

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



Cyberattack: Reports of patient records published online 'credible and accurate'

© Wed, May 19, 2021, 10:25

Tim O'Brien

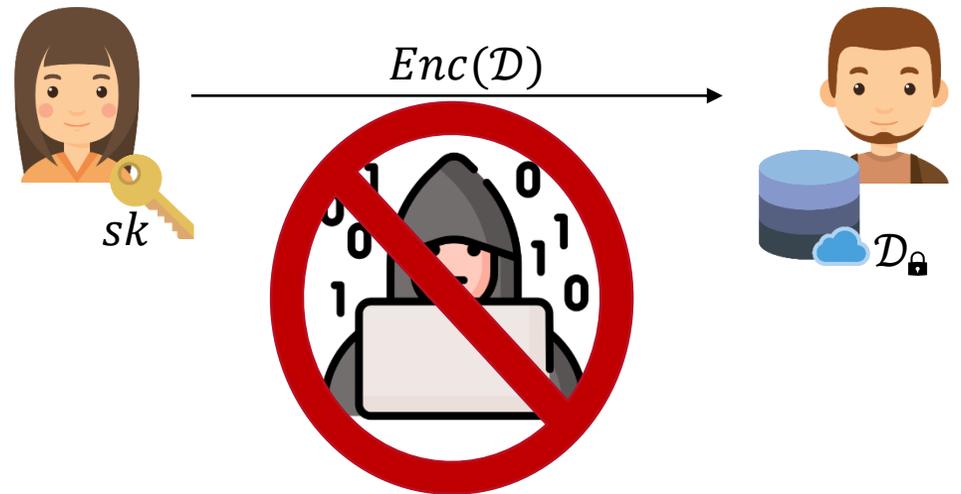


The release of stolen data, including medical records, is viewed as 'credible and accurate' by some users, according to a survey by the cybersecurity firm.

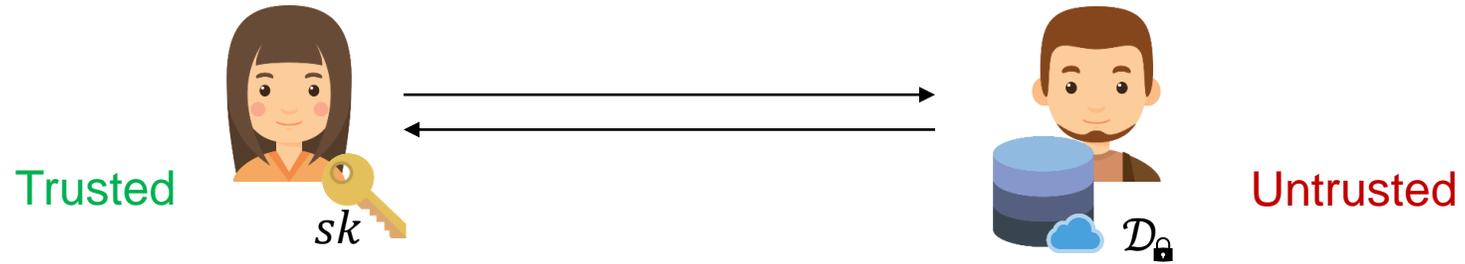


Emotet Returns in Malspam Attacks Dropping TrickBot, OakBot

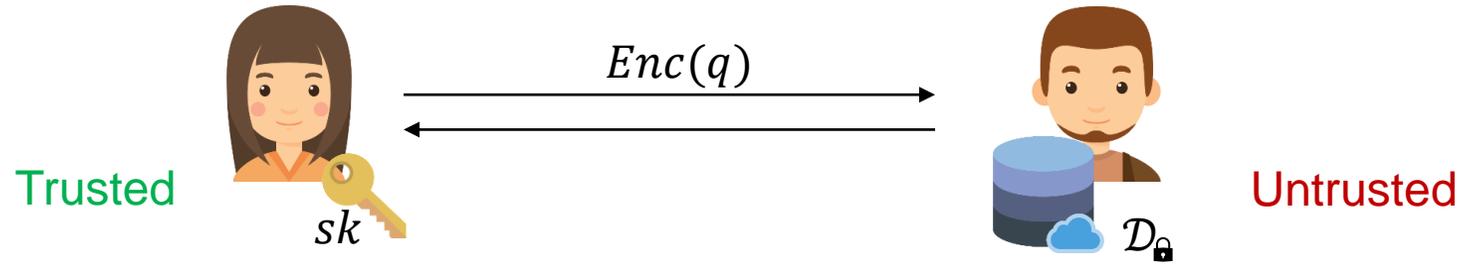
Leak Exposes Private Data of Genealogy Service Users



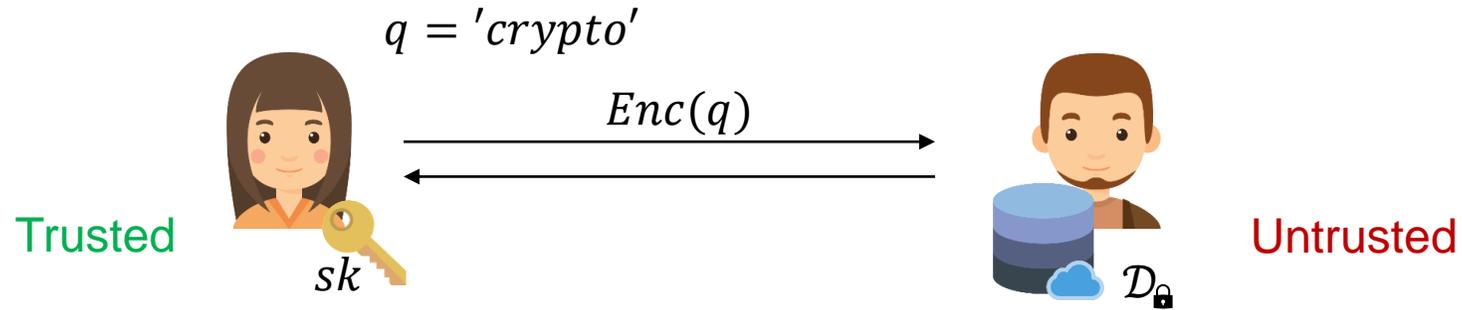
Encrypted Search Algorithms (ESAs)



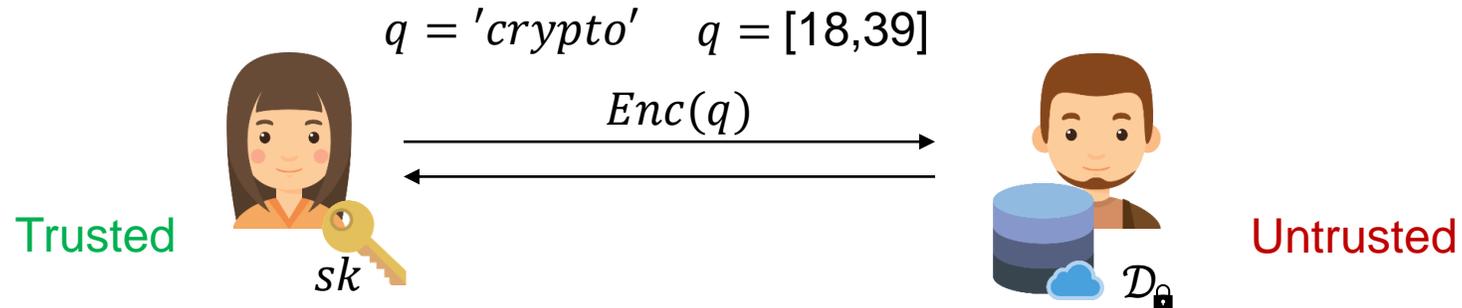
Encrypted Search Algorithms (ESAs)



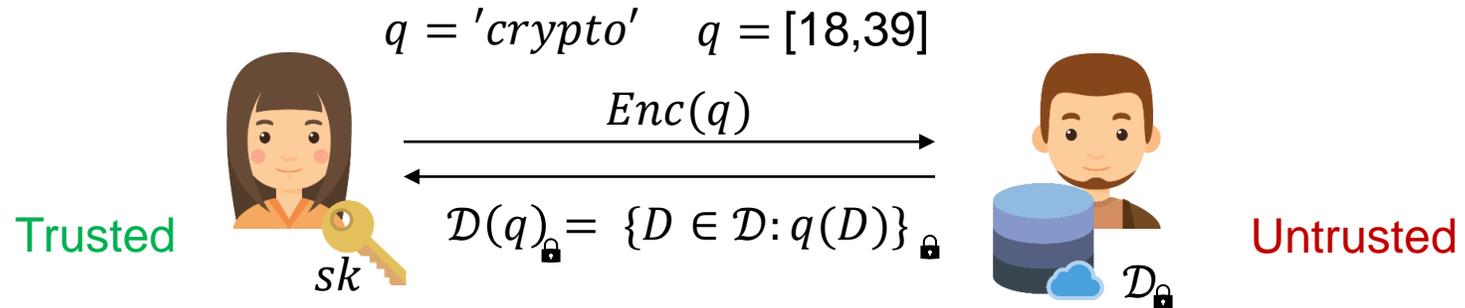
Encrypted Search Algorithms (ESAs)



