The State of the Uniform: Attacks on Encrypted Databases Beyond the Uniform Query Distribution

EVGENIOS M. KORNAROPOULOS
UC BERKELEY

Joint work with:
Charalampos (Babis) Papamanthou
Roberto Tamassia
INTRO

ENCRYPTED SEARCH

\[ \text{Enc}_K(q) \]

\[ \text{Enc}_K(\text{database}) \]

\[ \text{Enc}_K(\text{database}) \]
Leakage: Information Revealed by Construction
WHAT IS LEAKAGE?
INTRO
WHAT IS LEAKAGE?

Client

Server

Tokens

Responses

5
WHAT IS LEAKAGE?

Client

Server

Tokens

Responses

\[ \text{PRF}_K(\bullet) = t \]
WHAT IS LEAKAGE?

Tokens

PRF_K(●) = t
PRF_K(●) = t'
PRF_K(●) = t''
PRF_K(●) = t

Responses
WHAT IS LEAKAGE?

Client

Server

Tokens

PRF\_K(\bullet) = t

PRF\_K(\bullet) = t'

PRF\_K(\bullet) = t''

PRF\_K(\bullet) = t
INTRO
WHAT IS LEAKAGE?

Client

Server

 Tokens

PRF$_K$(•) = t

PRF$_K$(•) = t'

PRF$_K$(•) = t''

PRF$_K$(•) = t

Responses

Search Pattern
Leakage
**INTRO**

**WHAT IS LEAKAGE?**

The diagram illustrates the concept of leakage, specifically focusing on access patterns and search patterns. The process begins with the client generating a token, which is then processed through a PRF function with key $K$.

**Tokens**

- $PRF_K(\bullet) = t$
- $PRF_K(\bullet) = t'$
- $PRF_K(\bullet) = t''$
- $PRF_K(\bullet) = t$

**Responses**

The responses are stored in a database, with each row representing a response that is potentially leaked.

**Search Pattern Leakage**

- The search pattern leakage involves the client's search requests, which are used to infer information about the database.

**Access Pattern Leakage**

- The access pattern leakage involves the pattern of accesses made by the client, which can also be used to infer information about the database.

**Server**

The server processes the incoming requests and generates responses, which are then stored in the database.
INTRO

LEAKAGE-ABUSE ATTACKS: STATE OF THE ART
## INTRO

**LEAKAGE-ABUSE ATTACKS: STATE OF THE ART**

<table>
<thead>
<tr>
<th>Value Reconstruction Attack Algorithms</th>
<th>Query Type</th>
<th>Assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Query Type</td>
<td>Data Values in a Fixed Region</td>
</tr>
<tr>
<td>KPT</td>
<td>$k$-NN</td>
<td>Uniform</td>
</tr>
</tbody>
</table>

Data Recovery on Encrypted Databases With $k$-Nearest Neighbor Query Leakage

S&P'19
### LEAKAGE-ABUSE ATTACKS: STATE OF THE ART

<table>
<thead>
<tr>
<th>Value Reconstruction Attack Algorithms</th>
<th>Query Type</th>
<th>Query Distribution</th>
<th>Data Values in a Fixed Region</th>
<th>Known Data Distribution</th>
<th>Known Query Distribution</th>
<th>Dense Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>KPT</td>
<td>$k$-NN</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KKNO</td>
<td>Range</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
### LEAKAGE-ABUSE ATTACKS: STATE OF THE ART

<table>
<thead>
<tr>
<th>Value Reconstruction Attack Algorithms</th>
<th>Query Type</th>
<th>Query Distribution</th>
<th>Data Values in a Fixed Region</th>
<th>Known Data Distribution</th>
<th>Known Query Distribution</th>
<th>Dense Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>KPT</td>
<td>$k$-NN</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KKNO</td>
<td>Range</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LMP</td>
<td>Range</td>
<td>Agnostic</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>●</td>
</tr>
</tbody>
</table>
LEAKAGE-ABUSE ATTACKS: STATE OF THE ART
# Leakage-Abuse Attacks: State of the Art

