

Identification over Multiple Databases

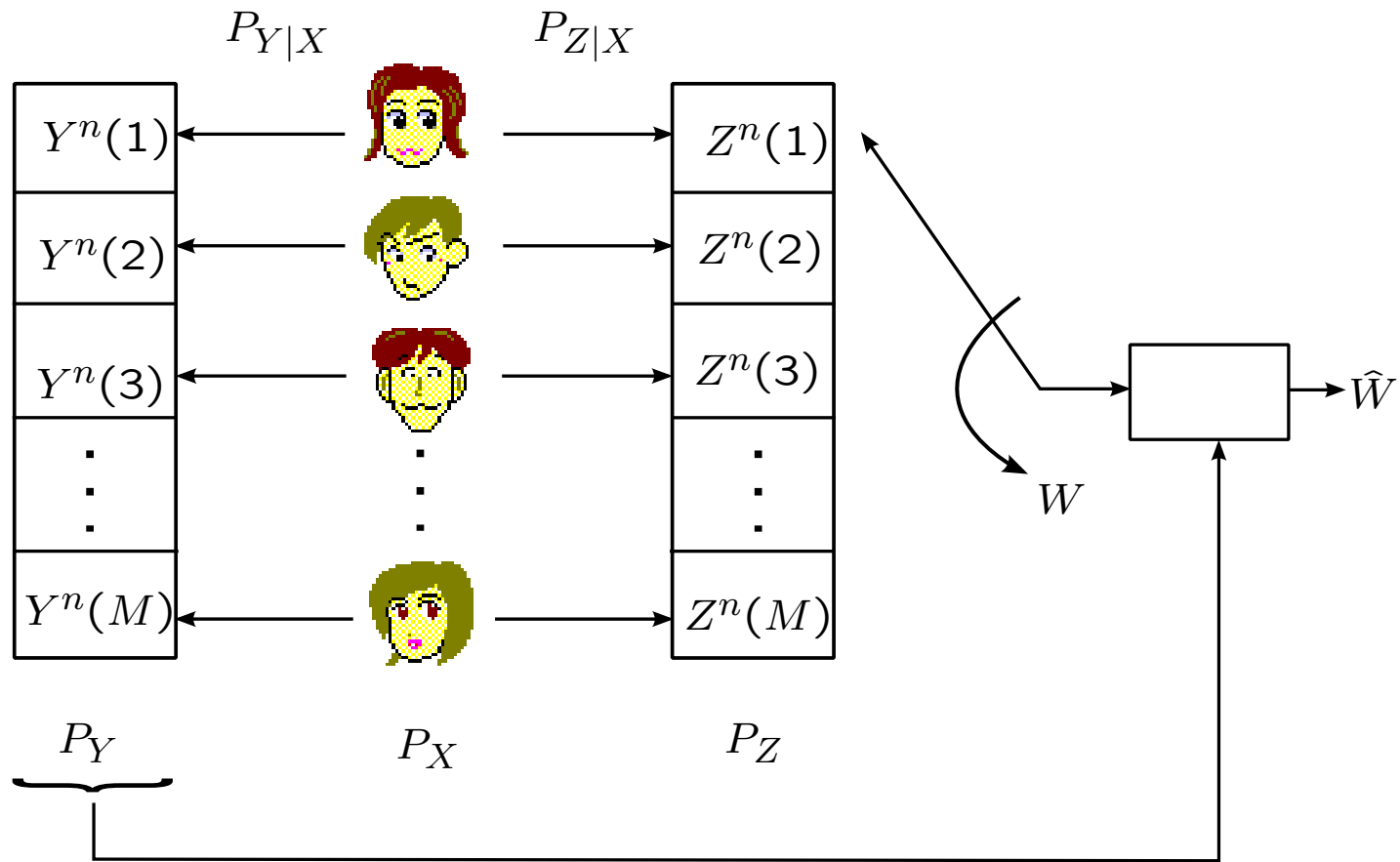
Deniz Gündüz^{†‡}, Ertem Tuncel[#], Andrea Goldsmith[†],
H. Vincent Poor[‡]

[†] Stanford University, [‡] Princeton University,
[#] University of California, Riverside

Motivation

- Volume of data keeps growing: genomic information, social networks, online data, marketing information, ...
- We need efficient storage and quick search methods
- Fundamental limits?

