Information-Theoretical Analysis of Private Content Identification

S. Voloshynovskiy, O. Koval, F. Beekhof, F. Farhadzadeh, T. Holotyak
Stochastic Information Processing Group,
University of Geneva,
Switzerland
Outline

1. Introduction
2. Identification setup
3. Error events
4. Identification capacity and privacy leak
5. Complexity of fingerprint-based identification
6. Conclusions
Introduction

Physical Objects
- watches
- packaging

Humans

Digital Content
- Images
- Videos
- Audios
- Text docs

Online sharing services

Main concerns
- Identification: identity, authenticity, origin, ownership, ...
- Tracking and tracing
- Automatic tagging
Digital fingerprinting (a.k.a. robust perceptual hashing) is a technique for computing a compact robust, secure and private binary representation of physical or digital content.
Introduction

Objectives

- Performance analysis (probability of error, achievable rate);
- Privacy leak evaluation;
- Complexity for large-scale applications.

Related works: F. Willems, T. Kalker, J. Goseling, and J.-P. Linnartz, ISIT2003
Identification Setup (1)

**Enrollment**

\[ X(m) \in \mathcal{X}^N, X \sim p(x) \]

\[ X' \in \mathcal{X}^M \]

\[ \tilde{X} = WX \]

\[ W \in \mathbb{R}^{L \times N}, W_{ij} \sim \mathcal{N}\left(0, \frac{1}{N}\right) \]

\[ B_x = \text{sign}(\tilde{X}) \]

\[ B_x \sim \mathcal{B}\left(L, \frac{1}{2}\right) \]

\[ P_y|x \]

**Identification**

\[ Y \in \mathcal{Y}^N \]

\[ \psi_y \]

\[ B_y \in \{0,1\}^L \]

\[ \text{Decoder} \]

\[ \hat{m} \]

**Codebook cardinality:** \( M = 2^{LR} \) with \( R \) being the rate.

**W** - dim. reduction (sensing matrix)  
**B** - binarization  
**PA** - privacy amplification
**Identification Setup (2)**

**Identification as composite hypothesis testing**

\[
\begin{align*}
\mathcal{H}_0 &: p(b_y | \mathcal{H}_0) = p(b_y | b'_x), \\
\mathcal{H}_m &: p(b_y | \mathcal{H}_m) = p(b_y | b_u(m)), m = 1, \ldots, M
\end{align*}
\]

**Binary fingerprinting**

\[
\begin{align*}
\mathcal{H}_0 &: p(b_y | b'_x) = \frac{1}{2L}, \\
\mathcal{H}_m &: p(b_y | b_u(m)) = P_{b_e}^{d^R(b_y, b_u(m))} (1 - P_{b_e})^{L-d^R(b_y, b_u(m))}
\end{align*}
\]

\[
P_{b_e} = P_b \ast \lambda = P_b (1 - \lambda) + \lambda (1 - P_b)
\]

- \(\lambda\) - enrollment bit error probability
- \(P_b\) - identification bit error probability
Forney’s erasure/list decoder [Forney’68]

\[ \tau \leq -H_2(P_{b_e}) \text{ - binary entropy.} \]

\[ p(b_y | b_u(m)) \geq 2^{\tau L} \]

Binary fingerprinting – Bounded Distance Decoder (BDD)

\[ d^H(b_y, b_u(m)) \leq L\gamma \]

with \( \gamma = \frac{-\tau + \log_2(1 - P_{b_e})}{\log_2 \frac{1 - P_{b_e}}{P_{b_e}}} \)
Properties

Correct acceptance

\[ x(m) \quad y \]

\[ \Pr \left[ D^H(b_u(m), b_y) \leq \gamma L \right] \to 1 \]

Binomial distribution

\[ D^H(b_u(m), b_y) \sim B(L, P_b) \]

Correct rejection

\[ x(m) \quad x(m') \]

\[ \Pr \left[ D^H(b_u(m), b_{x'(m')}) \leq \gamma L \right] \to 0 \]

Binomial distribution

\[ D^H(b_u(m), b_y) \sim B \left( L, \frac{1}{2} \right) \]
Identification Setup (5)

- **Robustness**
  - Errorless recovery of fingerprint index
  - Probability of error
    \[
    \Pr[\hat{M} \neq M] \leq \varepsilon
    \]

