All About That Data: Towards a Practical Assessment Of Attacks on Encrypted Search
Seny Kamara, Abdelkarim Kati, Tarik Moataz, Thomas Schneider, Amos Treiber, and Michael Yonli
Encrypted Search Algorithms (ESAs)

Trusted

Untrusted
Encrypted Search Algorithms (ESAs)
Encrypted Search Algorithms (ESAs)

q = 'crypto'

Trusted

Enc(q)

Untrusted

sk
Encrypted Search Algorithms (ESAs)

\[ q = 'crypto' \quad q = [18,39] \]

\[ Enc(q) \]

Trusted

Untrusted
Encrypted Search Algorithms (ESAs)

\( q = 'crypto' \quad q = [18,39] \)

\[ \mathcal{D}(q) = \{ D \in \mathcal{D} : q(D) \} \]

Trusted

Untrusted
Encrypted Search Algorithms (ESAs)

\[ q = 'crypto' \quad q = [18,39] \]

\[ \text{Enc}(q) \quad D(q) = \{ D \in \mathcal{D} : q(D) \} \]

Leakage
Encrypted Search Algorithms (ESAs)

$q = 'crypto'$  \quad q = [18,39]

\[ D(q) = \{ D \in \mathcal{D} : q(D) \} \]

This work

- Structured Encryption (STE)
- Searchable Symmetric Encryption (SSE)
- Oblivious RAM (ORAM)
Encrypted Search Algorithms (ESAs)

- $q = 'crypto'$
- $q = [18,39]$

$$Enc(q)$$

$D(q) = \{D \in D: q(D)\}$

This work
- Structured Encryption (STE)
- Searchable Symmetric Encryption (SSE)
- Oblivious RAM (ORAM)

Leakage attack

(Auxiliary information)

$q$ or $D$
Encrypted Search Algorithms (ESAs)

$q = 'crypto'$  \quad q = [18,39]

$Enc(q)$

$D(q) = \{D \in \mathcal{D} : q(D)\}$

Leakage

(Auxiliary information)

$q$ or $\mathcal{D}$

Leakage attack

This work

- Structured Encryption (STE)
- Searchable Symmetric Encryption (SSE)
- Oblivious RAM (ORAM)

Persistent & passive
Encrypted Search Algorithms (ESAs)

$q = 'crypto' \quad q = [18,39]$

$Enc(q)$

$D(q) = \{D \in D : q(D)\}$

Leakage

(Auxiliary information)

$q$ or $D$

Leakage attack

This work

- Structured Encryption (STE)
- Searchable Symmetric Encryption (SSE)
- Oblivious RAM (ORAM)

This work

Persistent & passive
Encrypted Search Algorithms (ESAs): Uncertainty Of Security

Constructions

Attacks & Countermeasures
Encrypted Search Algorithms (ESAs): Uncertainty Of Security

Constructions

- Benign leakage
- Common leakage
- Standard leakage
- Accepted leakage

Attacks & Countermeasures
Encrypted Search Algorithms (ESAs): Uncertainty Of Security

Constructions

- Benign leakage
- Standard leakage

Attacks & Countermeasures

- Common leakage
- Accepted leakage

[Attacks] assume extremely strong adversarial models

Leakages [...] are not exploitable via leakage-abuse attacks in practice
# Encrypted Search Algorithms (ESAs): Uncertainty Of Security

## Constructions
- Benign leakage
- Common leakage
- Standard leakage
- Accepted leakage

## Attacks & Countermeasures
- Severe threat
- Devastating results
- [ESAs] are extremely vulnerable to [attacks]
- [ESA] schemes should no longer be used without countermeasures

[Attacks] assume extremely strong adversarial models

Leakages [...] are not exploitable via leakage-abuse attacks in practice
Encrypted Search Algorithms (ESAs): Uncertainty Of Security

**Constructions**
- Benign leakage
- Common leakage
- Standard leakage
- Accepted leakage

**Attacks & Countermeasures**
- Severe threat
- Devastating results
- [ESAs] are extremely vulnerable to [attacks]
- [ESA] schemes should no longer be used without countermeasures
- Our assumptions on background information are weak
- With some prior knowledge [...] an honest-but-curious server can recover the underlying keywords

[Attacks] assume extremely strong adversarial models
Leakages [...] are not exploitable via leakage-abuse attacks in practice
Encrypted Search Algorithms (ESAs): Uncertainty Of Security

Constructions

- Benign leakage
- Common leakage
- Accepted leakage

Attacks & Countermeasures

- Severe threat
- Devastating results
- [ESA] schemes should no longer be used without countermeasures