Identification over a Biometric Database



How many individuals can be identified reliably?

Capacity of an Identification System

- Error probability:

$$P_e = Pr\{\hat{W} \neq W\}$$

- Identification rate:

$$R^i = \frac{1}{n} \log M$$

- **Capacity** is C if for any $\epsilon > 0$, there exists large enough n s.t.

$$R^i \geq C - \epsilon$$

$$P_e \leq \epsilon$$

Theorem (Willems et al., ISIT'03) The capacity of the identification system is $I(Y; Z)$.

Capacity/Storage Tradeoff

- Compress observed feature vectors before storage
- Expedites identification process
- Reduces storage requirement
- But, degrades identification capacity
- Compression function: $f : \mathcal{Y}^n \rightarrow \mathcal{L} = \{1, \dots, L\}$
- Compression rate:

$$R^c = \frac{1}{n} \log L$$

- Tradeoff between R^c (compression) and R^i (identification)

Capacity/Storage Tradeoff

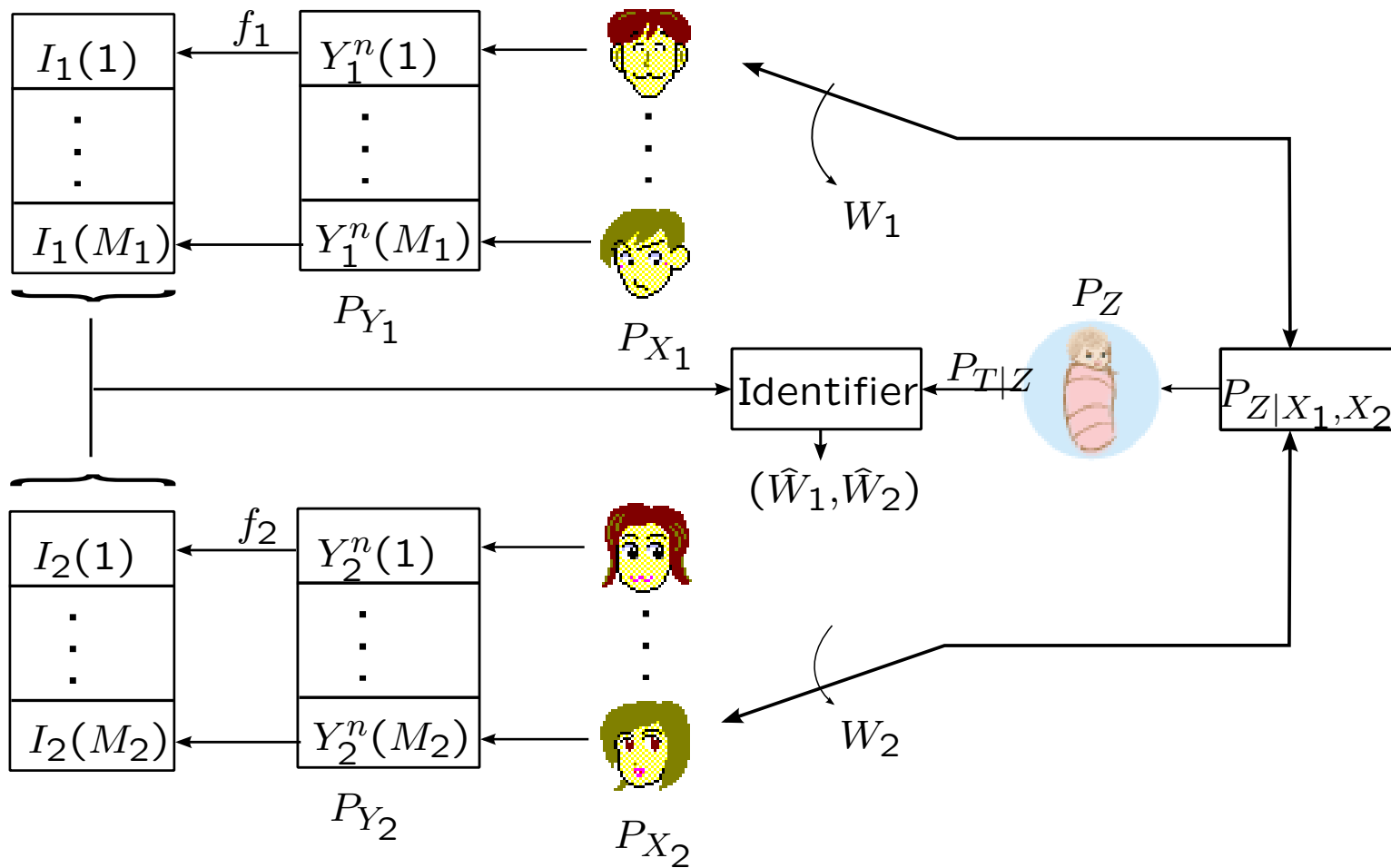
Theorem (Westover and O'Sullivan, ISIT'04, Tuncel, ISIT'06)
 (R^c, R^i) is an achievable compression/identification rate pair iff,
there exists an auxiliary random variable U such that