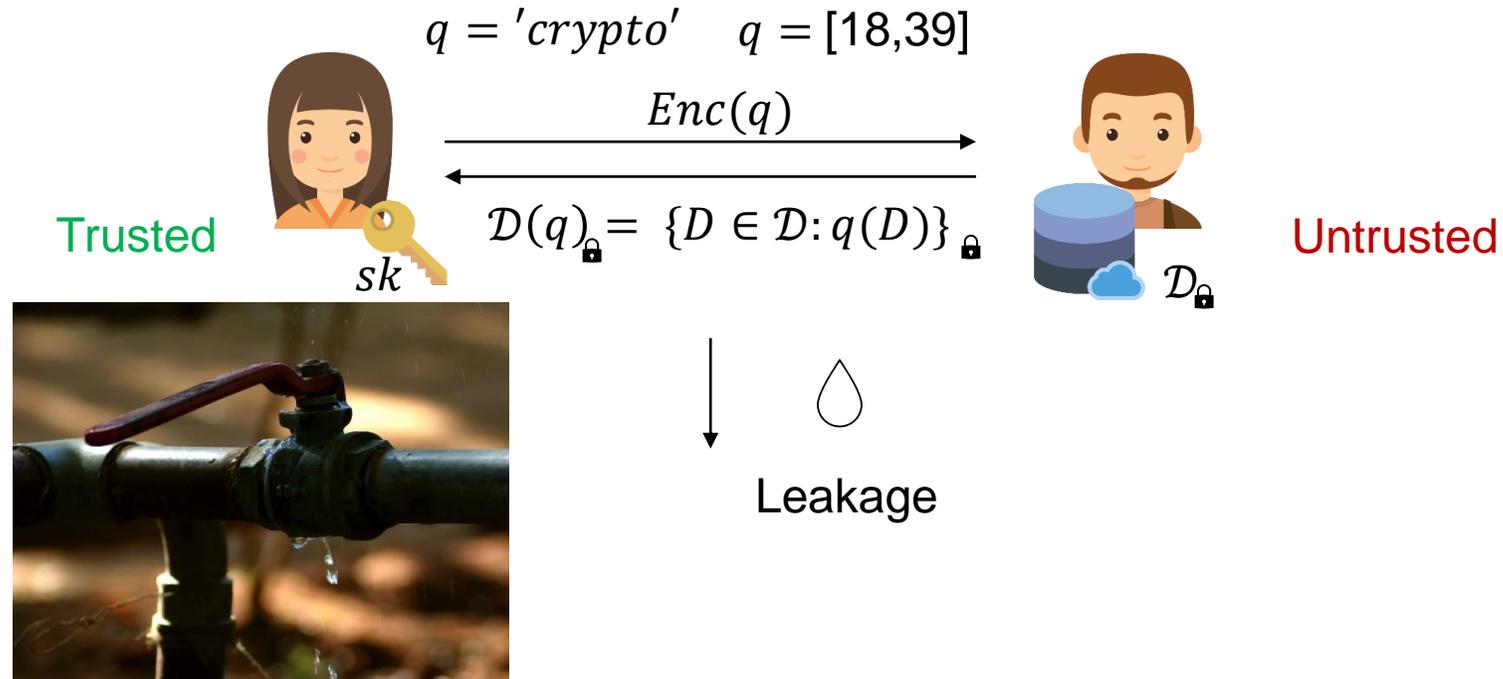
Encrypted Search Algorithms (ESAs)



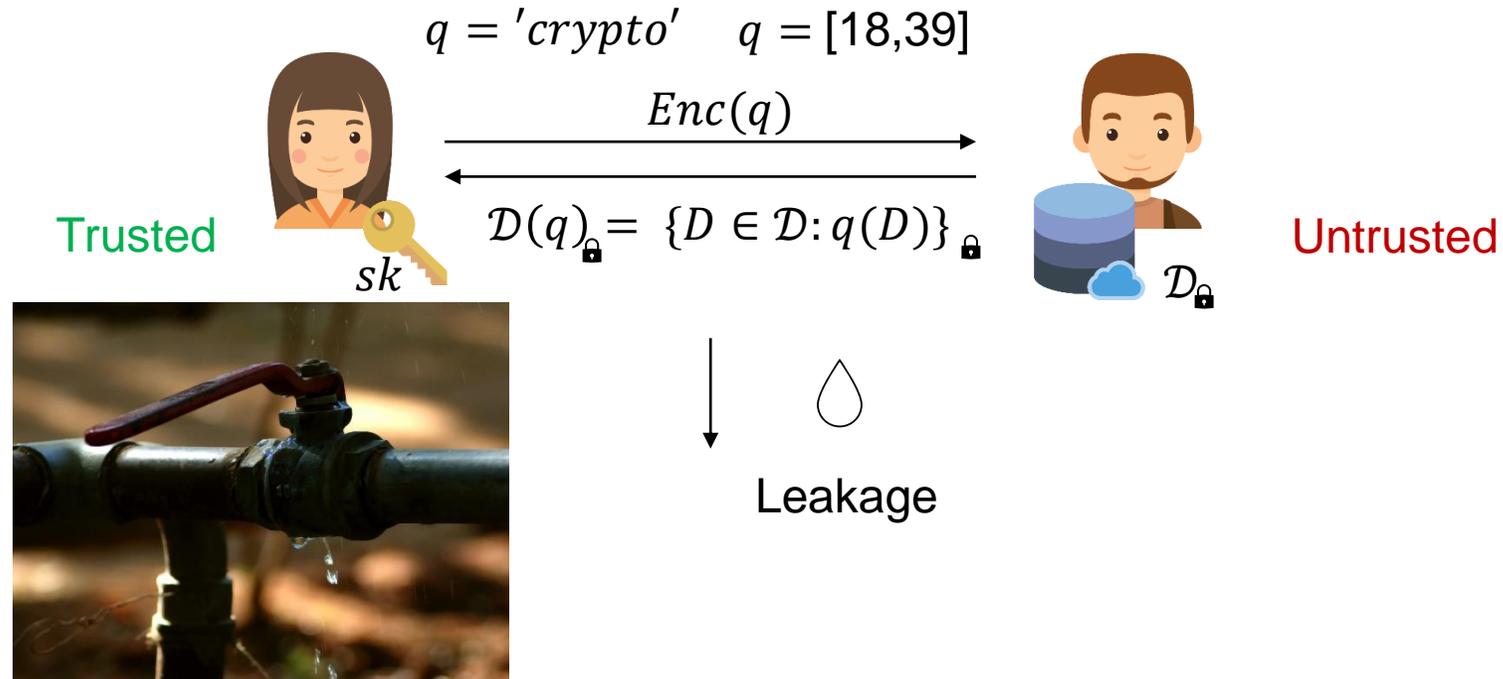
Encrypted Search Algorithms (ESAs)



Encrypted Search Algorithms (ESAs)



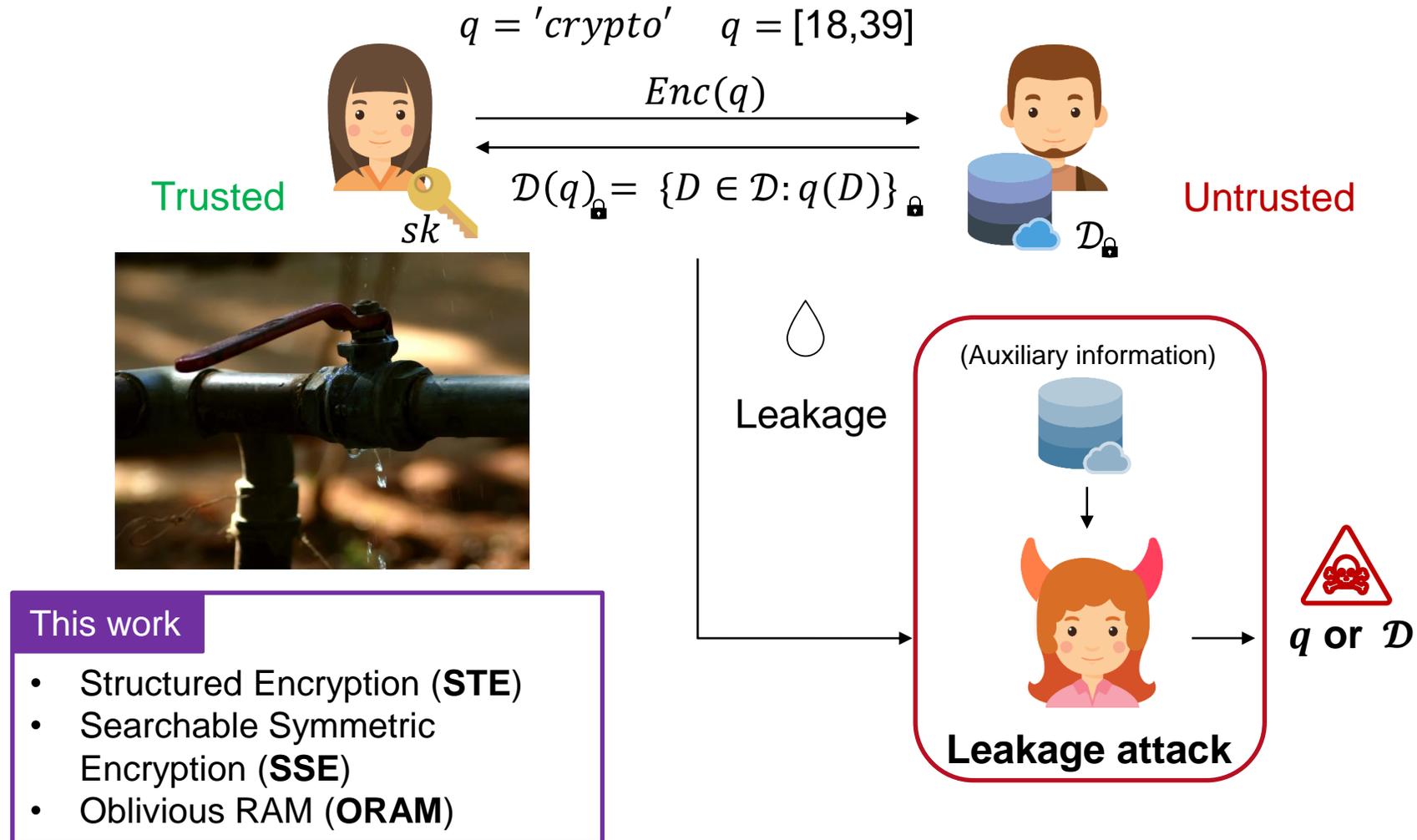
Encrypted Search Algorithms (ESAs)



