FUZZY LABELED PRIVATE SET INTERSECTION
WITH APPLICATIONS TO PRIVATE REAL-TIME BIOMETRIC SEARCH

Erkam Uzun, Simon Chung, Vladimir Kolesnikov, Alexandra Boldyreva and Wenke Lee
(Georgia Institute of Technology)
Fuzzy Labeled Private Set Intersection with Applications to Private Real-Time Biometric Search
Current practice: privacy risk

1. Detect faces
2. Send query to the cloud.
3. Similar face search
4. Display results to the client.

Cloud service obtains “query” and “result”.

No match.

Clearview AI hit with sweeping legal complaints over controversial face scraping in Europe

The privacy watchdogs believe Clearview’s image-scraping methods violate European laws

By Ian Carlos Coppel (Username) | May 27, 2021, 6:15am EDT

Fuzzy Labeled Private Set Intersection with Applications to Private Real-Time Biometric Search
Ban vs Keep Using

Fuzzy Labeled Private Set Intersection with Applications to Private Real-Time Biometric Search
Solution: Fuzzy Labeled PSI (FLPSI)

1. Detect faces
2. Send query to the cloud
3. Similar face search (under encryption)
4. Display results to the client.

FLPSI Client-side privacy layer
FLPSI Server-side privacy layer

Cloud service obtains random strings.

Pre-stored data (facial features) extracted from “persons of interest”.

Fuzzy Labeled Private Set Intersection with Applications to Private Real-Time Biometric Search
State-of-the-art

Exact private match: CHLR18
- (Labeled) Private Set Intersect.
  - E.g., contact list discovery
- Chen et al. (CCS’17, CCS’18)
  Sublinear communication.
  Efficient computation.
  Not directly be applied to fuzzy (e.g., biometrics) match.

Fuzzy private match: SANNS
- Secure Approximate NNS
  - E.g., top-k closest embedding vector search
- Chen et al. (Usenix’20)
  Accommodate fuzzy matching.
  High bandwidth requirement.
  - 1.7-5.4 GB communication to search a face over 1M-row DB.
Building FLPSI
Accommodating exact matching

Fuzzy Labeled Private Set Intersection with Applications to Private Real-Time Biometric Search
Accommodating exact matching

Fuzzy Labeled Private Set Intersection with Applications to Private Real-Time Biometric Search
Local binary encoding

SERVER

\[ d_1 \]
\[ d_2 \]
\[ d_3 \]

\[ x_1 = 01100 \]
\[ x_2 = 11011 \]
\[ x_3 = 00110 \]

CLIENT

\[ q \]

fuzzy item

\[ y = 11001 \]

bio-bit vector

Deep Learner + Locality Sensitive Hashing + Noise Removal

\[ \text{Euclidean to Hamming closeness.} \]

\[ \text{e.g., Hamming Distance}(y, x_2) = 1 \]

\[ \text{LPSI} \]

out

l2

label
Subsampling

SERVER

\[ d1 \rightarrow x1=01100 \]
\[ d2 \rightarrow x2=11011 \]
\[ d3 \rightarrow x3=00110 \]

CLIENT

\[ q \rightarrow y=11001 \]

**Binary Encoding**

\[ y=11001 \]
\[ x3=00110 \]
\[ x2=11011 \]
\[ x1=01100 \]

**Subsampling**

\[ S \text{ samples an AES key } k \text{ and masks.} \]

\[ \text{mask1}=10101 \]
\[ \text{mask2}=11001 \]
\[ \text{mask3}=01010 \]

---

\[ S \text{ locally computes:} \]

\[ x_{ij} = AES_k(x_i \land mask_j) \]

C and S computes via 2PC:

\[ y_j = AES_k(y \land mask_j) \]
Subsampling

Fuzzy Labeled Private Set Intersection with Applications to Private Real-Time Biometric Search
k-out-of-N Secret Sharing

Fuzzy Labeled Private Set Intersection with Applications to Private Real-Time Biometric Search

Client does not learn L1, but partial matches are still leaked!
Set Threshold LPSI: k-out-of-N private match

**SERVER**
- d1
- d2
- d3

**CLIENT**
- q
- fuzzy item

**Binary Encoding**
- x1=01100
- x2=11011
- x3=00110

**Subsampling**
- x11, ss11
- x12, ss12
- x13, ss13
- x21, ss21
- x22, ss22
- x23, ss23
- x31, ss31
- x32, ss32
- x33, ss33

