Statistical Analysis and Modeling of Content Identification and Retrieval

Pierre Moulin

University of Illinois at Urbana-Champaign

Electrical and Computer Engineering

CUHK Information Engineering Dept

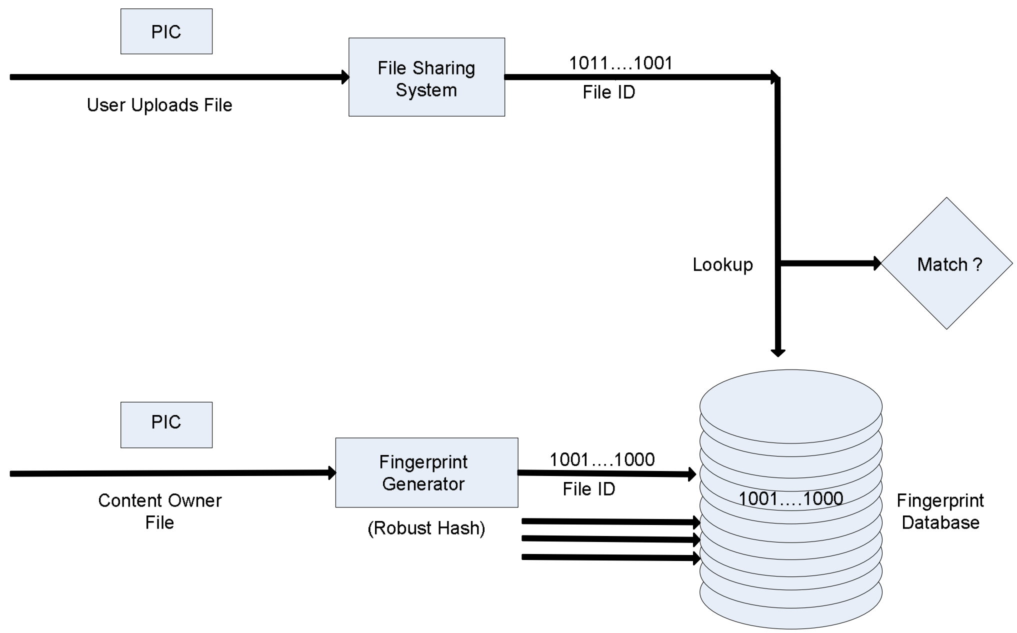
January 26, 2010

University of Illinois at Urbana-Champaign

- Beckman Institute and Coordinated Science Laboratory

Content Identification

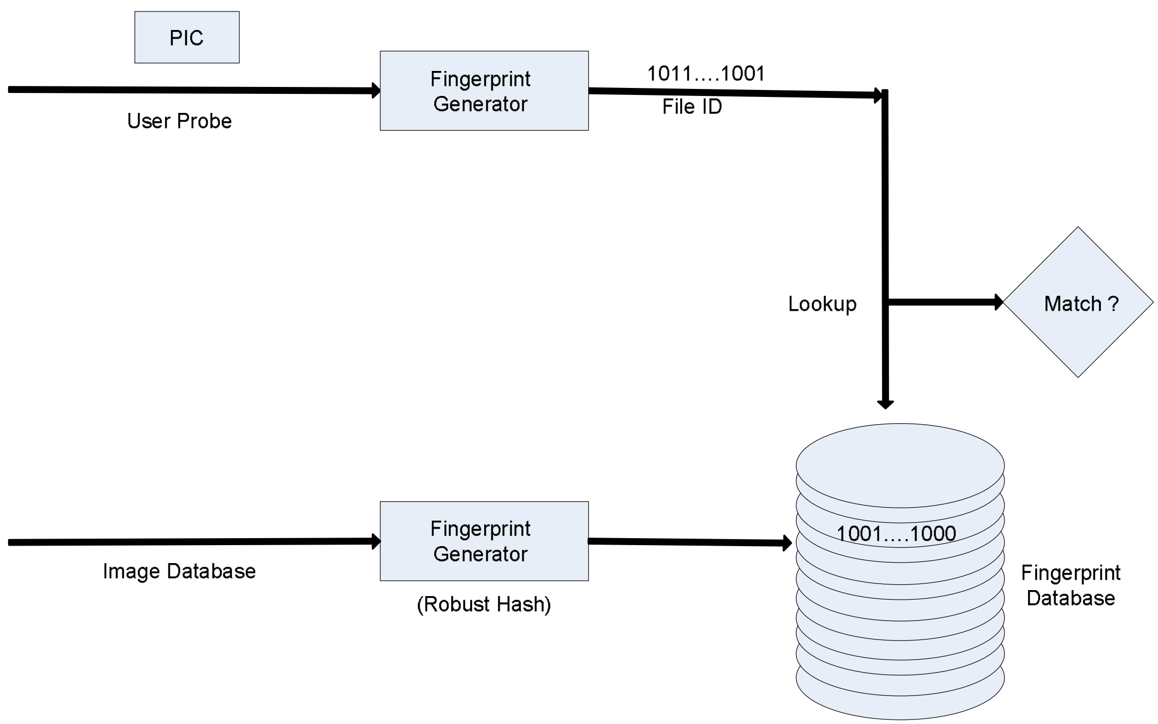
- YouTube & other User Generated Content (UGC) sharing sites
- registration of copyrighted content → fingerprint database