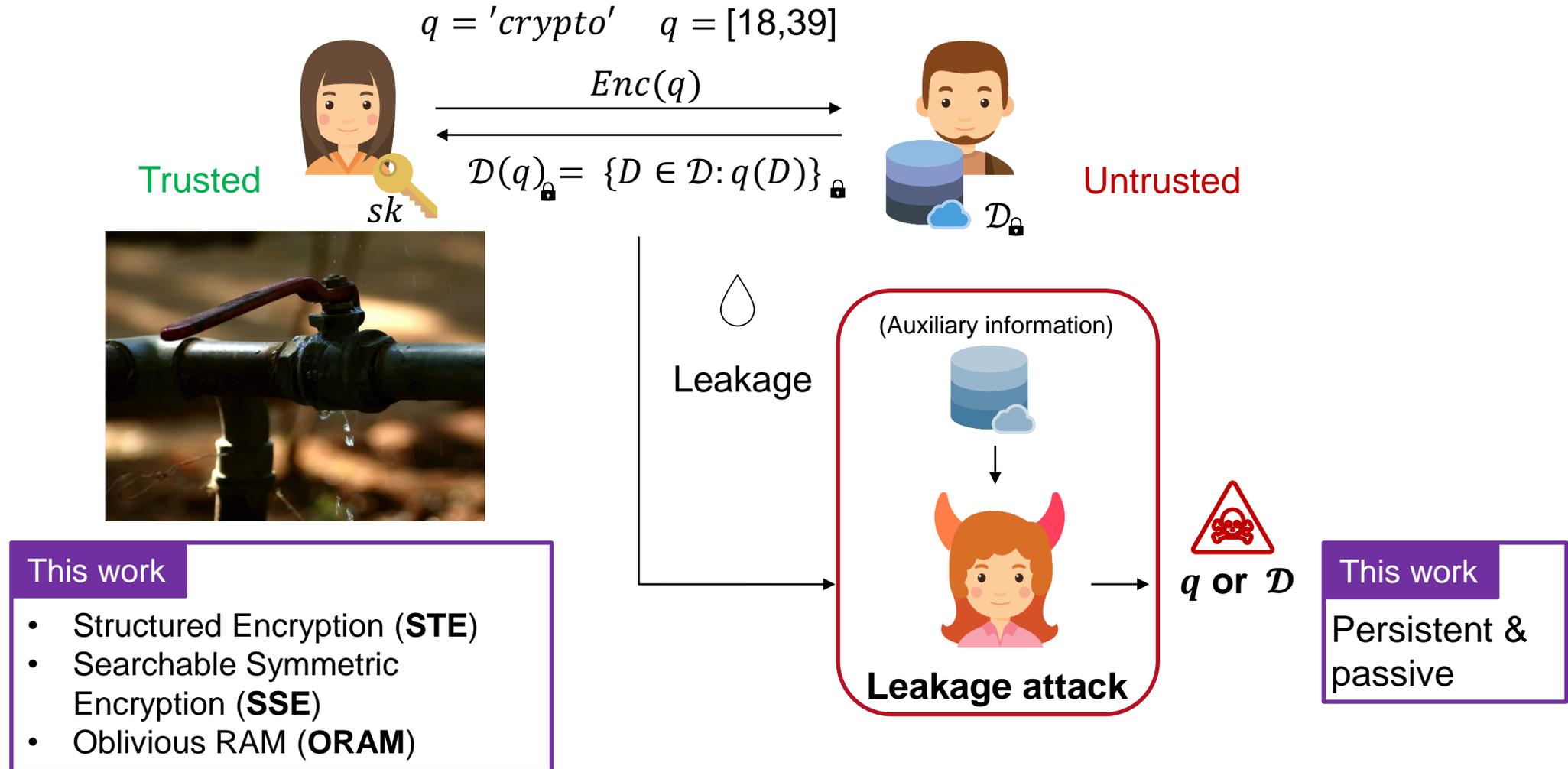
This work

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

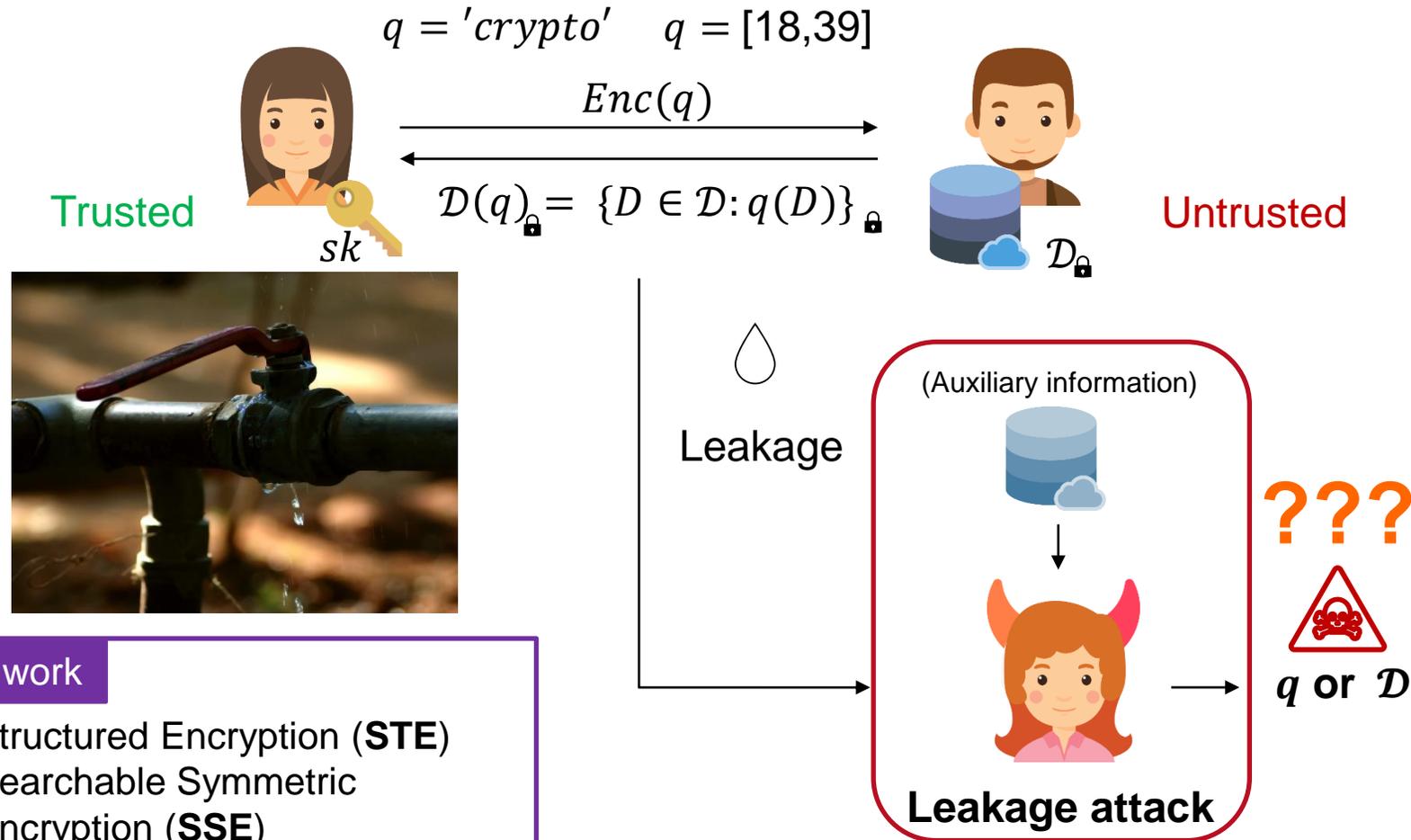
Encrypted Search Algorithms (ESAs)



Encrypted Search Algorithms (ESAs)



Encrypted Search Algorithms (ESAs)



Encrypted Search Algorithms (ESAs): Uncertainty Of Security



Constructions

Attacks & Countermeasures

Encrypted Search Algorithms (ESAs): Uncertainty Of Security

Constructions

Attacks & Countermeasures

“ Benign leakage ”

“ Common leakage ”

“ Standard leakage ”

“ Accepted leakage ”

Encrypted Search Algorithms (ESAs): Uncertainty Of Security

Constructions

Attacks & Countermeasures

“ Benign leakage ”

“ Common leakage ”

“ Standard 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] assume extremely strong adversarial models ”

“ Leakages [...] are not exploitable via leakage-abuse attacks in practice ”

Attacks & Countermeasures

“ Severe threat ”

“ Devastating results ”

“ [ESAs] are extremely vulnerable to [attacks] ”

“ [ESA] schemes should no longer be used without countermeasures ”

Encrypted Search Algorithms (ESAs): Uncertainty Of Security

Constructions

“ Benign leakage ”

“ Common leakage ”

“ Standard leakage ”

“ Accepted leakage ”

“ [Attacks] assume extremely strong adversarial models ”

“ Leakages [...] are not exploitable via leakage-abuse attacks in practice ”

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 ”

Encrypted Search Algorithms (ESAs): Uncertainty Of Security

Constructions

Attacks & Countermeasures

Benign leakage

Common

Severe threat

Devastating results

Standard leakage

Accepted

Extremely
attacks]

[ESA] schemes
should no longer be
used without
countermeasures

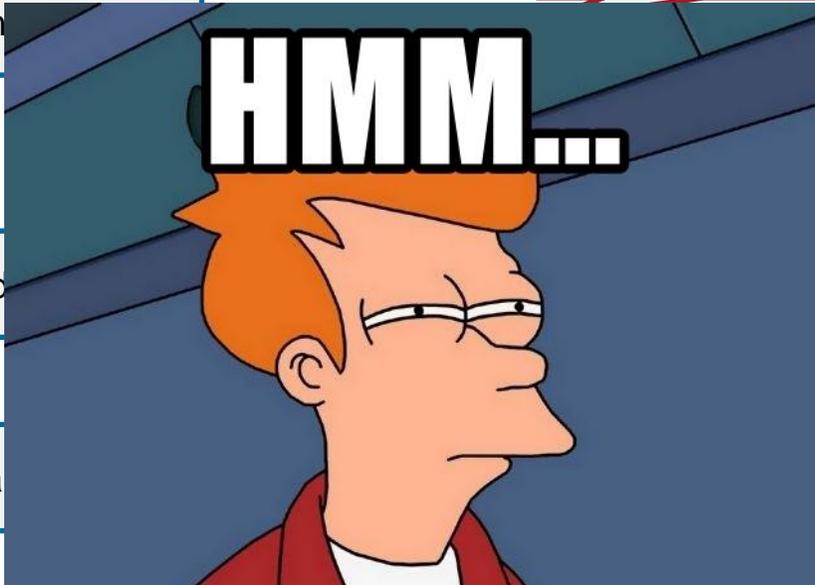
[Attacks] assume extremely strong adversa

on background information are weak

Leakages [...] are not exploitable via leakage-abuse
attacks in practice

With some prior knowledge [...] an honest-but-curious
server can recover the underlying keywords

HMM...



