) \text{ and } R_1^i \leq I(U_1;T)$$

for some auxiliary random variable U_1 such that $U_1 - Y_1 - X_1 - Z - T$ forms a Markov chain.

Remark 1: It is pointed out in [6] that the capacity/storage tradeoff problem (or, the pattern recognition problem as stated in [3]) for a single database as in Corollary 2 is inherently related to the information bottleneck (IB) method introduced in [7]. Hence, the solution of the IB problem also constitutes a solution for the rate region in Corollary 2. However, this equivalence is established based on the single-letter characterization of the capacity region. For identification over multiple databases we do not have a capacity characterization. On the other hand, we can identify an equivalent multivariate IB problem [8] for the achievable rate region. Hence, we can use the algorithms proposed for the multivariate IB problem in [8] to numerically evaluate these regions.

IV. Proof of Theorem 1

We start with the proof of the inner bound, that is, $\bar{\mathcal{R}}_{in} \subseteq \mathcal{R}$. We assume that $(R_1^c, R_2^c, R_1^i, R_2^i) \in \mathcal{R}_{in}$. Fix any $\epsilon > 0$.

Codebook generation: For database $j = 1, 2$, generate a codebook consisting of L_j length-n codewords i.i.d. with distribution p_{U_j}. Enumerate these codewords as $U_j^n(l_j)$ where $l_j \in \{1, \ldots, L_j\}$. We will determine L_j later.

Enrollment: Given the noisy observation of a feature vector $y_j^n \in \mathcal{Y}_j^n$, define the enrollment function f_j as the smallest index l_j such that $(y_j^n, U_j^n(l_j)) \in T^n_{[U_jY_j]_\epsilon}$ [1]. We set $f_j(y_j^n) = 1$ if no such codeword exits.

Identification: In the identification phase, given any $t^n \in \mathcal{T}^n$, $\mathbf{I}_1$, and $\mathbf{I}_2$, we define the identification function g as the smallest pair of indices w_1 and w_2 such that $(t^n, U_1^n(I_1(w_1)), U_2^n(I_2(w_2))) \in T^n_{[TU_1U_2]_\epsilon}$. We set $g(t^n, \mathbf{I}_1, \mathbf{I}_2) = (1, 1)$ if no such pair can be found. We define $(\hat{w}_1, \hat{w}_2) = g(t^n, \mathbf{I}_1, \mathbf{I}_2)$.

Probability of Error Analysis: Define the following events

$$E_{1j}(w_j) \triangleq \left\{ \left(Y_j^n(w_j), U_j^n(I_j(w_j))\right) \in T^n_{[U_jY_j]_\epsilon} \right\}$$

[1] We use strong typicality arguments in the proof, where the set of all x^n strongly ϵ-typical with X is denoted by $T^n_{[X]_\epsilon}$. See [9] for further details.

for $j = 1, 2$, and

$$E_2(w_1, w_2) \triangleq \left\{ \left(T^n, U_1^n(I_1(w_1)), U_2^n(I_2(w_2))\right) \in T^n_{[TU_1U_2]_\epsilon} \right\}.$$

The probability of error can be bounded as follows:

$$\begin{aligned} P_e^n &\leq P[E_{11}^c(w_1)] + P[E_{12}^c(w_2)] \\ &\quad + P[E_2^c(w_1, w_2)|E_{11}(w_1), E_{12}(w_2)] \\ &\quad + \sum_{m_1 \neq w_1} P[E_2(m_1, w_2)] + \sum_{m_2 \neq w_2} P[E_2(w_1, m_2)] \\ &\quad + \sum_{m_1 \neq w_1, m_2 \neq w_2} P[E_2(m_1, m_2)]. \end{aligned}$$

It is easy to see that $P[E_{1j}^c(w_j)]$, $j = 1, 2$, can be made arbitrarily small for a sufficiently large n if

$$\frac{1}{n} \log L_j \geq I(U_j; Y_j) + \frac{\epsilon}{2};$$

hence we set $L_j = 2^{n(R_j^c + \epsilon)}$. Similarly, we can also let $P[E_2^c(w_1, w_2)|E_{11}(w_1), E_{12}(w_2)]$ go to zero for sufficiently large n which follows from the Markov Lemma [10].