**Secret Shares**
- ss11, ss12, ss13
- ss21, ss22, ss23
- ss31, ss32, ss33

**STLPSI**

- C can distinguish a secret share from random *iff* there are at least k-out-of-N matching subsamples.
- C must try all k-out-of-N combinations to recover the secret.
Set Threshold LPSI: k-out-of-N private match

Fuzzy Labeled Private Set Intersection with Applications to Private Real-Time Biometric Search

- Evaluate polynomial $P(c_j) = r_{ij}(c_j - x_{ij}) + ss_{ij}$
  - $r_{ij}$ is a random
  - $c_j = FHE. Enc(y_j)$

- Optimize for millions of DB records. **Sublinear** DB comm.
Set Threshold LPSI: \( k \)-out-of-\( N \) private match

Fuzzy Labeled Private Set Intersection with Applications to Private Real-Time Biometric Search
Evaluating FLPSI
Security of FLPSI: in semi-honest model

Fuzzy Labeled Private Set Intersection with Applications to Private Real-Time Biometric Search
## Datasets

<table>
<thead>
<tr>
<th>Used for</th>
<th>Query</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face-1M</td>
<td>YouTube Face-YTF (1.6K)</td>
<td>YTF (1.6K) + StyleGAN (1M)</td>
</tr>
<tr>
<td>Deep1B-1M</td>
<td>10K image descriptors</td>
<td>1M</td>
</tr>
<tr>
<td>Deep1B-10M</td>
<td>10K</td>
<td>10M</td>
</tr>
<tr>
<td>AT&amp;T</td>
<td>40 people</td>
<td>40</td>
</tr>
</tbody>
</table>
Environment and parameters

- Parameters are tuned to preserve plaintext accuracy.
  - 2-out-of-64 matching.
  - 0.67/0.75% of FRR for plaintext/FLPSI @10 false matches/query over Face-1M
- Same environment settings with SANNS (Chen et al. from Usenix’20).
  - Network settings: *fast* (500 MB/s) and *slow* (40 MB/s).
  - Azure F72s_v2 instance: 72 virtual cores, 144 GB of RAM
Performance results: Face-1M database

• Communication overhead : 40.8 MB
• Computation time:
  • @1 thread: 44 sec.
  • @72 threads: 1.36 sec.
• Best response time:
  • @fast network: 1.46 sec.
  • @slow network: 1.66 sec.
Comparison with threshold matching systems

**Distance thresholding**

- On AT&T dataset, single thread and same network speed (*fast*).
- Comparison with 7 systems.
  - 7.2x – 90x network save.
  - 121x – 7086x resp. time speed up.

**k-out-of-N matching**

- Asymptotic comparison with 3 systems.
- FLPSI is the *first* achieving communication sublinear to DB.
Comparison with kNN systems: SANNS

- The state-of-the-art: SANNS (Chen et al. from Usenix’20).
  - SANNS-linear: Searching over all DB items.
  - SANNS-approx: Searching over sub-DB items with slight accuracy penalty.

<table>
<thead>
<tr>
<th>Database</th>
<th>Protocol</th>
<th>Communication</th>
<th>Response time (fast/slow)</th>
<th>Speed-up</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Total</td>
<td>Saving</td>
<td>Total (sec.)</td>
</tr>
<tr>
<td>Deep1B-1M</td>
<td>FLSPI</td>
<td>40.8 MB</td>
<td>-</td>
<td>1.46/1.66</td>
</tr>
<tr>
<td></td>
<td>SANNS-linear</td>
<td>5.39 GB</td>
<td>132x</td>
<td>5.79/41.7</td>
</tr>
<tr>
<td></td>
<td>SANNS-approx</td>
<td>1.72 GB</td>
<td>42x</td>
<td>1.70/15.1</td>
</tr>
<tr>
<td>Deep1B-10M</td>
<td>FLSPI</td>
<td>128 MB</td>
<td>-</td>
<td>12.7/13.5</td>
</tr>
<tr>
<td></td>
<td>SANNS-linear</td>
<td>57.7 GB</td>
<td>452x</td>
<td>73.1/446</td>
</tr>
<tr>
<td></td>
<td>SANNS-approx</td>
<td>6.07 GB</td>
<td>48x</td>
<td>5.27/41.8</td>
</tr>
</tbody>
</table>
Limitations

• Requires offline preprocessing before each query
  • 501 MB storage and 37.5 sec preprocessing for 1M-row DB.

• Client requires a public DL model.

• Not resilient against malicious attacks.
  • Server can return random outputs
  • Client can exploit allowed false matches to learn entire DB.
  • But prior systems are also semi-honest.
Questions?

Email: euzun@gatech.edu

Project page: https://sites.gatech.edu/euzun/projects/biometrics-surveillance