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

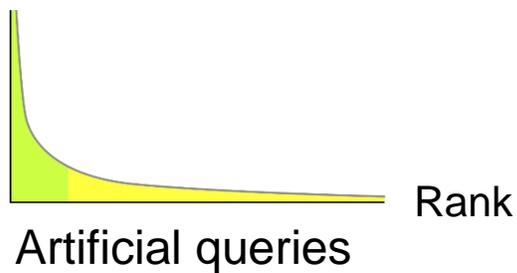


Few comparisons



Small/restricted data

Frequency



Rank

Previous Evaluations & Our Contributions

Previous evaluations



Closed-source code



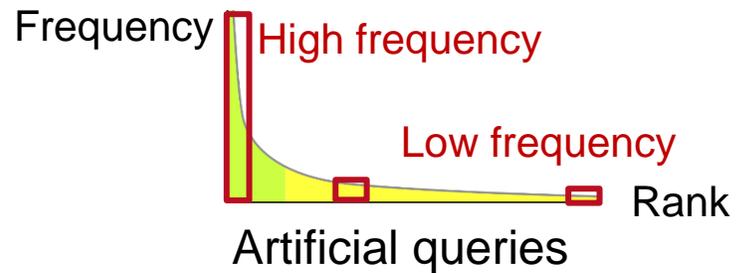
Single use case



Few comparisons



Small/restricted data



Previous Evaluations & Our Contributions

Previous evaluations



Closed-source code



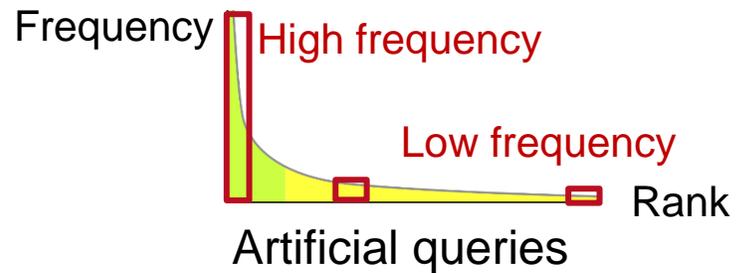
Single use case



Few comparisons



Small/restricted data



This work



Open-source
framework



Multiple use cases



Systematic re-
evaluation



Large data

Previous Evaluations & Our Contributions

Previous evaluations



Closed-source code



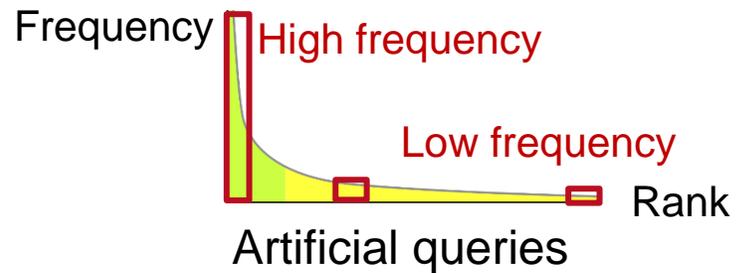
Single use case



Few comparisons



Small/restricted data



This work



Open-source
framework



Multiple use cases



Systematic re-
evaluation



Large data

```
User, Query  
216, 'crypto'  
216, 'amsterdam'  
106, 'doctor'  
216, 'hotel'
```

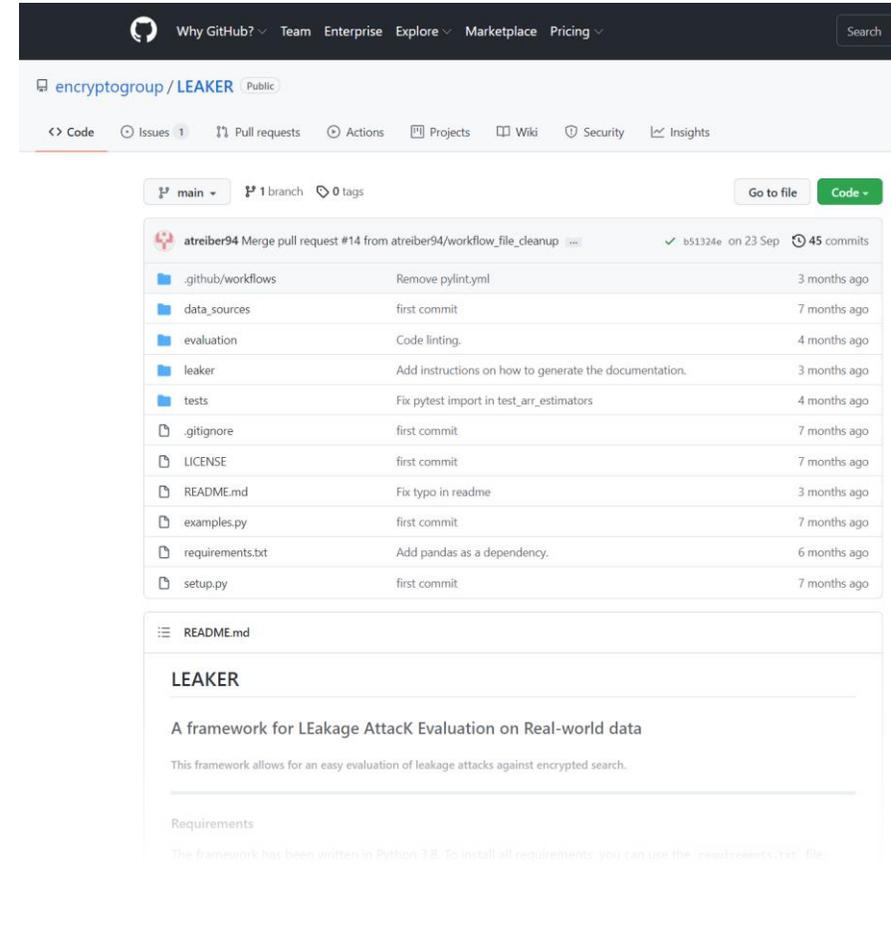
First real-world query logs

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>



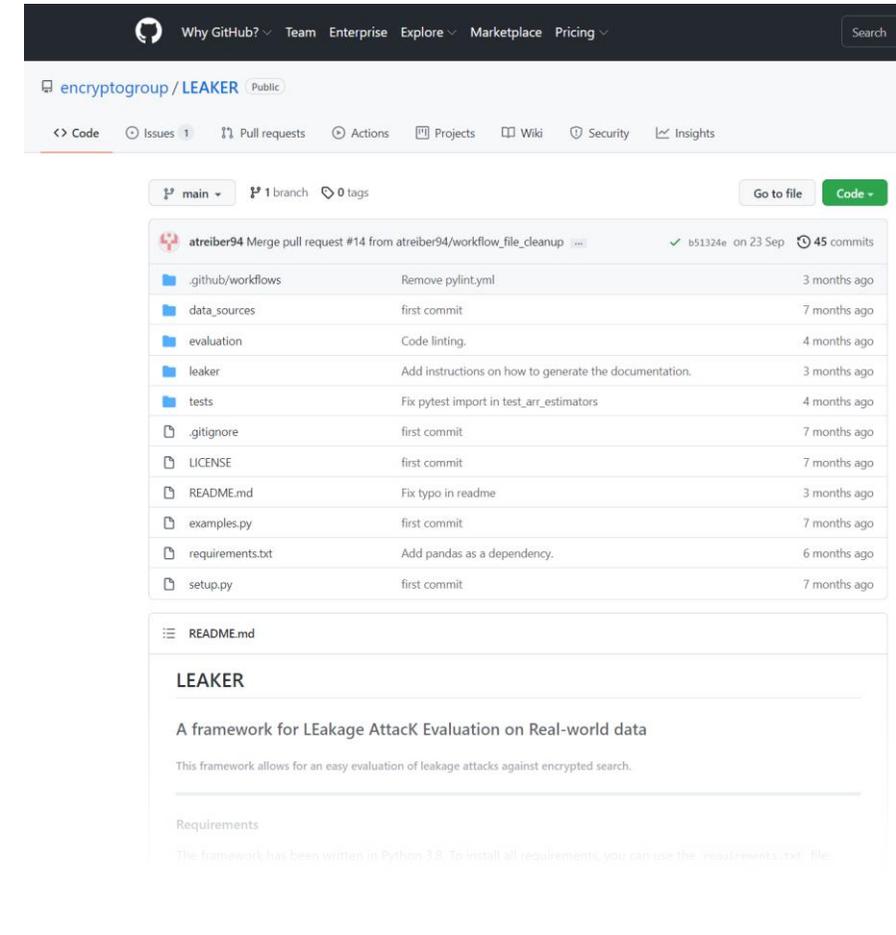
New Software: LEAKER



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

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

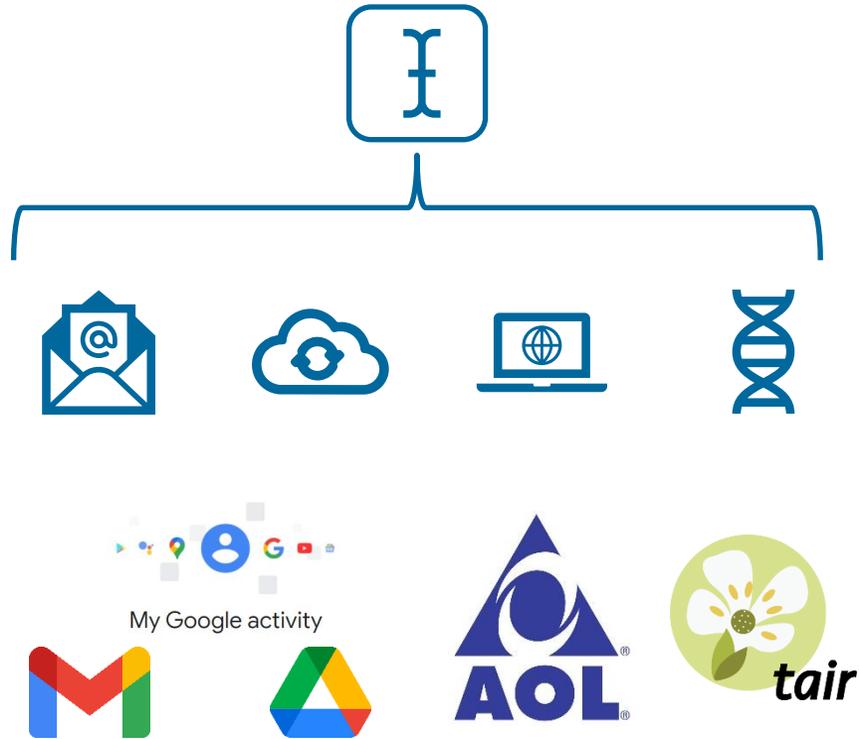
- Modular design & interoperability
- Easy to implement new attacks & countermeasures
- Easy to pre-process & use new data