- Related application: connected audio (Shazaam on I-phones)

Content Retrieval

- User seeks similar contents (audio, video) in large database
- Can search based on fingerprints/hashes



Cryptographic vs. Robust Hashes

- A cryptographic hash function $\Phi_K : \mathcal{X} \rightarrow \{0,1\}^k$ satisfies the following property:

$$Pr_K[\Phi_K(x) = \Phi_K(x')] = 2^{-k} \quad \forall x \neq x'$$

- In contrast, a robust hash function should return the same hash if x and x' are "perceptually similar":

$$\begin{aligned} Pr_K[\Phi_K(x) = \Phi_K(x')] &> 1 - \epsilon \quad \forall x \sim x' \\ Pr_K[\Phi_K(x) = \Phi_K(x')] &< \epsilon \qquad\quad \forall x \not\sim x' \end{aligned}$$

Formulation of Content ID Problem

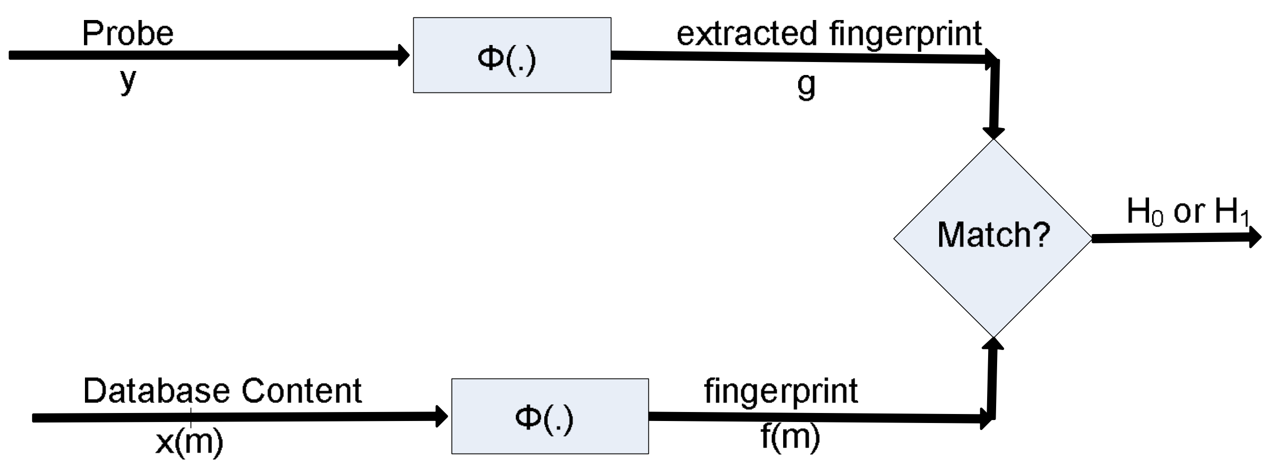
- Content database = $\{\mathbf{x}(m),\ 1 \le m \le M\}$
- Each $\mathbf{x}(m) = \{x_1(m), x_2(m), \cdots, x_N(m)\} \in \mathcal{X}^N$ is a collection of N frames. For audio ID,
  - frames are short audio snippets (370 msec) with 31/32 temporal overlap.
  - A 3-minute song is represented by $N \approx 15,500$ frames
  - desired granularity $\approx$ 3 sec ($L = 258$ frames)
- Probe $\mathbf{y} \in \mathcal{X}^L$ consisting of $L \ll N$ frames
- Is the probe related to one of the database elements?
- Construct $\psi(\mathbf{y}) \in \{0, 1, 2, \cdots, M\}$