We also have,

$$\sum_{m_1 \neq w_1} P[E_2(m_1, w_2)] \leq M_1 2^{-n(I(T;U_1|U_2) - \frac{\epsilon}{2})}, \tag{6}$$

which can be made arbitrarily small for sufficiently large n if,

$$\frac{1}{n} \log M_1 \leq I(T; U_1|U_2) - \epsilon \leq R_1^i - \epsilon. \tag{7}$$

Similarly, we also need

$$\frac{1}{n} \log M_2 \leq I(T; U_2|U_1) - \epsilon \leq R_2^i - \epsilon. \tag{8}$$

Finally, for the last term in the error probability bound

$$\sum_{m_1 \neq w_1, m_2 \neq w_2} P[E_2(m_1, m_2)] \leq M_1 M_2 2^{-n(I(T;U_1,U_2) - \frac{\epsilon}{2})}, \tag{9}$$

can be made arbitrarily small for sufficiently large n if,

$$\frac{1}{n}(\log M_1 + \log M_2) \leq I(T; U_1, U_2) - \epsilon \leq R_1^i + R_2^i - \epsilon. \tag{10}$$

Since we can choose M_1 and M_2 such that (7), (8) and (10) are all satisfied, we have shown that $P_e^n \to 0$ as $n \to \infty$.

This proves that the average probability of error, averaged over the ensembles of codebooks, can be made arbitrarily small given $(R_1^c, R_2^c, R_1^i, R_2^i) \in \mathcal{R}_{in}$. Hence, there exists at least one code with arbitrarily small average probability of error. The convex hull $\bar{\mathcal{R}}_{in}$ is achieved based on the usual time-sharing arguments.

Next, we prove the outer bound, that is, $\mathcal{R} \subseteq \mathcal{R}_{out}$. Assume that $(R_1^c, R_2^c, R_1^i, R_2^i) \in \mathcal{R}$. Then, for any $\epsilon > 0$ and sufficiently large n, there exist enrollment and identification functions f_1, f_2 and g such that

$$L_j \leq 2^{nR_j^c} \text{ and } M_j \geq 2^{nR_j^i}, \tag{11}$$

for $j = 1, 2$, and $P_e^n < \epsilon$. We have

$$\begin{aligned} \log M_1 &= H(W_1|\mathbf{I}_1, \mathbf{I}_2, T^n) + I(W_1; \mathbf{I}_1, \mathbf{I}_2, T^n) && (12) \\ &\leq H(W_1|\hat{W}_1) + I(W_1; \mathbf{I}_1, \mathbf{I}_2, T^n) && (13) \\ &\leq 1 + P_e^n \log M_1 + I(W_1; \mathbf{I}_1, \mathbf{I}_2, T^n) && (14) \end{aligned}$$

where (13) follows since $\hat{W}_1$ is a deterministic function of $\mathbf{I}_1, \mathbf{I}_2$ and T^n; (14) follows from Fano's inequality. From here we can obtain

$$\begin{aligned} (1-\epsilon) \log M_1 - 1 &\leq I(W_1; \mathbf{I}_1, \mathbf{I}_2, T^n) && (15) \\ &= I(W_1; T^n|\mathbf{I}_1, \mathbf{I}_2) && (16) \\ &= H(W_1|\mathbf{I}_1, \mathbf{I}_2) - H(W_1|\mathbf{I}_1, \mathbf{I}_2, T^n) && (17) \\ &\leq H(W_1|\mathbf{I}_1, \mathbf{I}_2, W_2) - H(W_1|\mathbf{I}_1, \mathbf{I}_2, T^n, W_2) && (18) \\ &= I(W_1; T^n|W_2, \mathbf{I}_1, \mathbf{I}_2) && (19) \\ &\leq H(T^n|W_2, \mathbf{I}_2) - H(T^n|W_1, W_2, \mathbf{I}_1, \mathbf{I}_2) && (20) \\ &= H(T^n|I_2(W_2)) - H(T^n|I_1(W_1), I_2(W_2)) && (21) \end{aligned}$$