<https://encrypto.de/code/LEAKER>



Keyword (*point*) queries



Keyword (*point*) queries



My Google activity



tair

Range queries



DATA.GOV.UK
Opening up Government



Keyword (*point*) queries



Have query logs

Range queries



Evaluation: Summary – Keyword Search



(subjective)

| Leakage  | Attack Success  | Risk  |
|---|--|--|
| <ul style="list-style-type: none">• Response length• Response volume | <ul style="list-style-type: none">• High adversarial knowledge | Low |
| <ul style="list-style-type: none">• Co-occurrence | <ul style="list-style-type: none">• High adversarial knowledge | Low |
| <ul style="list-style-type: none">• Response identifiers• Response volumes (of individual documents) | <ul style="list-style-type: none">• Low adversarial knowledge | High |

Evaluation: Summary – Keyword Search



(subjective)

| Leakage  | Attack Success  | Risk  |
|---|--|--|
| <ul style="list-style-type: none">• Response length• Response volume | <ul style="list-style-type: none">• High adversarial knowledge | Low |
| <ul style="list-style-type: none">• Co-occurrence | <ul style="list-style-type: none">• High adversarial knowledge | Low |
| <ul style="list-style-type: none">• Response identifiers• Response volumes (of individual documents) | <ul style="list-style-type: none">• Low adversarial knowledge | High |

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

Evaluation: Summary – Keyword Search



(subjective)

| Leakage  | Attack Success  | Risk  |
|---|--|--|
| <ul style="list-style-type: none">Response lengthResponse volume | <ul style="list-style-type: none">High adversarial knowledge | Low |
| <ul style="list-style-type: none">Co-occurrence | <ul style="list-style-type: none">High adversarial knowledge | Low |
| <ul style="list-style-type: none">Response identifiersResponse volumes (of individual documents) | <ul style="list-style-type: none">Low adversarial knowledge | High |

Subgraph attacks [BKM20]

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

Evaluation: Highlights – Keyword Search



“ ”

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

[BKM20]

“ ”

Users are more likely to search for a
specific email

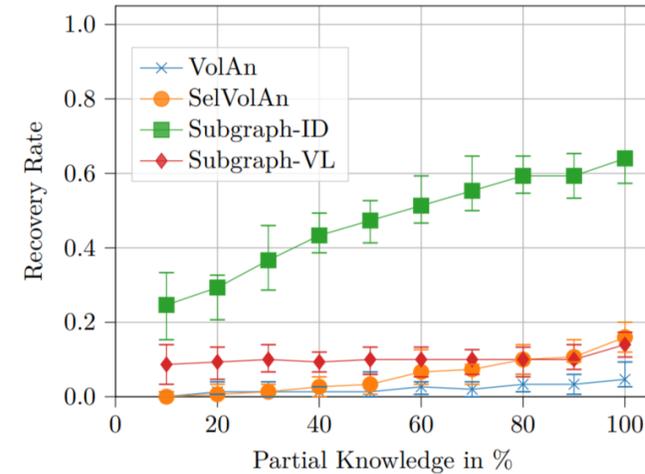
[RPH21]

Evaluation: Highlights – Keyword Search



“
None of the attacks worked against low-
[frequency] keywords
”

[BKM20]



Mean
frequency:
1.54!
(on TAIR)

“
Users are more likely to search for a
specific email
”

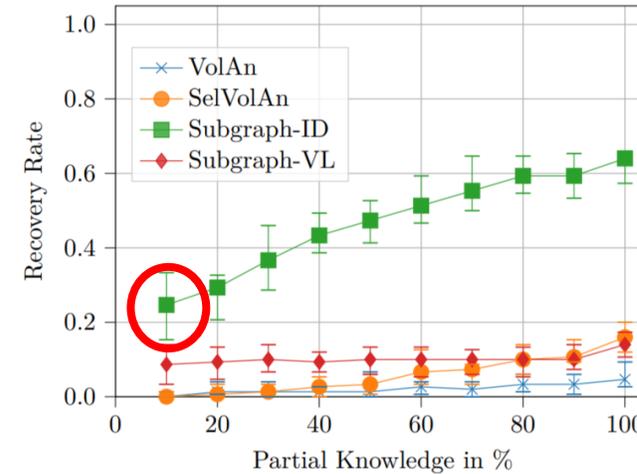
[RPH21]

Evaluation: Highlights – Keyword Search



“
None of the attacks worked against low-
[frequency] keywords
”

[BKM20]



Mean
frequency:
1.54!
(on TAIR)

“
Users are more likely to search for a
specific email
”

[RPH21]

Evaluation: Highlights – Keyword Search

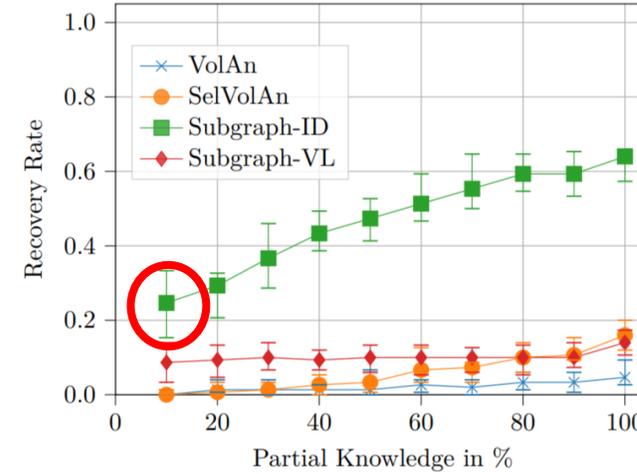


“
None of the attacks worked against low-
[frequency] keywords
”

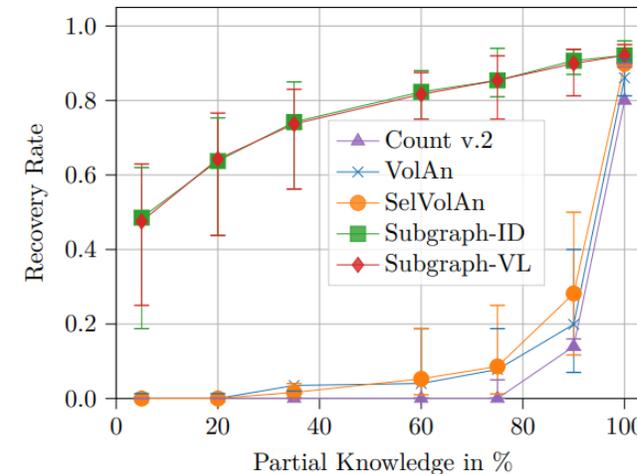
[BKM20]

“
Users are more likely to search for a
specific email
”

[RPH21]



Mean
frequency:
1.54!
(on TAIR)



Mean
frequency:
326!
(on GMail)

Evaluation: Highlights – Keyword Search

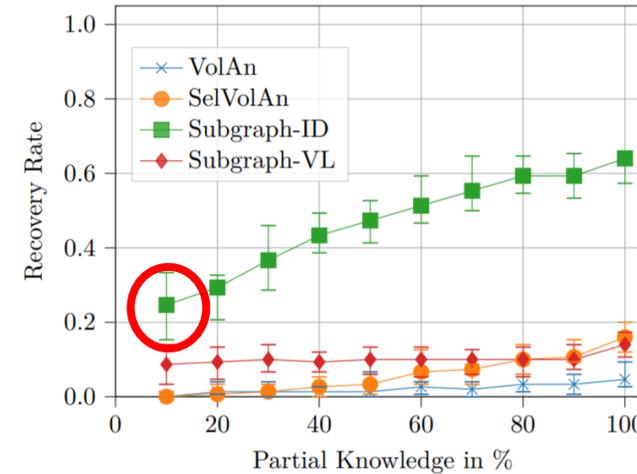


“
None of the attacks worked against low-
[frequency] keywords
”

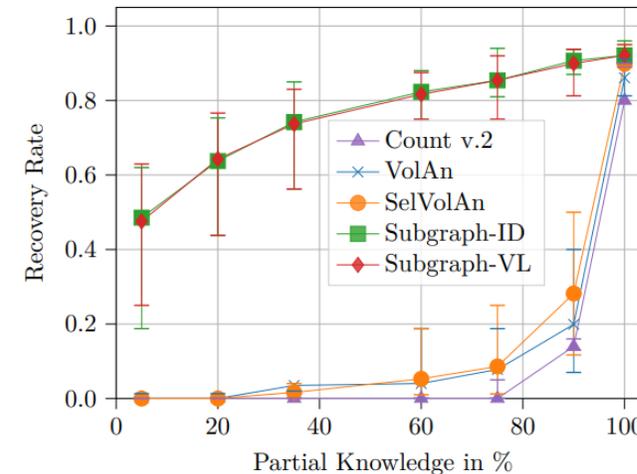
[BKM20]

“
Users are more likely to search for a
specific email
”

[RPH21]



Mean
frequency:
1.54!
(on TAIR)



Mean
frequency:
326!
(on GMail)