<table>
<thead>
<tr>
<th>Value Reconstruction Attack Algorithms</th>
<th>Query Type</th>
<th>Query Distribution</th>
<th>Data Values in a Fixed Region</th>
<th>Known Data Distribution</th>
<th>Known Query Distribution</th>
<th>Dense Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>KPT</td>
<td>$k$-NN</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KKNO</td>
<td>Range</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LMP</td>
<td>Range</td>
<td>Agnostic</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>●</td>
</tr>
<tr>
<td>GLMP GENERALIZEDKKNO</td>
<td>Range</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GLMP APPROXVALUE</td>
<td>Range</td>
<td>Uniform</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>-</td>
</tr>
<tr>
<td>GLMP AOR to ADR</td>
<td>Range</td>
<td>Agnostic</td>
<td>-</td>
<td>●</td>
<td>●</td>
<td>-</td>
</tr>
</tbody>
</table>
INTRO
LEAKAGE-ABUSE ATTACKS: STATE OF THE ART

<table>
<thead>
<tr>
<th>Value Reconstruction Attack Algorithms</th>
<th>Query Type</th>
<th>Query Distribution</th>
<th>Data Values in a Fixed Region</th>
<th>Known Data Distribution</th>
<th>Known Query Distribution</th>
<th>Dense Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>KPT</td>
<td>$k$-NN</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KKNO</td>
<td>Range</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LMP</td>
<td>Range</td>
<td>Agnostic</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GLMP GeneralizedKKNO</td>
<td>Range</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GLMP ApproxValue</td>
<td>Range</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GLMP AOR to ADR</td>
<td>Range</td>
<td>Agnostic</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

“While there has been some progress on designing leakage attacks against STE [9, 24, 30, 32], these attacks remain mostly of theoretical interest due to the strong assumptions they rely on.”
# Leakage-Abuse Attacks: State of the Art

<table>
<thead>
<tr>
<th>Value Reconstruction Attack Algorithms</th>
<th>Query Type</th>
<th>Query Distribution</th>
<th>Data Values in a Fixed Region</th>
<th>Known Data Distribution</th>
<th>Known Query Distribution</th>
<th>Dense Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>KPT</td>
<td>$k$-NN</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KKNO</td>
<td>Range</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LMP</td>
<td>Range</td>
<td>Agnostic</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>●</td>
</tr>
<tr>
<td>GLMP GeneralizedKKNO</td>
<td>Range</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GLMP ApproxValue</td>
<td>Range</td>
<td>Uniform</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>-</td>
</tr>
<tr>
<td>GLMP AOR to ADR</td>
<td>Range</td>
<td>Agnostic</td>
<td>-</td>
<td>●</td>
<td>●</td>
<td>-</td>
</tr>
</tbody>
</table>
# INTRO

## LEAKAGE-ABUSE ATTACKS: STATE OF THE ART

<table>
<thead>
<tr>
<th>Value Reconstruction Attack Algorithms</th>
<th>Query Type</th>
<th>Query Distribution</th>
<th>Data Values in a Fixed Region</th>
<th>Known Data Distribution</th>
<th>Known Query Distribution</th>
<th>Dense Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>KPT</td>
<td>$k$-NN</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KKNO</td>
<td>Range</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LMP</td>
<td>Range</td>
<td>Agnostic</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>●</td>
</tr>
<tr>
<td>GLMP GeneralizedKKNO</td>
<td>Range</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GLMP ApproxValue</td>
<td>Range</td>
<td>Uniform</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>-</td>
</tr>
<tr>
<td>GLMP AOR to ADR</td>
<td>Range</td>
<td>Agnostic</td>
<td>-</td>
<td>●</td>
<td>●</td>
<td>-</td>
</tr>
</tbody>
</table>
# Leakage-Abuse Attacks: State of the Art

<table>
<thead>
<tr>
<th>Value Reconstruction Attack Algorithms</th>
<th>Query Type</th>
<th>Assumptions</th>
<th>Query Distribution</th>
<th>Known Data Distribution</th>
<th>Known Query Distribution</th>
<th>Dense Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>KPT</td>
<td>$k$-NN</td>
<td>1</td>
<td>Uniform</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>KKNO</td>
<td>Range</td>
<td>1</td>
<td>Uniform</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>LMP</td>
<td>Range</td>
<td>1</td>
<td>Agnostic</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GLMP GeneralizedKKNO</td>
<td>Range</td>
<td>1</td>
<td>Uniform</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GLMP ApproxValue</td>
<td>Range</td>
<td>1</td>
<td>Uniform</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GLMP AOR to ADR</td>
<td>Range</td>
<td>1</td>
<td>Agnostic</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
# Leakage-Abuse Attacks: State of the Art