i) $U - Y - X - Z$ forms a Markov chain,

ii) $I(U; Y) \leq R^c$

iii) $I(U; Z) \geq R^i$

Identification over Multiple Databases



Identification over Multiple Databases

$(R_1^c, R_2^c, R_1^i, R_2^i)$ 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 such that

$$\begin{aligned}\frac{1}{n} \log L_j &\leq R_j^c \\ \frac{1}{n} \log M_j &\geq R_j^i,\end{aligned}$$

for $j = 1, 2$, and $P_e^n \leq \epsilon$.

Main Result

For a given joint distribution

$$P_{X_1, X_2, Y_1, Y_2, Z, T} = P_{X_1} P_{X_2} P_{Y_1|X_1} P_{Y_2|X_2} P_{Z|X_1, X_2} P_{T|Z}$$

define two sets:

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

and

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

Main Result

For a given pair of (U_1, U_2) , define the rate region:

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

Main Result

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

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

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

Achievability

Codebook generation: For $j = 1, 2$, generate codebooks of L_j length- n codewords i.i.d. with distribution p_{U_j} .

Enrollment: Given $y_j^n \in \mathcal{Y}_j^n$, define f_j as the smallest index l_j such that $(y_j^n, U_j^n(l_j)) \in T_{[U_j Y_j]_\epsilon}^n$. Set $f_j(y_j^n) = 1$ if no such codeword exists.

Identification: Given any $t^n \in \mathcal{T}^n$ and the database entries, define the identification function as the smallest pair of indices w_1 and w_2 such that $(t^n, U_1^n(I_1(w_1)), U_2^n(I_2(w_2))) \in T_{[TU_1 U_2]_\epsilon}^n$. We set $g(t^n, \mathbf{I}_1, \mathbf{I}_2) = (1, 1)$ if no such pair can be found.

Information Bottleneck Problem

- Proposed by Tishby et al., Allerton'99
- Tradeoff between accuracy (relevant information) and complexity (compression)
- For given joint distribution P_{XY} ,

$$R^{IB}(\tau) = \min_{\substack{p(u|y): I(U; X) \geq \tau, \\ U-Y-X}} I(U; Y)$$

- Found applications in document clustering, feature selection, learning, gene expression data analysis, and protein sequence analysis

- For single database system, minimum storage required for given identification rate R^i is

$$C(R^i) = \min_{\substack{p(u|y): I(U;Z) \geq R^i, \\ U-Y-X-Z}} I(U;Y)$$

- Relations with hypothesis testing established in Tian and Chen, IT'08

Multiple Databases

- Consider the total identification rate
- Achievable rate region is equivalent to a multivariate information bottleneck problem (Slonim et al., Neural Comp.'06)
- Extension: Multiple databases for noisy observations of a single phenomena

Conclusions

- Analyzed the fundamental tradeoff between identification capacity and required storage for multiple databases
- Single-letter inner and outer bounds