Evaluation: Summary – Range Search



(subjective)

| Leakage  | Attack Success  | Risk  |
|--|--|--|
| <ul style="list-style-type: none">• Response length | <ul style="list-style-type: none">• None | Very low |
| <ul style="list-style-type: none">• Response length• Query equality | <ul style="list-style-type: none">• Evenly distributed data | Medium |
| <ul style="list-style-type: none">• Co-occurrence | <ul style="list-style-type: none">• Large widths• Skewed values | Medium |
| <ul style="list-style-type: none">• Co-occurrence• Order | <ul style="list-style-type: none">• Most cases | High |

Evaluation: Summary – Range Search



(subjective)

| Leakage  | Attack Success  | Risk  |
|--|--|--|
| <ul style="list-style-type: none">• Response length | <ul style="list-style-type: none">• None | Very low |
| <ul style="list-style-type: none">• Response length• Query equality | <ul style="list-style-type: none">• Evenly distributed data | Medium |
| <ul style="list-style-type: none">• Co-occurrence | <ul style="list-style-type: none">• Large widths• Skewed values | Medium |
| <ul style="list-style-type: none">• Co-occurrence• Order | <ul style="list-style-type: none">• Most cases | High |

=> 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

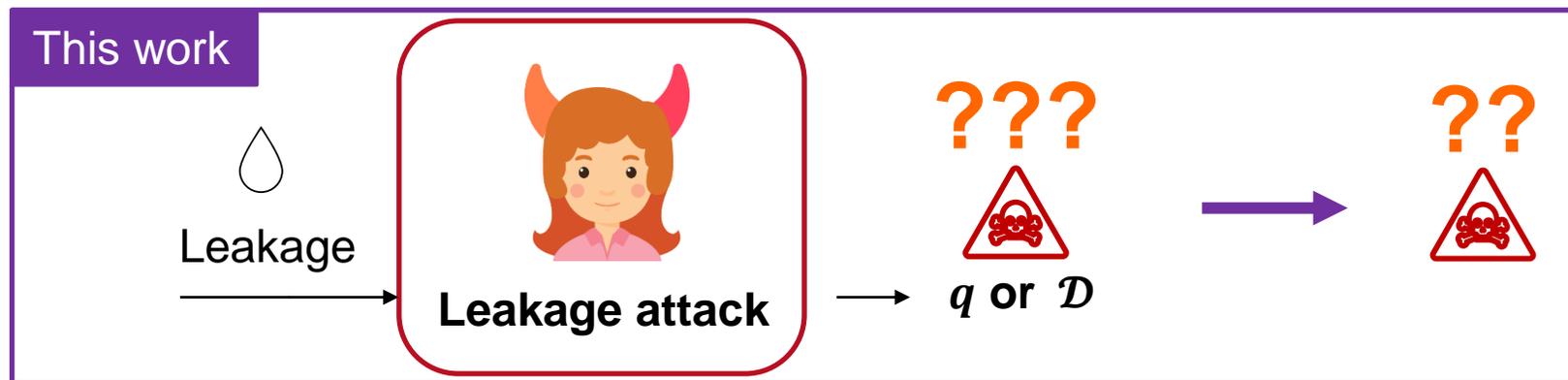
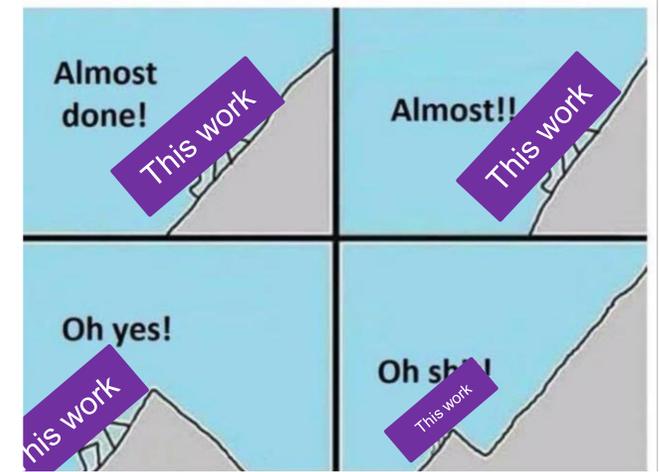
Conclusions

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



Conclusions

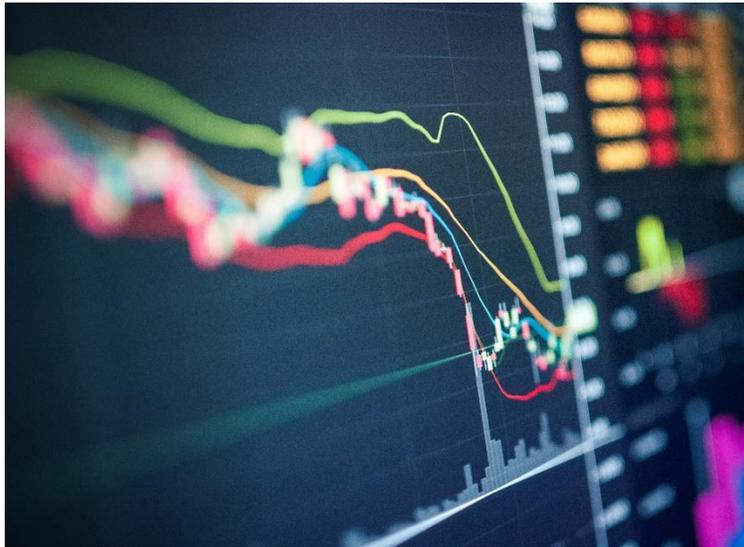
- Extensible **open-source** framework LEAKER
- **First** usage of real-world queries
- **Systematic** empirical analysis of leakage attacks
- **Contradict** some previous conclusions



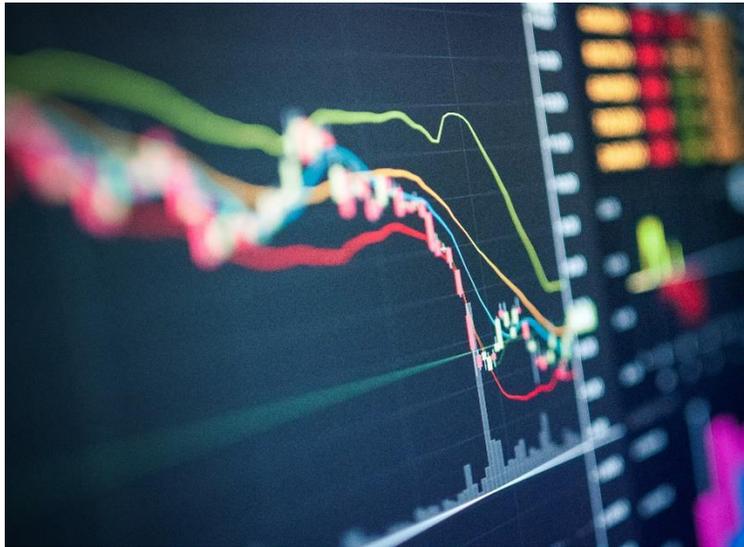
What needs to be done



What needs to be done



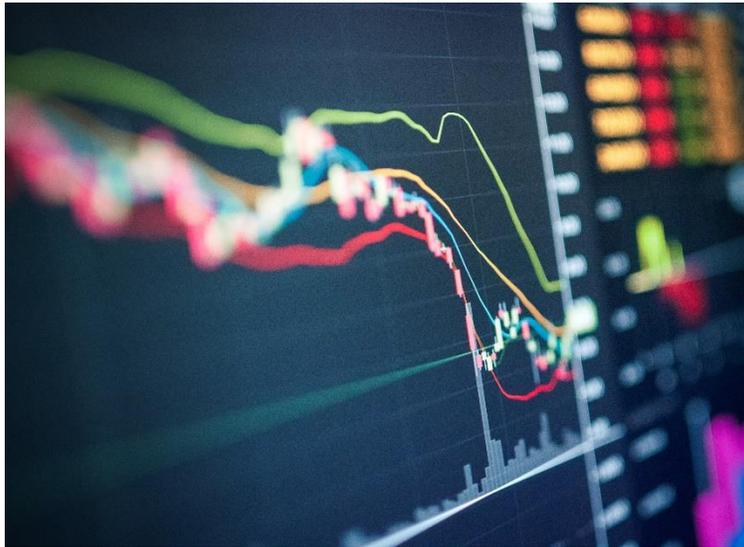
What needs to be done



+



What needs to be done

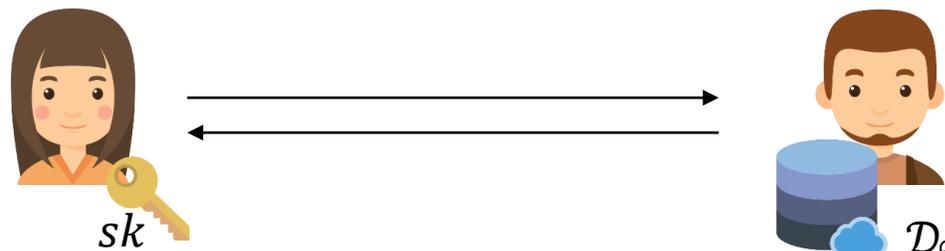


+



⇒





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>



- Icons & pics by *Flaticons (FreePik, Becris, Darius Dan, Surang, Vectors Market, Becris), FreePNG, PNGItem, <https://memegenerator.net/>, Futurama - "The Lesser of Two Evils", 2011 by 20th Television, Rawpixel.com / Shutterstock*
- [BKM20] Laura Blackstone, Seny Kamara, and Tarik Moataz. Revisiting leakage abuse attacks. In Network and Distributed System Security Symposium (NDSS), 2020
- [CGPR15] David Cash, Paul Grubbs, Jason Perry, and Thomas Ristenpart. Leakage-abuse attacks against searchable encryption. In ACM SIGSAC Conference on Computer and Communications Security (CCS), 2015.
- [DHP21] Marc Damie, Florian Hahn, and Andreas Peter. A Highly Accurate Query-Recovery Attack against Searchable Encryption using Non-Indexed Documents. In USENIX Security Symposium (USENIX Security), 2021.
- [GLMP18] Paul Grubbs, Marie-Sarah Lacharité, Brice Minaud, and Kenneth G Paterson. Pump up the volume: Practical database reconstruction from volume leakage on range queries. In ACM SIGSAC Conference on Computer and Communications Security (CCS), 2018.
- [GLMP19] Paul Grubbs, Marie-Sarah Lacharité, Brice Minaud, and Kenneth G Paterson. Learning to reconstruct: Statistical learning theory and encrypted database attacks. In IEEE Symposium on Security and Privacy (S&P), 2019.
- [GJW19] Zichen Gui, Oliver Johnson, and Bogdan Warinschi. Encrypted databases: New volume attacks against range queries. In ACM SIGSAC Conference on Computer and Communications Security (CCS), 2019.
- [GPP21] Zichen Gui, Kenneth G Paterson, and Sikhar Patranabis. Leakage Perturbation is Not Enough: Breaking Structured Encryption Using Simulated Annealing. In IACR ePrint, 879, 2021
- [IKK12] Mohammad Saiful Islam, Mehmet Kuzu, and Murat Kantarcioglu. Access pattern disclosure on searchable encryption: Ramification, attack and mitigation. In Network and Distributed System Security Symposium (NDSS), 2012.
- [KKNO16] Georgios Kellaris, George Kollios, Kobbi Nissim, and Adam O'Neill. Generic attacks on secure outsourced databases. In ACM SIGSAC Conference on Computer and Communications Security (CCS), 2016