<table>
<thead>
<tr>
<th>Value Reconstruction Attack Algorithms</th>
<th>Query Type</th>
<th>Query Distribution</th>
<th>Data Values in a Fixed Region</th>
<th>Known Data Distribution</th>
<th>Known Query Distribution</th>
<th>Dense Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>KPT</td>
<td>$k$-NN</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KKNO</td>
<td>Range</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LMP</td>
<td>Range</td>
<td>Agnostic</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>●</td>
</tr>
<tr>
<td>GLMP GENERALIZEDKKNO</td>
<td>Range</td>
<td>Uniform</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GLMP APPROXValue</td>
<td>Range</td>
<td>Uniform</td>
<td>●</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GLMP AOR to ADR</td>
<td>Range</td>
<td>Agnostic</td>
<td>-</td>
<td>●</td>
<td>●</td>
<td>-</td>
</tr>
</tbody>
</table>

This Work $k$-NN & Range Agnostic - - - - -
STATE OF THE UNIFORM
OVERVIEW

NEW INSIGHTS ON LEAKAGE EXPLOITATION

Synergy between Search Pattern Leakage + Access Pattern Leakage
Non-parametric estimation techniques on the Search Pattern Leakage information
STATE OF THE UNIFORM
OVERVIEW

NEW INSIGHTS ON LEAKAGE EXPLOITATION

Synergy between Search Pattern Leakage + Access Pattern Leakage
Non-parametric estimation techniques on the Search Pattern Leakage information

REVISIT RANGE APPROXIMATE RECONSTRUCTION

Combination of new tools and Optimization formulation and no assumptions about the query or data distribution
NEW INSIGHTS ON LEAKAGE EXPLOITATION

Synergy between Search Pattern Leakage + Access Pattern Leakage
Non-parametric estimation techniques on the Search Pattern Leakage information

REVISIT RANGE APPROXIMATE RECONSTRUCTION

Combination of new tools and Optimization formulation and no assumptions about the query or data distribution

REVISIT k-NN APPROXIMATE RECONSTRUCTION

Smaller number of samples, larger k values and no assumptions about the query or data distribution
STATE OF THE UNIFORM
OVERVIEW

NEW INSIGHTS ON LEAKAGE EXPLOITATION

Synergy between Search Pattern Leakage + Access Pattern Leakage
Non-parametric estimation techniques on the Search Pattern Leakage information

REVISIT RANGE APPROXIMATE RECONSTRUCTION

Combination of new tools and Optimization formulation and no assumptions about the query or data distribution

REVISIT k-NN APPROXIMATE RECONSTRUCTION

Smaller number of samples, larger k values and no assumptions about the query or data distribution
STATE OF THE UNIFORM

ASSUMPTIONS OF THE ATTACKS

BOUNDARIES:
   Known boundaries $\alpha$ and $\beta$

STATIC:
   No updates in the database

EXACT RESPONSES:
   No false positives records or missing records

QUERY DISTRIBUTION:
   Fixed distribution with non-zero probabilities. Queries are i.i.d.
WHAT CAN THE ADVERSARY LEARN FROM THE SEARCH PATTERN LEAKAGE?
• Consider a token as an “encrypted pair of boundaries”