- **Security**
  - Ability of attacker to learn key \(K\) (confusion)
  - Secrecy leak
    \[
    I(K;B_u) \leq LL_K
    \]

- **Privacy**
  - Ability of attacker to learn private data \(X\) (diffusion)
  - Privacy leak
    \[
    I(B_x;B_u) \leq LL_p
    \]

- **Information-theoretic rate**
  - Number of MM contents that can be uniquely fingerprinted
    \[
    2^{LR}
    \]
  - Achievable rate
    \[
    I(B_y;B_u) = LR
    \]

- **Algorithmic complexity**
  - Search in large database (exponential in input length)
    \[
    O(2^L)
    \]
  - approximate search
Proposition 1

The optimal threshold $\tau$ for unique content identification under Forney’s erasure rule should satisfy $\tau \leq -H_2(P_e)$ to guarantee a minimum of overall identification error $P_e$.

Proof:

$$P_e = \frac{1}{2} P_f + \frac{1}{2} P_{ic}$$

$P_f$ - probability of false acceptance

$P_{ic}$ - probability of incorrect decoding

$$P_f(\gamma) = \Pr \left[ \bigcup_{m=1}^{M} d^H(B_u(m), B_y) \leq \gamma L \mid H_0 \right]$$

$$\leq 2^{-L(1-H_2(\gamma)-R)}$$

$$P_{ic}(\gamma) = \Pr \left[ d^H(B_u(m), B_y) > \gamma L \cup \bigcup_{n=m}^{M} d^H(B_u(n), B_y) \leq \gamma L \mid H_m \right]$$

$$\leq 2^{-LD(\gamma \mid \mid P_e)} + 2^{-L(1-H_2(\gamma)-R)}$$

$$D(\gamma \mid \mid P_e) = \gamma \log_2 \frac{\gamma}{P_e} + (1-\gamma) \log_2 \frac{1-\gamma}{1-P_e}$$

- divergence
Error events (2)

\[ P_e = \frac{1}{2} P_f + \frac{1}{2} P_{ic} \]

\[ \leq \frac{1}{2} \left( 2^{-LD(\gamma||P_{b_e})} + 2^{2L(1-H_2(\gamma)-R)} \right) \]

that is minimized by:

\[ \gamma_{opt} = \frac{1 - R + \log_2 \left( 1 - P_{b_e} \right) - 1 / L}{\log_2 \left( \frac{1 - P_{b_e}}{P_{b_e}} \right)} \]

Forney’s threshold

\[ \gamma = \frac{-\tau + \log_2 \left( 1 - P_{b_e} \right)}{\log_2 \left( \frac{1 - P_{b_e}}{P_{b_e}} \right)} \]

For large \( L \), \( \tau \leq -(1 - R) = -H_2(P_{b_e}) \) for the identification rates \( R \leq 1 - H_2(P_{b_e}) \).

Remark 1

For the identification rate satisfying \( R \leq 1 - H_2(P_{b_e}) \), the above optimal threshold yields \( \gamma_{opt} = P_{b_e} \).
Proposition 2

For $P_{b_e} \leq \gamma \leq \frac{1}{2}$ and if $H_2(\gamma) \leq 1 - R$ there exist codes with rate $R$ and error probability $P_e$ such that:

$$\lim_{L \to \infty} P_e = 0$$

As soon as $\gamma$ is arbitrarily close to $P_{b_e}$, the rate $R = 1 - H_2(P_{b_e})$ is achievable, and it is referred to as:

**private identification capacity:** $C_{id} = I(B_u; B_y) = 1 - H_2(P_{b_e})$,

**privacy leak:** $\mathcal{L}_p = I(B_u; B_x) = 1 - H_2(\lambda)$

Remark 2 /result coincides with F. Willems et al ISIT2003/

If privacy amplification is not applied, i.e., $\lambda = 0$ and $b_u = b_x$, one is interested in the maximization of the identification capacity that yields:

$$C_{id} = I(B_x; B_y) = 1 - H_2(P_b),$$

$$\mathcal{L}_p = I(B_u; B_x) = 1$$
Identification capacity and privacy leak (2)

\[ R_{id} \]

\[ \mathcal{L}_p \]

\[ P_b = 0.0 \]
\[ P_b = 0.1 \]
\[ P_b = 0.2 \]
\[ P_b = 0.3 \]
\[ P_b = 0.4 \]
Complexity of fingerprint based identification (1)