[Attacks] assume extremely strong adversarial models
Leakages [...] are not exploitable via leakage-abuse attacks in practice
With some prior knowledge [...] an honest-but-curious server can recover the underlying keywords

Our assumptions on background information are weak

3
Previous Evaluations & Our Contributions

Previous evaluations
Previous Evaluations & Our Contributions

Previous evaluations

- Closed-source code
- Single use case
- Few comparisons
- Small/restricted data
Previous Evaluations & Our Contributions

Previous evaluations

- Closed-source code
- Single use case
- Few comparisons
- Small/restricted data

Artificial queries

Rank

Frequency
Previous Evaluations & Our Contributions

Previous evaluations:
- Closed-source code
- Single use case
- Few comparisons
- Small/restricted data
- Frequency
  - High frequency
  - Low frequency
  - Rank
- Artificial queries
Previous Evaluations & Our Contributions

**Previous evaluations**
- Closed-source code
- Few comparisons
- Small/restricted data
- High frequency
- Artificial queries

**This work**
- Open-source framework
- Multiple use cases
- Systematic re-evaluation
- Large data
Previous Evaluations & Our Contributions

### Previous evaluations

- Closed-source code
- Few comparisons
- Small/restricted data
- Artificial queries
- High frequency
- Low frequency

### This work

- Open-source framework
- Multiple use cases
- Systematic re-evaluation
- Large data
- First real-world query logs

User, Query:
- 216,'crypto'
- 216,'amsterdam'
- 106,'doctor'
- 216,'hotel'
New Software: LEAKER

- Re-implementation of 17 major attacks in open-source framework

  [IKK12, CGPR15, LMP18, GLMP18, GLMP19, GJW19, BKM20, KPT20, KPT21, RPH21]

https://encrypto.de/code/LEAKER

![Python Logo]
New Software: LEAKER

- Re-implementation of 17 major attacks in open-source framework

  [IKK12, CGPR15, LMP18, GLMP18, GLMP19, GJW19, BKM20, KPT20, KPT21, RPH21]

- Modular design & interoperability
- Easy to implement new attacks & countermeasures
- Easy to pre-process & use new data

https://encrypto.de/code/LEAKER
New Data

Keyword *(point)* queries

- @
- Cloud
- Laptop
- DNA
- My Google activity
- AOL
- tair
New Data

Keyword \((point)\) queries

Range queries

- My Google activity
- AOL
- taIR
- SDSS
- MIMIC
- DATA.GOV.UK
- Walmart
- OpenData
New Data

**Keyword (point) queries**

- Email
- Cloud
- Computer
- DNA

**Range queries**

- Microscope
- School
- Social network
- Shopping cart
- Car

Have query logs

Keywords: My Google activity, AOL, tair, SDSS, MIMIC, OpenData
### Evaluation: Summary – Keyword Search

<table>
<thead>
<tr>
<th>Leakage</th>
<th>Attack Success</th>
<th>Risk</th>
</tr>
</thead>
</table>
| • Response length  
  • Response volume | • High adversarial knowledge | Low |
| • Co-occurrence | • High adversarial knowledge | Low |
| • Response identifiers  
  • Response volumes (of individual documents) | • Low adversarial knowledge | High |

(subjective)
# Evaluation: Summary – Keyword Search

<table>
<thead>
<tr>
<th>Leakage</th>
<th>Attack Success</th>
<th>Risk</th>
</tr>
</thead>
</table>
| • Response length  
• Response volume | • High adversarial knowledge | Low |
| • Co-occurrence | • High adversarial knowledge | Low |
| • Response identifiers  
• Response volumes (of individual documents) | • Low adversarial knowledge | High |

=> Suppression of identifier and volume leakage of responses necessary!
**Evaluation: Summary – Keyword Search**

<table>
<thead>
<tr>
<th>Leakage</th>
<th>Attack Success</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Response length&lt;br&gt;• Response volume</td>
<td>• High adversarial knowledge</td>
<td>Low</td>
</tr>
<tr>
<td>• Co-occurrence</td>
<td>• High adversarial knowledge</td>
<td>Low</td>
</tr>
<tr>
<td>• Response identifiers&lt;br&gt;• Response volumes (of individual documents)</td>
<td>• Low adversarial knowledge</td>
<td>High</td>
</tr>
</tbody>
</table>

=> Suppression of identifier and volume leakage of responses necessary!