<table>
<thead>
<tr>
<th>Tokens</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{PRF}_K((\text{ }, \text{ }) )$</td>
<td><img src="image" alt="Responses" /></td>
</tr>
<tr>
<td>$\text{PRF}_K((\text{ }, \text{ }) )$</td>
<td><img src="image" alt="Responses" /></td>
</tr>
<tr>
<td>$\text{PRF}_K((\text{ }, \text{ }) )$</td>
<td><img src="image" alt="Responses" /></td>
</tr>
<tr>
<td>$\text{PRF}_K((\text{ }, \text{ }) )$</td>
<td><img src="image" alt="Responses" /></td>
</tr>
<tr>
<td>$\text{PRF}_K((\text{ }, \text{ }) )$</td>
<td><img src="image" alt="Responses" /></td>
</tr>
<tr>
<td>$\text{PRF}_K((\text{ }, \text{ }) )$</td>
<td><img src="image" alt="Responses" /></td>
</tr>
<tr>
<td>$\text{PRF}_K((\text{ }, \text{ }) )$</td>
<td><img src="image" alt="Responses" /></td>
</tr>
<tr>
<td>$\text{PRF}_K((\text{ }, \text{ }) )$</td>
<td><img src="image" alt="Responses" /></td>
</tr>
<tr>
<td>$\text{PRF}_K((\text{ }, \text{ }) )$</td>
<td><img src="image" alt="Responses" /></td>
</tr>
<tr>
<td>$\text{PRF}_K((\text{ }, \text{ }) )$</td>
<td><img src="image" alt="Responses" /></td>
</tr>
</tbody>
</table>
Consider a token as an “encrypted pair of boundaries”

**Tokens**

- \( \text{PRF}_K((\text{●}, \text{○})) \)
- \( \text{PRF}_K((\text{●}, \text{●})) \)
- \( \text{PRF}_K((\text{○}, \text{○})) \)
- \( \text{PRF}_K((\text{○}, \text{●})) \)
- \( \text{PRF}_K((\text{●}, \text{○})) \)
- \( \text{PRF}_K((\text{○}, \text{●})) \)
- \( \text{PRF}_K((\text{●}, \text{●})) \)
- \( \text{PRF}_K((\text{○}, \text{○})) \)
- \( \text{PRF}_K((\text{○}, \text{●})) \)
- \( \text{PRF}_K((\text{●}, \text{○})) \)
- \( \text{PRF}_K((\text{○}, \text{●})) \)
- \( \text{PRF}_K((\text{●}, \text{●})) \)

**Responses**
• Consider a token as an “encrypted pair of boundaries”
- Consider a token as an “encrypted pair of boundaries”
- Partition the token-response with respect to the response
• Consider a token as an “encrypted pair of boundaries”
• Partition the token-response with respect to the response
STATE OF THE UNIFORM
SEARCH PATTERN + ACCESS PATTERN

- Consider a token as an “encrypted pair of boundaries”
- Partition the token-response with respect to the response
STATE OF THE UNIFORM
SEARCH PATTERN + ACCESS PATTERN

- Consider a token as an “encrypted pair of boundaries”
- Partition the token-response with respect to the response

\[ \text{PRF}_K \left( (\cdot, \cdot), (\cdot, \cdot) \right) \]

\[ \text{PRF}_K \left( (\cdot, \cdot), (\cdot, \cdot) \right) \]

\[ \text{PRF}_K \left( (\cdot, \cdot), (\cdot, \cdot) \right) \]
STATE OF THE UNIFORM
SEARCH PATTERN + ACCESS PATTERN

- Consider a token as an “encrypted pair of boundaries”
- Partition the token-response with respect to the response

\[
\text{PRF}_K((\text{gray}, \text{blue}))
\]

\[
\text{PRF}_K((\text{gray}, \text{red}))
\]

\[
\text{PRF}_K((\text{green}, \text{blue}))
\]

\[
\text{PRF}_K((\text{green}, \text{red}))
\]

\[
\text{PRF}_K((\text{white}, \text{blue}))
\]

How many distinct tokens exist that return response?
STATE OF THE UNIFORM

SEARCH PATTERN + ACCESS PATTERN

- Consider a token as an “encrypted pair of boundaries”
- Partition the token-response with respect to the response

How many distinct tokens exist that return response?
STATE OF THE UNIFORM
SEARCH PATTERN + ACCESS PATTERN

- Consider a token as an “encrypted pair of boundaries”
- Partition the token-response with respect to the response

How many distinct tokens exist that return response?
Plaintext “Universe”

Number of distinct range queries with response $v_1$, $v_2$, $v_3$?
STATE OF THE UNIFORM
SEARCH PATTERN + ACCESS PATTERN

Plaintext “Universe”

Number of distinct range queries with response $v_1, v_2, v_3$?

$$|Q_r| = d(v_0, v_1) \cdot d(v_3, v_4)$$
Plaintext “Universe”

Number of distinct range queries with response \( v_1, v_2, v_3 \) ?