Performance Metrics

- Probability of false positives
- Probability of false negatives
- Robustness
- Granularity
- Database size
- Storage requirements
- Execution time

Fingerprint-Based Content ID

- Hash function Φ returns fingerprint $\mathbf{f}(m)$ for each input $\mathbf{x}(m)$ and fingerprint $\mathbf{g}$ for input probe $\mathbf{y}$
- Decisions are made based on fingerprints only



Research Challenges

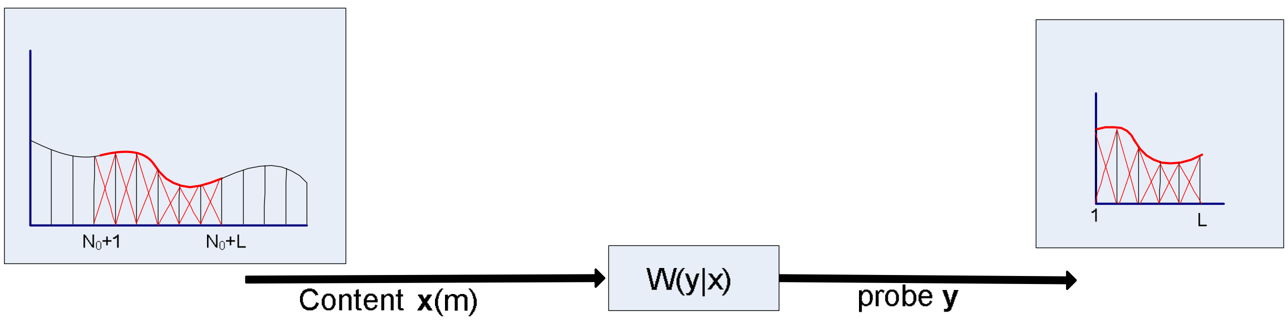
- signal processing primitives for robust hashes
- efficient string matching algorithms
- information-theoretic challenge: what is the fundamental relation between database size, hash length, and robustness?
- general framework for hash function design

Statistical Model for Content Database

- Database elements $\mathbf{x}(m)$, $1 \le m \le M$ are drawn independently from stationary probability distribution $P_{\mathbf{X}}$ on $\mathcal{X}^N$

Statistical Model for Probe

- $M+1$ hypotheses $H_0, \cdots, H_M$
- under $H_m,\ 1 \le m \le M$:



$$W^L(\mathbf{y}|\mathbf{x}(m), N_0) \triangleq \prod_{i=1}^{L} W(y_i|x_{i+N_0}(m))$$

and N_0 is drawn uniformly from $\{0, 1, \cdots, N-L-1\}$

- under H_0, probe $\mathbf{y}$ is drawn from same $P_{\mathbf{Y}}$

Statistical Model for Hash Function

- Let $\mathbf{F} = \phi(\mathbf{X}) \in \mathcal{F}^N$ and $\mathbf{G} = \phi(\mathbf{Y})$ where $|\mathcal{F}| \ll |\mathcal{X}|$
- Fingerprint storage cost $\leq N \log |\mathcal{F}|$ bits
- Assume
  - the samples F_i, $1 \leq i \leq N$ are iid with pmf p_F
  - the conditional pmf of $\mathbf{g}$ given $\mathbf{f}(m)$ and N_0 is

$$p^L_{G|F}(\mathbf{g}|\mathbf{f}(m), N_0) \triangleq \prod_{i=1}^{L} p_{G|F}(g_i|f_{i+N_0}(m))$$