- [KPT20] Evgenios M Kornaropoulos, Charalampos Papamanthou, and Roberto Tamassia. The state of the uniform: Attacks on encrypted databases beyond the uniform query distribution. In IEEE Symposium on Security and Privacy (S&P), 2020.
- [KPT21] Evgenios M Kornaropoulos, Charalampos Papamanthou, and Roberto Tamassia. Response-hiding encrypted ranges: Revisiting security via parametrized leakage-abuse attacks. In IEEE Symposium on Security and Privacy (S&P), 2021.
- [LMP18] Marie-Sarah Lacharité, Brice Minaud, and Kenneth G Paterson. Improved reconstruction attacks on encrypted data using range query leakage. In IEEE Symposium on Security and Privacy (S&P), 2018.
- [LZWT14] Chang Liu, Liehuang Zhu, Mingzhong Wang, and Yu-An Tan. Search pattern leakage in searchable encryption: Attacks and new construction. Information Sciences, 265, 2014.
- [NKW15] Muhammad Naveed, Seny Kamara, and Charles V Wright. Inference attacks on property-preserving encrypted databases. In ACM SIGSAC Conference on Computer and Communications Security (CCS), 2015.
- [OK21a] Simon Oya and Florian Kerschbaum. Hiding the access pattern is not enough: Exploiting search pattern leakage in searchable encryption. In USENIX Security Symposium (USENIX Security), 2021.
- [OK21b] Simon Oya and Florian Kerschbaum. IHOP: Improved Statistical Query Recovery against Searchable Symmetric Encryption through Quadratic Optimization. In arXiv 2110.04180, 2021.
- [PWLP20] Rishabh Poddar, Stephanie Wang, Jianan Lu, and Raluca Ada Popa. Practical volume-based attacks on encrypted databases. In IEEE European Symposium on Security and Privacy (EuroS&P), 2020.
- [RPH21] Ruben Groot Roessink, Andreas Peter, and Florian Hahn. Experimental review of the IKK query recovery attack: Assumptions, recovery rate and improvements. In International Conference on Applied Cryptography and Network Security (ACNS), 2021.
- [ZKP16] Yupeng Zhang, Jonathan Katz, and Charalampos Papamanthou. All your queries are belong to us: The power of file-injection attacks on searchable encryption. In USENIX Security Symposium (USENIX Security), 2016.

| Leakage  | Information |
|--|-----------------------------|
| Response Length | $ D(q) $ |
| Query Equality | $q_i = q_j$ |
| Co-Occurrence | $ D(q_i) \cap D(q_j) $ |
| Response Identifiers | $\{i: D_i \in q(D)\}$ |
| Response Volumes | $\{ D_i _b: D_i \in q(D)\}$ |

(Simplified)

Leakage Attacks Types



Keyword (point) queries
[IKK12,CGPR15,BKM20,RPH21]



| Keyword | Document IDs |
|----------|-----------------|
| 'real' | 2,5,11,13,20,31 |
| 'world' | 3,5,10,11,13,25 |
| 'crypto' | 5,11,21,27 |

$$q = w$$

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

Recover q $q = \text{'crypto'}$

Known-data: Adversary knows subset of \mathcal{D}



Range queries
[KKNO16,LMP18,GLMP18,
GLMP19,GJW19,KPT20,KPT21]



| ID | Age |
|----|-----|
| 1 | 65 |
| 2 | 7 |
| 3 | 27 |

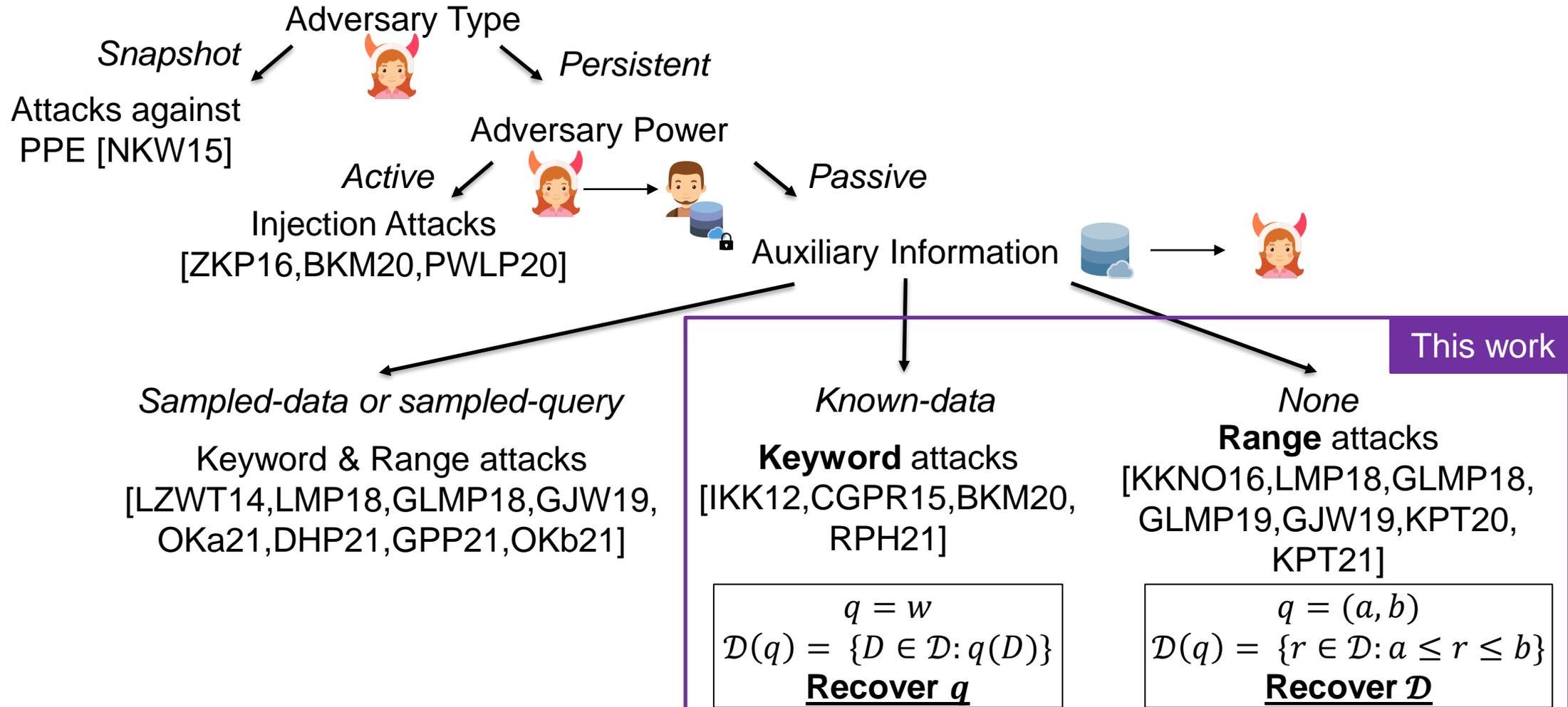
$$q = (a, b)$$

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

Recover \mathcal{D} $q = (18, 39)$

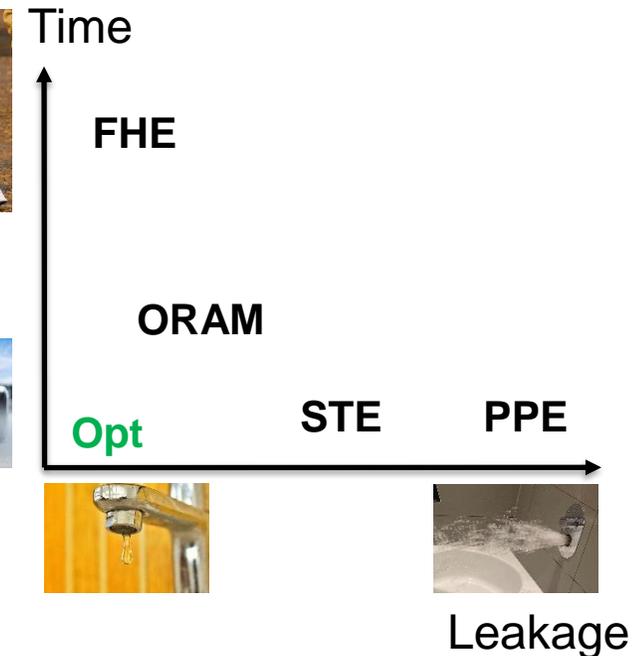
No auxiliary knowledge

Overview of Leakage Attacks on ESAs



Overview of Techniques for ESAs (Extremely informal)

| Technique | Leakage  | Query Time | |
|--------------------------------------|---|------------|---|
| Fully Homomorphic Encryption (FHE) | <ul style="list-style-type: none"> None | Linear |  <p>Considered secure but inefficient</p> |
| Oblivious RAM (ORAM) | <ul style="list-style-type: none"> Response Length | Sublinear | |
| Structured Encryption (STE) | <ul style="list-style-type: none"> Query Equality Responses' Equality | Optimal |  <p>This work</p> <p>Considered efficient and ???</p> |
| Property-Preserving Encryption (PPE) | <ul style="list-style-type: none"> Ciphertext Equality Ciphertext Order | Optimal | |