Decoding algorithms

- Exhaustive implementation of BDD
- Hamming sphere decoding
- Reliability-based decoding
**Complexity of fingerprint based identification (2)**

**Exhaustive implementation of BDD**

- Given: $b_u(m)$ codewords, $1 \leq m \leq M$ with $M \leq 2^{L(1-H_2(P_e))}$
- Compute: $d^H(b_y, b_u(m)) \leq L\gamma$
- Complexity: $O\left(\frac{L(1-H_2(P_e))}{2} \log_2 M\right)$ exponential

**Identification = exhaustive computation**

<table>
<thead>
<tr>
<th>Fingerprint</th>
<th>$b_y$</th>
<th>$d^H(b_y, b_u(m)) \leq L\gamma$</th>
<th>Codebook/Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td></td>
<td></td>
<td>$b_u(1)$, $b_u(2)$, $b_u(M)$</td>
</tr>
<tr>
<td>$K$</td>
<td></td>
<td></td>
<td>$b_y$, $b_u(1)$, $b_u(2)$, $b_u(M)$</td>
</tr>
</tbody>
</table>

**Remark 3**

Complexity does not depend on data quality (both privacy amplification and acquisition).
Complexity of fingerprint based identification (3)

Hamming sphere decoding

Observation: the most likely codewords $b_u(\hat{m})$ are within a Hamming sphere with radius $\gamma L$ around $b_y$.

Identification = codeword presence verification
Proposition 3

The cardinality of the list of candidates contained in the sphere of radius \( \gamma L \), where \( 0 \leq \gamma \leq \frac{1}{2} \), is:

\[
\sum_{t=0}^{\gamma L} \left( \begin{array}{c} L \\ t \end{array} \right) \leq 2^{LH_2(\gamma)}
\]

Asymptotically:

\[
\lim_{L \to \infty} \sum_{t=0}^{\gamma L} \left( \begin{array}{c} L \\ t \end{array} \right) \approx 2^{LH_2(\gamma)}
\]

For unique decoding \( \gamma = P_e \cdot 2^{LH_2(\gamma)} \rightarrow 2^{LH_2(P_e)} \)

Remark 4

Complexity depends on data quality \( P_e \).

Remark 5

Analogy: binary fingerprint = computer memory address with flag.
Reliability based decoding

Observation: the reliability of bits of $b_y$ within Hamming sphere can be estimated based on the fingerprint magnitude.

Random projections: impact of distortions

$$y = x + z$$

$$P_{b|x} = Q\left(\frac{|\bar{x}|}{\sigma_z}\right)$$

sign($\bar{x}_i$) $\neq$ sign($\bar{y}_i$)

Concept of random projection bit reliability (sign-magnitude decomp.)

$$\tilde{x} = \text{sign}(\bar{x})|\bar{x}|$$

$\tilde{x}'$
Identification = soft decoding or soft verification

Two-channel splitting:
- Good channel with $P_b \rightarrow 0$
- Bad channel with $P_b \rightarrow 0.5$

Multi-channel splitting, randomization and decoding
[Voloshynovskiy et al, IEEE WIFS2010]
Complexity of fingerprint based identification (7)

Comparison of decoding strategies
\[ \mathcal{O}(LM^{\alpha_1}) \rightarrow \mathcal{O}(LM^{\alpha_2}) \rightarrow \mathcal{O}(LM^{\alpha_3}) \]

- **Exhaustive search**
  \[ \alpha_1 = 1 \]

- **Hamming sphere decoding**
  \[ \alpha_2 = \frac{H_2(P_{b_e})}{1 - H_2(P_{b_e})} \]

- **Reliability based decoding**
  \[ \alpha_3 = \frac{P_{b_e}}{1 - H_2(P_{b_e})} \]

Fixed \( P_{b_e} = 0.05 \)

Remark 6
- Reference performance: \( \mathcal{O}(LM^{0.5}) \) [F. Willems, ISIT2009]
## Conclusions

- Content identification = coding problem with random codes.
- We have analyzed the performance-privacy trade-off of content identification framework.
- We have investigated the complexity of three decoding strategies based on the BDD.
- The obtained results can be of interest for security and content-based retrieval system analysis as well as low-complexity approximate search implementations.

## Extensions

- Compression (memory storage) – Identification Rate – Privacy – Complexity
- List decoding vs unique decoding
- Performance under soft decoding.