where (16) follows since W_1 is independent of the database entries $(\mathbf{I}_1, \mathbf{I}_2)$; (18) follows since W_1 is independent of W_2 and conditioning reduces entropy; and (21) follows since T^n is independent of $I_1(m_1)$ with $m_1 \neq W_1$ and $I_2(m_2))$ with $m_2 \neq W_2$.

We define, for $j = 1, 2$, $U_{j,i} \triangleq (T^{i-1}, I_j(W_j))$. Using this definition and (11), we obtain

$$\begin{aligned} (1-\epsilon)nR_1^i - 1 &\leq H(T^n|I_2(W_2)) - H(T^n|I_1(W_1), I_2(W_2)) \\ &= \sum_{i=1}^n \big[H(T_i|T^{i-1}, I_2(W_2)) \\ &\qquad - H(T_i|T^{i-1}, I_1(W_1), I_2(W_2))\big] && (22) \\ &\leq \sum_{i=1}^n [H(T_i|U_{2,i}) - H(T_i|U_{1,i}, U_{2,i})] && (23) \\ &= \sum_{i=1}^n [I(T_i; U_{1,i}|U_{2,i})]. && (24) \end{aligned}$$

Hence, we have, for $(j, k) \in \{(1, 2), (2, 1)\}$

$$(1-\epsilon)R_j^i \leq \frac{1}{n} \sum_{i=1}^n [I(T_i; U_{j,i}|U_{k,i})] + \frac{1}{n}. \tag{25}$$

For the total rate of identification we have

$$\begin{aligned} \log M_1 M_2 &= H(W_1, W_2) && (26) \\ &= H(W_1, W_2|\mathbf{I}_1, \mathbf{I}_2, T^n) + I(W_1, W_2; \mathbf{I}_1, \mathbf{I}_2, T^n) && (27) \\ &\leq H(W_1, W_2|\hat{W}_1, \hat{W}_2) + I(W_1, W_2; \mathbf{I}_1, \mathbf{I}_2, T^n) && (28) \\ &\leq 1 + P_e^n \log M_1 M_2 + I(W_1, W_2; \mathbf{I}_1, \mathbf{I}_2, T^n) && (29) \end{aligned}$$

where (29) follows from Fano's inequality. Then, we can write

$$\begin{aligned} (1-\epsilon) \log M_1 M_2 - 1 &\leq I(W_1, W_2; \mathbf{I}_1, \mathbf{I}_2, T^n) && (30) \\ &= I(W_1, W_2; T^n|\mathbf{I}_1, \mathbf{I}_2) && (31) \\ &\leq H(T^n) - H(T^n|W_1, W_2, \mathbf{I}_1, \mathbf{I}_2) && (32) \\ &= H(T^n) - H(T^n|I_1(W_1), I_2(W_2)) && (33) \\ &= \sum_{i=1}^n \big[H(T_i|T^{i-1}) - H(T_i|T^{i-1}, I_1(W_1), I_2(W_2))\big] \\ &\leq \sum_{i=1}^n [H(T_i) - H(T_i|U_{1,i}, U_{2,i})] && (34) \\ &= \sum_{i=1}^n [I(T_i; U_{1,i}, U_{2,i})]. && (35) \end{aligned}$$

where (31) follows since W_1 and W_2 are independent of the database entries $(\mathbf{I}_1, \mathbf{I}_2)$; and (33) follows since T^n is independent of $I_1(m_1)$ with $m_1 \neq W_1$ and $I_2(m_2)$ with $m_2 \neq W_2$. Finally, we can obtain

$$(1-\epsilon)(R_1^i + R_2^i) \leq \frac{1}{n}\sum_{i=1}^{n}\left[I(T_i; U_{1,i}, U_{2,i})\right] + \frac{1}{n}. \qquad (36)$$