\[ |Q_r| = d(v_0, v_1) \cdot d(v_3, v_4) \]

If we infer the number of distinct range queries we learn the product of distances!
STATE OF THE UNIFORM
SEARCH PATTERN + ACCESS PATTERN

- Consider a token as an “encrypted pair of boundaries”
- Partition the token-response with respect to the response

How many distinct tokens exist that return response?
STATE OF THE UNIFORM

SEARCH PATTERN + ACCESS PATTERN

- Consider a token as an “encrypted pair of boundaries”
- Partition the token-response with respect to the response

PRF_K((?, ?))
PRF_K((?, ?))
PRF_K((?, ?))
PRF_K((?, ?))
PRF_K((?, ?))
PRF_K((?, ?))
STATE OF THE UNIFORM

SEARCH PATTERN + ACCESS PATTERN

- Consider a token as an “encrypted pair of boundaries”
- Partition the token-response with respect to the response

Can we estimate how many distinct tokens exist that return response?
Can we estimate how many distinct tokens exist that return response?

T: Random variable takes values from universe of tokens
R: Random variable takes values from universe of Responses

STATE OF THE UNIFORM
SEARCH PATTERN + ACCESS PATTERN

- Consider a token as an “encrypted pair of boundaries”
- Partition the token-response with respect to the response
STATE OF THE UNIFORM
SEARCH PATTERN + ACCESS PATTERN

- Consider a token as an “encrypted pair of boundaries”
- Partition the token-response with respect to the \textit{response}

Can we estimate how many distinct tokens exist that return response? 

\textbf{T:} Random variable takes values from universe of tokens 
\textbf{R:} Random variable takes values from universe of Responses

Sample from $\text{Pr}(T \mid R = \{ , , \})$
Can we estimate how many distinct tokens exist that return response?

T: Random variable takes values from universe of tokens
R: Random variable takes values from universe of Responses

Sample from \( \Pr(T | R = \{\text{ }, \text{ }, \text{ }\}) \)
STATE OF THE UNIFORM
SEARCH PATTERN + ACCESS PATTERN

- Consider a token as an “encrypted pair of boundaries”
- Partition the token-response with respect to the response

Can we estimate how many distinct tokens exist that return response?

T: Random variable takes values from universe of tokens
R: Random variable takes values from universe of Responses

Sample from \( \text{Pr}(T \mid R = \{,\}) \)
STATE OF THE UNIFORM
SEARCH PATTERN + ACCESS PATTERN

- Consider a token as an “encrypted pair of boundaries”
- Partition the token-response with respect to the response

\[
\text{PRF}_K((\text{gray}, \text{blue})), \text{PRF}_K((\text{blue}, \text{red})), \text{PRF}_K((\text{green}, \text{red})), \text{PRF}_K((\text{green}, \text{blue})), \\
\text{Sample from } \Pr(T|R=\{\text{blue}, \text{red}, \text{green}\})
\]
STATE OF THE UNIFORM
SEARCH PATTERN + ACCESS PATTERN

- Consider a token as an “encrypted pair of boundaries”
- Partition the token-response with respect to the response

Sample from $\Pr(T | R = \{\ ,\ ,\ ,\ })$

Set of frequencies of responses $\{2, 2, 1\}$

Support Size Estimator for Distributions
STATE OF THE UNIFORM
SEARCH PATTERN + ACCESS PATTERN

- Consider a token as an “encrypted pair of boundaries”
- Partition the token-response with respect to the response

Sample from Pr(T | R = \{(\(
\), \(
\)), \ldots\})

Set of frequencies of responses \{2, 2, 1\}

Support Size Estimator for Distributions

\(d(v_0, v_1) d(v_3, v_4)\)
LET’S TALK ABOUT SUPPORT SIZE

ESTIMATORS

JACKKNIFE

- Non-parametric
- Frequency of each token as input
- Based on bias reduction, order decided based on the sample

VALIANT-VALIANT

- Non-parametric
- Frequency of each token as input
- “Simplest Histogram”
WHAT CAN THE ADVERSARY LEARN FROM THE SEARCH PATTERN LEAKAGE?
WHAT CAN THE ADVERSARY LEARN FROM THE SEARCH PATTERN LEAKAGE?