(subjective)

Subgraph attacks [BKM20]
None of the attacks worked against low-[frequency] keywords

Users are more likely to search for a specific email
None of the attacks worked against low-frequency keywords

Users are more likely to search for a specific email

Mean frequency: 1.54!
(on TAIR)
None of the attacks worked against low-frequency keywords

[BKM20]

Users are more likely to search for a specific email

[RPH21]

Mean frequency: 1.54!
(on TAIR)
None of the attacks worked against low-frequency keywords

Users are more likely to search for a specific email

Mean frequency: 1.54!
(on TAIR)

Mean frequency: 326!
(on GMail)
None of the attacks worked against low-[frequency] keywords

Users are more likely to search for a specific email

Mean frequency: 1.54!
(on TAIR)

Mean frequency: 326!
(on GMail)
### Evaluation: Summary – Range Search

<table>
<thead>
<tr>
<th>Leakage</th>
<th>Attack Success</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Response length</td>
<td>• None</td>
<td>Very low</td>
</tr>
<tr>
<td>• Response length</td>
<td>• Evenly distributed data</td>
<td>Medium</td>
</tr>
<tr>
<td>• Query equality</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Co-occurrence</td>
<td>• Large widths</td>
<td>Medium</td>
</tr>
<tr>
<td>• Order</td>
<td>• Most cases</td>
<td>High</td>
</tr>
</tbody>
</table>

(subjective)
### Evaluation: Summary – Range Search

<table>
<thead>
<tr>
<th>Leakage</th>
<th>Attack Success</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Response length</td>
<td>• None</td>
<td>Very low</td>
</tr>
<tr>
<td>• Response length</td>
<td>• Evenly distributed data</td>
<td>Medium</td>
</tr>
<tr>
<td>• Query equality</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Co-occurrence</td>
<td>• Large widths</td>
<td>Medium</td>
</tr>
<tr>
<td>• Order</td>
<td>• Most cases</td>
<td>High</td>
</tr>
</tbody>
</table>

=> Leakage suppression for range case!
Conclusions

- Extensible open-source framework LEAKER
Conclusions

• Extensible open-source framework LEAKER
• First usage of real-world queries
Conclusions

• Extensible **open-source** framework LEAKER
• **First** usage of real-world queries
• **Systematic** empirical analysis of leakage attacks
Conclusions

• Extensible **open-source** framework LEAKER
• **First** usage of real-world queries
• **Systematic** empirical analysis of leakage attacks
• **Contradict** some previous conclusions
Conclusions

- Extensible **open-source** framework LEAKER
- **First** usage of real-world queries
- **Systematic** empirical analysis of leakage attacks
- **Contradict** some previous conclusions

Leakage → Leakage attack → ???

$q$ or $D$
Conclusions

- Extensible **open-source** framework LEAKER
- **First** usage of real-world queries
- **Systematic** empirical analysis of leakage attacks
- **Contradict** some previous conclusions

This work

Leakage → Leakage attack → $q$ or $D$
What needs to be done
What needs to be done
What needs to be done
What needs to be done
THANK YOU!

https://encrypto.de/treiber

More details: https://ia.cr/2021/1035
(to appear at EuroS&P’22)

Code: https://encrypto.de/code/LEAKER
Resources


Resources

# Leakage Patterns

<table>
<thead>
<tr>
<th>Leakage Pattern</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response Length</td>
<td>$</td>
</tr>
<tr>
<td>Query Equality</td>
<td>$q_i = q_j$</td>
</tr>
<tr>
<td>Co-Occurrence</td>
<td>$</td>
</tr>
<tr>
<td>Response Identifiers</td>
<td>${i : D_i \in q(D)}$</td>
</tr>
<tr>
<td>Response Volumes</td>
<td>${</td>
</tr>
</tbody>
</table>

(Simplified)
**Leakage Attacks Types**

**Keyword (point) queries**

- `real` 2, 5, 11, 13, 20, 31
- `world` 3, 5, 10, 11, 13, 25
- `crypto` 5, 11, 21, 27

\[ D(q) = \{ D \in \mathcal{D} : q \in D \} \]

**Unknown data:** Adversary knows subset of \( \mathcal{D} \)

**Range queries**

- `real` 2, 5, 11, 13, 20, 31
- `world` 3, 5, 10, 11, 13, 25
- `crypto` 5, 11, 21, 27

\[ D(q) = \{ r \in \mathcal{D} : a \leq r \leq b \} \]

**No auxiliary knowledge**
Overview of Leakage Attacks on ESAs

Adversary Type
- Persistent
- Active
- Passive

Adversary Power
- Snapshot
- Persistent

Injection Attacks
- [ZKP16, BKM20, PWLP20]