We need to show that $U_{1,i}$ and $U_{2,i}$ satisfy the Markov chains in the outer bound. We show that $U_{1,i} - Y_{1,i}(W_1) - X_{1,i}(W_1) - Z_i - T_i$ form a Markov chain. Since we already know that $Y_{1,i}(W_1) - X_{1,i}(W_1) - Z_i - T_i$ and $U_{1,i} - Y_{1,i}(W_1) - X_{1,i}(W_1)$ form Markov chain relationships, it is sufficient to show that $U_{1,i} - (Y_{1,i}(W_1), X_{1,i}(W_1)) - Z_i$ and $U_{1,i} - (Y_{1,i}(W_1), X_{1,i}(W_1), Z_i) - T_i$ form two Markov chains.

$$\begin{aligned}
&I(U_{1,i}; Z_i | Y_{1,i}(W_1), X_{1,i}(W_1)) = H(U_{1,i}|Y_{1,i}(W_1), X_{1,i}(W_1)) \\
&\qquad\qquad - H(U_{1,i}|Y_{1,i}(W_1), X_{1,i}(W_1), Z_i), \\
&= H(T^{i-1}, I_1(W_1)|Y_{1,i}(W_1), X_{1,i}(W_1)) \\
&\quad - H(T^{i-1}, I_1(W_1)|Y_{1,i}(W_1), X_{1,i}(W_1), Z_i), \\
&= H(T^{i-1}|Y_{1,i}(W_1), X_{1,i}(W_1)) \\
&\quad + H(I_1(W_1)|T^{i-1}, Y_{1,i}(W_1), X_{1,i}(W_1)) \\
&\quad - H(T^{i-1}|Y_{1,i}(W_1), X_{1,i}(W_1), Z_i) \\
&\quad - H(I_1(W_1)|T^{i-1}, Y_{1,i}(W_1), X_{1,i}(W_1), Z_i), \\
&= H(I_1(W_1)|T^{i-1}, Y_{1,i}(W_1), X_{1,i}(W_1)) \\
&\quad - H(I_1(W_1)|T^{i-1}, Y_{1,i}(W_1), X_{1,i}(W_1), Z_i), \\
&= 0, \qquad (37)
\end{aligned}$$

where the last equality follows since $T_i - Z_i - (X_{1i}, Y_{1,i}, T^{i-1}) - I_1(W_1)$. Similarly, it can be shown that $I(U_{1,i}; T_i | Y_{1,i}(W_1), X_{1,i}(W_1), Z_i) = 0$.

Next, we bound the enrollment rates. For $j = 1, 2$, we have the following:

$$\begin{aligned}
nR_j^c &\geq \log L_j \\
&\geq H(I_j(W_j)) \\
&= H(I_j(W_j)) - H(I_j(W_j)|Y_j^n(W_j)) \\
&= I(I_j(W_j); Y_j^n(W_j)) \\
&= H(Y_j^n(W_j)) - H(Y_j^n(W_j)|I_j(W_j)) \\
&= \sum_{i=1}^{n}\left[H(Y_{j,i}(W_j)) - H(Y_{j,i}(W_j)|Y_j^{i-1}, I_j(W_j))\right] \\
&= \sum_{i=1}^{n}\left[H(Y_{j,i}(W_j)) - H(Y_{j,i}(W_j)|Y_j^{i-1}, I_j(W_j), T^{i-1})\right] \qquad (38) \\
&\geq \sum_{i=1}^{n}\left[H(Y_{j,i}(W_j)) - H(Y_{j,i}(W_j)|I_j(W_j), T^{i-1})\right] \\
&\geq \sum_{i=1}^{n} I(Y_{j,i}(W_j); U_{j,i})
\end{aligned}$$

where (38) follows from the fact that $Y_{j,i}(W_j) - (Y_j^{i-1}, I_j(W_j)) - T^{i-1}$ forms a Markov chain.