Answer: From **Frequencies** of tokens we can estimate the product of pairwise distances
OVERVIEW OF THE ATTACK
OVERVIEW OF THE ATTACK

Plaintext:

\[ d(\alpha, \nu_0) = L_0 \quad d(\nu_0, \nu_3) = L_1 \quad d(\nu_1, \beta) = L_2 \]
OVERVIEW OF THE ATTACK

Plaintext:

\[ d(\alpha, \nu_0) = L_0 \]
\[ d(\nu_0, \nu_1) = L_1 \]
\[ d(\nu_1, \beta) = L_2 \]
OVERVIEW OF THE ATTACK

Plaintext:

\[ d(\alpha, v_0) = L_0 \]
\[ d(v_0, v_3) = L_1 \]
\[ d(v_1, \beta) = L_2 \]

Support Size Estimation on Tokens:

\( \hat{L}_0L_1 = 350 \)
\( \hat{L}_1L_2 = 1015 \)
\( \hat{L}_0L_2 = 290 \)
OVERVIEW OF THE ATTACK

Plaintext:

\[ d(\alpha, v_0) = L_0 \]
\[ d(v_0, v_1) = L_1 \]
\[ d(v_1, \beta) = L_2 \]

Support Size Estimation on Tokens

\[ \hat{L}_0 L_1 = 350 \]
\[ \hat{L}_1 L_2 = 1015 \]
\[ \hat{L}_0 L_2 = 290 \]

Choose Lengths that agree with the Estimations
OVERVIEW OF THE ATTACK

Plaintext:

\[
\begin{align*}
\alpha & \quad V_0 \\
\beta & \quad V_1
\end{align*}
\]

Support Size Estimation on Tokens

Choose Lengths that agree with the Estimations

\[
\min_{L_0, L_1, L_2} \left( (L_0 \cdot L_1 - 350)^2 + (L_1 \cdot L_2 - 1015)^2 - (L_0 \cdot L_2 - 290)^2 \right)
\]

s.t. \( \sum L_i = N \)

\( L_i \geq 0 \)
OVERVIEW OF THE ATTACK

Plaintext:

Support Size Estimation on Tokens

\[ d(\alpha, v_0) = L_0 \]
\[ d(v_0, v_1) = L_1 \]
\[ d(v_1, \beta) = L_2 \]

Choose Lengths that agree with the Estimations

\[
\begin{align*}
\min_{L_0, L_1, L_2} & \quad (L_0 \cdot L_1 - 350)^2 + (L_1 \cdot L_2 - 1015)^2 - (L_0 \cdot L_2 - 290)^2 \\
\text{s.t.} & \quad \sum L_i = N \\
& \quad L_i \geq 0
\end{align*}
\]

\[
\begin{align*}
\min_{X_0, X_1, X_2} & \quad (X_0 + X_1 - \log 350)^2 + (X_1 + X_2 - \log 1015)^2 - (X_0 + X_2 - \log 290)^2 \\
\text{s.t.} & \quad \sum X_i = \log N
\end{align*}
\]
OVERVIEW OF THE ATTACK

Plaintext:

Support Size Estimation on Tokens

Choose Lengths that agree with the Estimations

\[
\min_{L_0, L_1, L_2} \left( (L_0 \cdot L_1 - 350)^2 + (L_1 \cdot L_2 - 1015)^2 - (L_0 \cdot L_2 - 290)^2 \right)
\]

s.t. \( \sum L_i = N \quad L_i \geq 0 \)

\[
\min_{X_0, X_1, X_2} \left( (X_0 + X_1 - \log 350)^2 + (X_1 + X_2 - \log 1015)^2 - (X_0 + X_2 - \log 290)^2 \right)
\]

s.t. \( \sum X_i = \log N \)
Approximate Reconstruction

Range Queries

Uniform Query Distribution, $N=10^3$, $Q=10^4$

- GeneralizedKKNO
- Agnostic-Reconstruction-Range

Mean-Square Error vs Database Density
RANGE QUERIES
APPROXIMATE RECONSTRUCTION
RANGE QUERIES
APPROXIMATE RECONSTRUCTION
First attacks that combine Search Pattern and Access Pattern Leakage to **overcome strong assumptions** such as uniform query distribution.