Auxiliary Information
- Sampled-data or sampled-query
- Known-data
  - Keyword attacks
    - [LZWT14, LMP18, GLMP18, GJW19, OKa21, DHP21, GPP21, OKb21]
  - Range attacks
    - [KKNO16, LMP18, GLMP18, GLMP19, GJW19, KPT20, KPT21]

This work

Keywords & Range attacks
- [LZWT14, LMP18, GLMP18, GJW19, OKa21, DHP21, GPP21, OKb21]

\[ q = w \]
\[ D(q) = \{ D \in D : q(D) \} \]
Recover \( q \)

Range attacks
- \( q = (a, b) \)
- \( D(q) = \{ r \in D : a \leq r \leq b \} \)
Recover \( D \)
### Overview of Techniques for ESAs (Extremely informal)

<table>
<thead>
<tr>
<th>Technique</th>
<th>Leakage</th>
<th>Query Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully Homomorphic Encryption (FHE)</td>
<td>• None</td>
<td>Linear</td>
</tr>
<tr>
<td>Oblivious RAM (ORAM)</td>
<td>• Response Length</td>
<td>Sublinear</td>
</tr>
</tbody>
</table>
| Structured Encryption (STE)      | • Query Equality  
• Responses’ Equality | Optimal   |
| Property-Preserving Encryption (PPE) | • Ciphertext Equality  
• Ciphertext Order | Optimal   |

Considered secure but inefficient

This work

Considered efficient and ???

Considered efficient but insecure [NKW15]
Previous Evaluations

- Usual evaluations for keyword attacks:
  1. Enron (\& Apache) email data collection
  2. Restrict data to 500-3000 keywords
  3. Draw 150 queries from data collection
  4. Evaluate on partial knowledge

Frequency

- High frequency
- Low frequency

Rank

- ???
Previous Evaluations

- Usual evaluations for range attacks:
  1. Subset of HCUP or artificial Data collection
  2. Pick Artificial query distribution (Uniform/Zipf/…)
  3. Evaluate for different amounts of queries

or

???
New Data

• 9 new data sources for more realistic evaluations
• Keyword setting:

Use Case: Email/Cloud
- GMail and Google Drive
  - 7 Query Logs & Data Collections
  - 7 Users
  - 16-100 Queries
  - 200-47k Documents
  - 19k-895k Keywords

Web
- AOL and Wikipedia
  - 1 Query Log & 1 Data Collection
  - 656k Users
  - 2.9M Queries
  - 151k Documents
  - 268k Keywords

Genetic
- The Arabidopsis Information Resource
  - 1 Query Log & 1 Data Collection
  - 1.3k Users
  - 54k Queries
  - 115k Documents
  - 690k Keywords
- **Range setting:**

### Scientific
- **Sloan Digital Sky Survey**
  - 3 Query Logs & 1 Data Collection
  - 3 Users
  - 215-8k Queries
  - 5M Records
  - Domain $N = 10^k$
  - Density 96%

### Medical
- **Medical Information Mart for Intensive Care**
  - 3 Data Collections
  - 2k-8k Records
  - Domain $N = 73 - 10^k$
  - Density 3.3%-81%

### Human Resources
- **Salaries of the UK Attorney General’s Office junior civil servants**
  - 1 Data Collection
  - 536 Records
  - Domain $N = 395$
  - Density 2.3%

### Sales
- **Walmart Sales Data**
  - 1 Data Collection
  - 143 Records
  - Domain $N = 6.3k$
  - Density 2.3%

### Insurance
- **NYDT Insurance Claims**
  - 1 Data Collection
  - 886 Records
  - Domain $N = 25k$
  - Density 1.2%
Table 5: Normalized mean errors on the entire SDSS query logs. For feasibility, the collection is sampled $25 \times$ uniformly at random with size $n = 10^4$ ($n = 10^3$ for APA and ARR).

<table>
<thead>
<tr>
<th>Instance</th>
<th>GKKNO</th>
<th>AVALUE</th>
<th>ARR</th>
<th>ARR-OR</th>
<th>APA-OR$^{BT}$</th>
<th>APA-OR$^{ABT}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDSS-S</td>
<td>0.413</td>
<td>0.432</td>
<td>0.473</td>
<td>0.249</td>
<td>0.242</td>
<td>0.239</td>
</tr>
<tr>
<td>SDSS-M</td>
<td>0.408</td>
<td>0.435</td>
<td>0.287</td>
<td>0.128</td>
<td>0.242</td>
<td>0.240</td>
</tr>
<tr>
<td>SDSS-L</td>
<td>0.417</td>
<td>0.456</td>
<td>0.286</td>
<td>0.141</td>
<td>0.241</td>
<td>0.242</td>
</tr>
</tbody>
</table>