Next, we introduce a time-sharing random variable Q independent of all the other random variables of interest and uniformly distributed over $\{1, \ldots, n\}$. We can rewrite (25) as

$$\begin{aligned}
(1-\epsilon)R_1^i &\leq \left[I(T_Q; U_{1,Q}|U_{2,Q}, Q)\right] + \frac{1}{n}. \\
&= \left[I(T; U_1|U_2, Q)\right] + \frac{1}{n},
\end{aligned}$$

where we defined the new random variables as $T \triangleq T_Q$, $U_1 \triangleq U_{1,Q}$ and $U_2 \triangleq U_{2,Q}$. Following similar steps to those for (36) and letting $n \to \infty$ and $\epsilon \to 0$, we obtain

$$R_1^i \leq I(T; U_1|U_2, Q), \qquad (39a)$$
$$R_2^i \leq I(T; U_2|U_1, Q) \text{ and} \qquad (39b)$$
$$R_1^i + R_2^i \leq I(T; U_1, U_2|Q). \qquad (39c)$$

Also defining $Y_1 \triangleq Y_{1,Q}$ and $Y_2 \triangleq Y_{2,Q}$, we obtain

$$R_1^c \geq I(Y_1; U_1, |Q) \text{ and} \qquad (40a)$$
$$R_2^c \geq I(Y_2; U_2, |Q). \qquad (40b)$$

It is possible to show that the set of rate points satisfying (39) and (40) is equivalent to $\bar{\mathcal{R}}_{out}$.

V. Conclusions

We have studied the tradeoff between the storage and the identification rates over multiple databases. We have considered the joint identification of ancestors over two separate databases, which consist of the compressed noisy observations of the data vectors. We have presented single-letter inner and outer bounds on the set of achievable rate points, which identify a tradeoff between the compression rates and the identification rate region; the lower the compression rates for the enrollment process, the larger the identification rate region.

References

[1] E. Tuncel, P. Koulgi, and K. Rose, "Rate-distortion approach to databases: Storage and content-based retrieval," *IEEE Trans. Inform. Theory*, vol. 50, no. 6, pp. 953-967, June 2004.

[2] F. Willems, T. Kalker, J. Goseling and J.-P. Linnartz, "On the capacity of a biometrical identification system," *Proc. IEEE Int'l Symp. Inform. Theory*, Yokohama, Japan, July 2003.

[3] M.B. Westover and J.A. O'Sullivan, "Achievable rates for pattern recognition," *IEEE Trans. Inform. Theory*, vol. 54, no. 1, pp. 299-320, Jan. 2008.

[4] E. Tuncel, "Capacity/storage tradeoff in high-dimensional identification systems," to appear, *IEEE Trans. Inform. Theory*.

[5] S. A. Krawetz and D. D. Womble, *Introduction to Bioinformatics: A Theoretical and Practical Approach*, Totowa, NJ: Humana Press, 2003.

[6] C. Tian and J. Chen, "Successive refinement for hypothesis testing and lossless one-helper problem," *IEEE Trans. Inform. Theory*, vol. 54, no. 10, pp. 4666-4681, Oct. 2008.

[7] N. Z. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," *Proc. 37th Allerton Conf. Commun., Control and Computing*, Monticello, IL, Sep. 1999.

[8] N. Slonim, N. Friedman and N. Tishby, "Multivariate information bottleneck," *Neural Comput.*, vol. 18, no. 8, pp. 1739-1789, Aug. 2006.

[9] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, New York: Academic, 1981.

[10] T. Berger, "Multiterminal source coding," *Lectures presented at CISM Summer School on the Inform. Theory Approach to Commun.*, July 1977.