$\Rightarrow$ the pairs (F_i, G_i), $1 \leq i \leq L$ are iid with pmf p_{FG}

- If $\mathbf{F}(m)$ and $\mathbf{G}$ are independent, then
  the pairs (F_i, G_i), $1 \leq i \leq L$ are iid with product pmf $p_F p_G$

General Definition of Content ID Code

- A (M, N, L) content ID code for a size-M database populated with $\mathcal{X}^N$-valued content items, and granularity L, is a pair consisting of an encoding function $\phi : \mathcal{X}^N \to \mathcal{F}^N$ returning a fingerprint $\mathbf{f} = \phi(\mathbf{x})$, and a constrained decoding function $\psi : \mathcal{X}^L \to \{0, 1, \cdots, M\}$ returning $\hat{m} = \psi(\mathbf{y})$, where the dependency on input $\mathbf{y}$ is via the fingerprint $\phi(\mathbf{y})$.
- The rate of the code is

$$R \triangleq \frac{1}{L} \log(MN)$$

  (**fundamental scaling parameter**)
- Neither M nor N necessarily dominates

List Decoder

- Define decoding metric $d(f, g)$ on $\mathcal{F} \times \mathcal{F}$
- Extend additively to sequences:

$$d(\mathbf{f}, \mathbf{g}|N_0) = \sum_{i=1}^{L} d(f_{i+N_0}, g_i)$$

- Choose decision threshold τ
- Decoder outputs list $\mathcal{L}$ of all m such that

$$\min_{0 \leq N_0 < N-L} d(\mathbf{f}(m), \mathbf{g}|N_0) < L\tau$$

Error Analysis for List Decoder

- Wlog assume $m = 1$
- Error event #1: **Miss**: The correct m does not appear on the decoder's list:

$$\forall N_0 \in \{0, \cdots, N\text{–}L-1\} \;:\quad d(\mathbf{f}(1), \mathbf{g}|N_0) > L\tau.$$

- Error event #2: **Incorrect Decoding**:

$$\exists m > 1,\; N_0 \in \{0, \cdots, N\text{–}L-1\} : \quad d(\mathbf{f}(m), \mathbf{g}|N_0) < L\tau$$

Let N_{i} = number of incorrect messages on the list

- Consider performance metrics P_{miss} and $\mathbb{E}[N_{\mathrm{i}}]$

Error Analysis for List Decoder (Cont'd)

- Wlog, assume $M = 1$. Then

$$\begin{aligned}
\mathbb{E}[N_{\mathrm{i}}] &= M\, Pr\left[\min_{0 \le N_0 < N-L} d(\mathbf{F}(2), \mathbf{G}|N_0) < L\tau\right] \\
&\le M(N-L) \max_{0 \le N_0 < N-L} Pr\left[d(\mathbf{F}(2), \mathbf{G}|N_0) < L\tau\right] \\
&= M(N-L) Pr\left[d(\mathbf{F}(2), \mathbf{G}|N_0 = 0) < L\tau\right] \\
&= M(N-L)\ \underbrace{P_F^L P_G^L \left[\sum_{i=1}^{L} d(F_i, G_i) < L\tau\right]}_{=?}
\end{aligned}$$

Large-Deviations Bounds on Error Probabilities

- Give iid random variables v_i, $1 \le i \le L$ with distribution P_V, a function h, and a threshold τ, evaluate

$$p \triangleq P_V^L \left[\sum_{i=1}^{L} h(v_i) < L\tau \right]$$

- Large-deviations bound:

$$p \le 2^{-LE(\tau)}$$

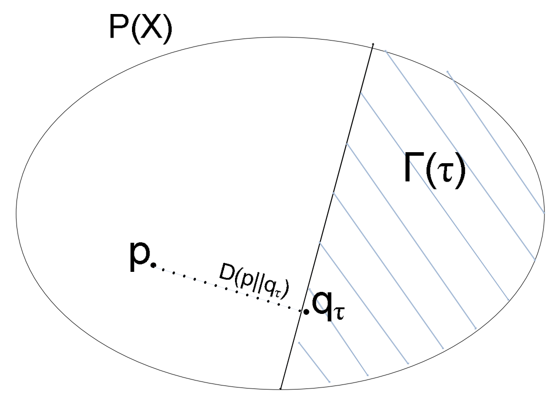
where

$$E(\tau) = \min_{Q \in \Gamma(\tau)} D(P_V \| Q)$$

and

$$\Gamma(\tau) \triangleq \{Q \ : \ \sum_{v} Q(v)h(v) < \tau\}$$

- Geometric view of $E(\tau) = \min_{Q \in \Gamma(\tau)} D(P_V \| Q)$:



Error Exponents

- For any sequence of (M, N, L) content ID codes such that $\lim \frac{1}{L} \log(MN) = R$, define the miss exponent

$$E_{miss}(P_F, P_{G|F}, \tau) = \liminf_{L\to\infty} -\frac{1}{L} \ln P_{miss}$$

and the incorrect-item exponent

$$E_{\mathrm{i}}(P_F, P_{G|F}, R, \tau) = \liminf_{L\to\infty} -\frac{1}{L} \ln \mathbb{E}[N_{\mathrm{i}}]$$

- Define convex set of pmf's over $\mathcal{F}^2$:

$$\Gamma(\tau) \triangleq \{Q \;:\; \sum_{f,g\in\mathcal{F}} Q(f,g)d(f,g) < \tau\}$$

- We have

$$E_{miss}(P_F, P_{G|F}, \tau) \;=\; \min_{P'_{FG}} \left[D(P'_{FG} \| P_F\, P_{G|F}) + \min_{Q\in\Gamma^c(\tau)} D(P'_{FG} \| Q) \right]$$

$$E_{\mathrm{i}}(P_F, P_{G|F}, R, \tau) \;=\; \min_{P'_{FG}} \left[D(P'_{FG} \| P_F\, P_G) + \min_{Q\in\mathring{\Gamma}(\tau)} D(P'_{FG} \| Q) - R \right]$$

Achievable Rates

- Define the set of conditional distributions

$$\begin{aligned}\mathscr{P}'_{G|F} \triangleq \{P'_{G|F} : P'_G = P_G, \\ \mathbb{E}_{P_F P'_{G|F}} d(F,G) = \mathbb{E}_{P_{FG}} d(F,G)\}\end{aligned}$$

  and the *generalized mutual information*

$$I_{\mathrm{GMI}}(P_F, P_{G|F}, d) \triangleq \min_{P'_{G|F} \in \mathscr{P}'_{G|F}} D(P_F P'_{G|F} \| P_F P_G)$$

  which also appears in information-theoretic analyses of channel capacity with mismatched decoders

- **Proposition**: The supremum of the values of R for which the error exponents are positive is $R = I_{\mathrm{GMI}}(P_F, P_{G|F}, d)$ and is achieved when $\tau = \mathbb{E}_{P_{FG}} d(F,G)$.

Matched Decoding

- If $p_{G|F}$ is known, choose

$$d(f,g) = -\log p_{G|F}(g|f) \quad \Rightarrow \quad I_{\text{GMI}} = I(F;G)$$

- Then the list decoder achieves positive error exponents for all

$$R < I(F;G)$$

- Converse?

Converse

- Recall $N_0 \in \{0, 1, \cdots, N-L-1\}$ = unknown nuisance parameter
- Is GLRT optimal?
- **Proposition:** For any sequence of of (M, N, L) content ID codes such that
$$\lim \frac{1}{L} \log M > I(F;G),$$
the average error probability $\overline{P}_e$ does not vanish.

  (Proof by Fano's inequality)
- This bound is unsatisfactory because
  - can achieve all $\frac{1}{L} \log M < I(F;G) - \frac{1}{L} \log N \quad \Rightarrow$ gap!
  - $\overline{P}_e$ criterion gives vanishing weight to H_0

Strong Converse

- Max error criterion:

$$P_{e,\max} \triangleq \max_{0 \le m \le M} Pr[\psi(\mathbf{Y}) \neq m \,|H_m]$$

- **Proposition:** For any sequence of of (M, N, L) content ID codes such that

$$\lim \frac{1}{L} \log(MN) > I(F; G),$$

  $P_{e,\max}$ tends to 1

- Lower and upper bounds now coincide