Previous Evaluations



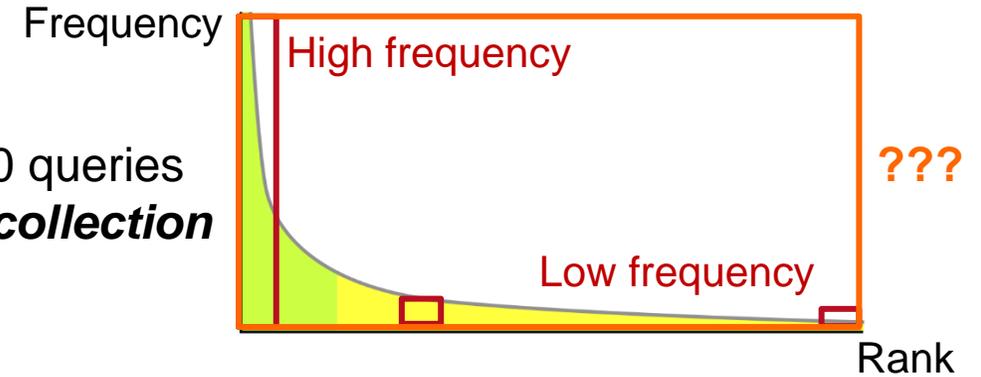
- Usual evaluations for keyword attacks:



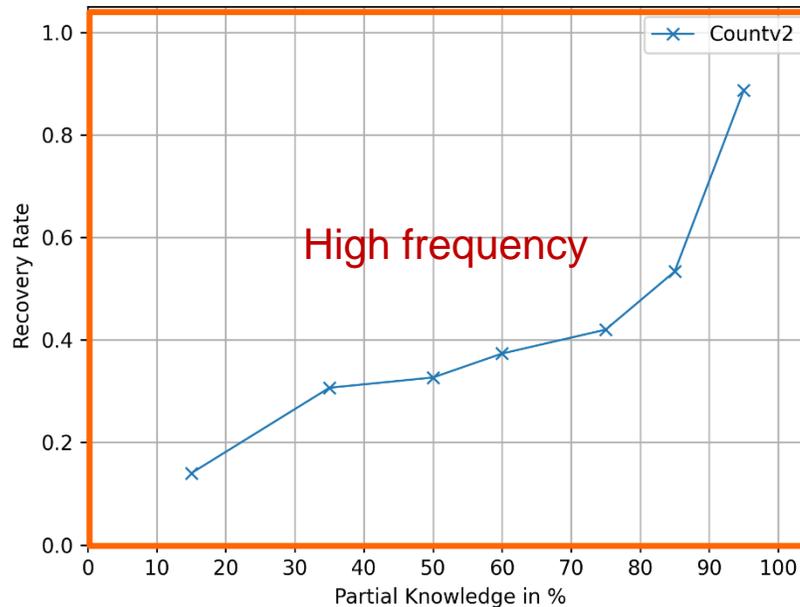
1. Enron (& Apache) email data collection

2. Restrict data to 500-3000 keywords

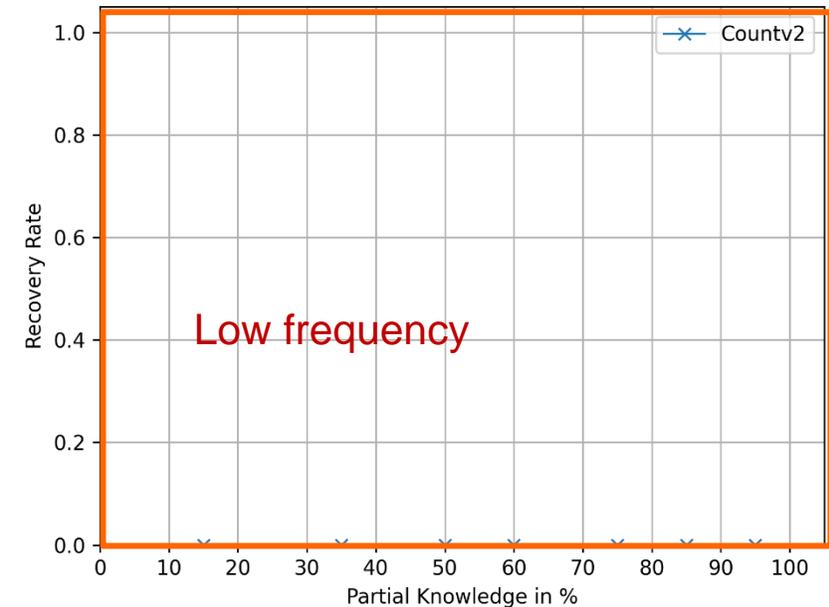
3. Draw 150 queries *from data collection*



4. Evaluate on partial knowledge



or
???



Previous Evaluations



- Usual evaluations for range attacks:

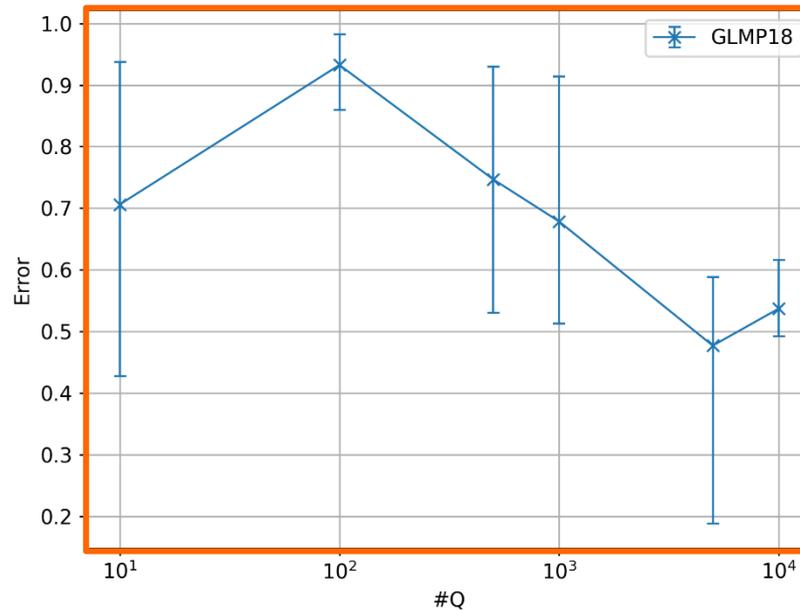


H-CUP
HEALTHCARE COST AND UTILIZATION PROJECT

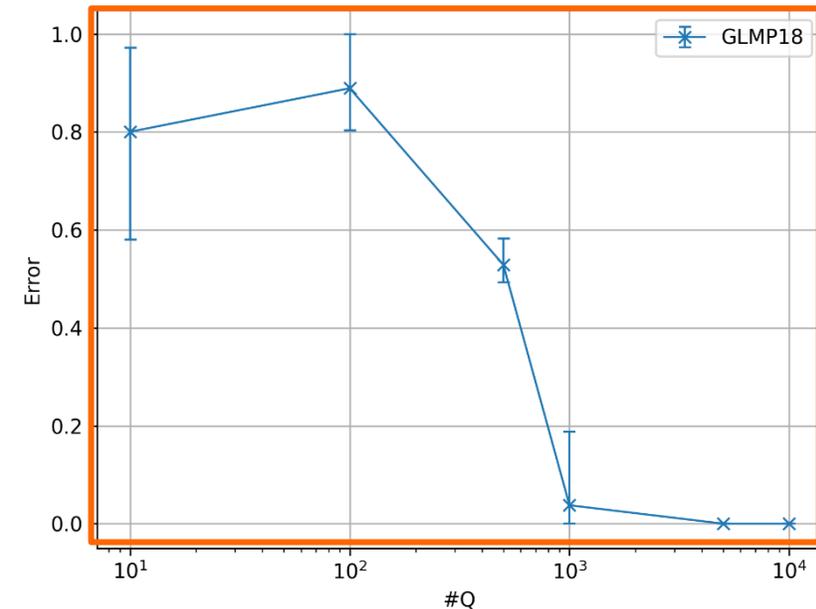
1. Subset of HCUP or artificial Data collection

2. Pick Artificial query distribution (Uniform/Zipf/...)

3. Evaluate for different amounts of queries



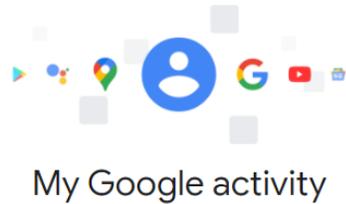
or
???





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

Use Case: *Email/Cloud*



The activity that you keep helps Google make services more useful for you, like helping you rediscover the things that you've searched for, read and watched.
You can see and delete your activity using the controls on this page.

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



Medical



Human Resources



Sales



Insurance



Sloan Digital Sky Survey

- 3 Query Logs & 1 Data Collection
- 3 Users
- 215-8k Queries
- 5M Records
- Domain $N = 10k$
- Density 96%

Medical Information Mart for Intensive Care

- 3 Data Collections
- 2k-8k Records
- Domain $N = 73 - 10k$
- Density 3.3%-81%

Salaries of the UK Attorney General's Office junior civil servants

- 1 Data Collection
- 536 Records
- Domain $N = 395$
- Density 2.3%

Walmart Sales Data

- 1 Data Collection
- 143 Records
- Domain $N = 6.3k$
- Density 2.3%

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).

| Instance | GKKNO | AVALUE | ARR | ARR-OR | APA-OR ^{BT} | APA-OR ^{ABT} |
|----------|-------|--------|-------|--------|----------------------|-----------------------|
| SDSS-S | 0.413 | 0.432 | 0.473 | 0.249 | 0.242 | 0.239 |
| SDSS-M | 0.408 | 0.435 | 0.287 | 0.128 | 0.242 | 0.240 |
| SDSS-L | 0.417 | 0.456 | 0.286 | 0.141 | 0.241 | 0.242 |