# Frequency-smoothing encryption

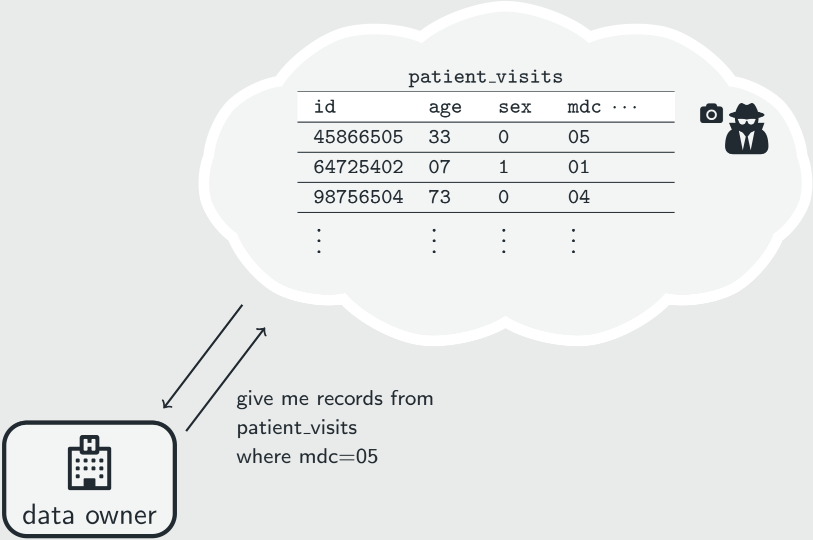
## Preventing snapshot attacks on deterministically encrypted data

**Marie-Sarah Lacharité** and Kenneth G. Paterson

Information Security Group, Royal Holloway, University of London

6 March 2018
FSE 2018, Bruges

# Outsourced database storage

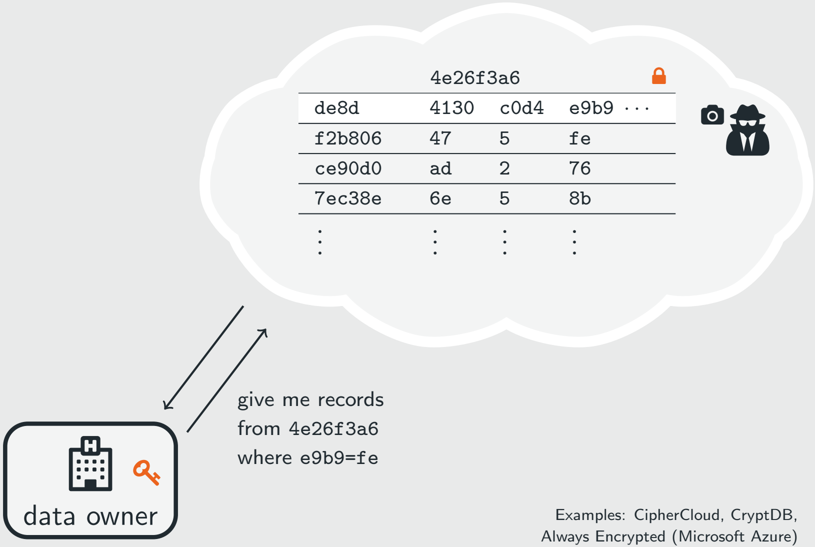


`patient_visits`

| id | age | sex | mdc ··· |
|---|---|---|---|
| 45866505 | 33 | 0 | 05 |
| 64725402 | 07 | 1 | 01 |
| 98756504 | 73 | 0 | 04 |
| ⋮ | ⋮ | ⋮ | ⋮ |

give me records from
patient_visits
where mdc=05

# Outsourced database storage with deterministic encryption



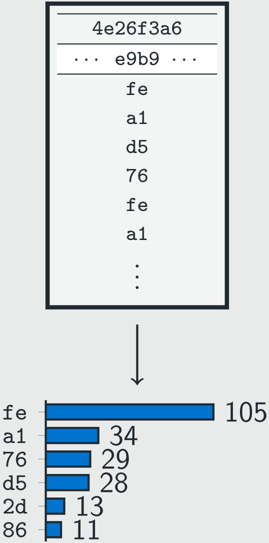
4e26f3a6

| de8d | 4130 | c0d4 | e9b9 ··· |
|---|---|---|---|
| f2b806 | 47 | 5 | fe |
| ce90d0 | ad | 2 | 76 |
| 7ec38e | 6e | 5 | 8b |
| ⋮ | ⋮ | ⋮ | ⋮ |

give me records
from 4e26f3a6
where e9b9=fe

Examples: CipherCloud, CryptDB,
Always Encrypted (Microsoft Azure)

# Inference attacks: an example



# Inference attacks: an example

4e26f3a6

··· e9b9 ···

fe

a1

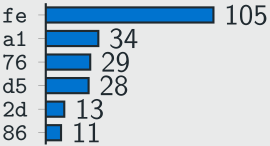
d5

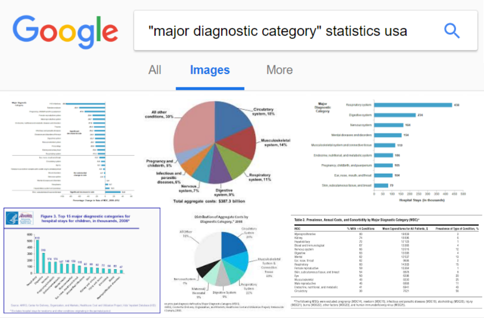
76

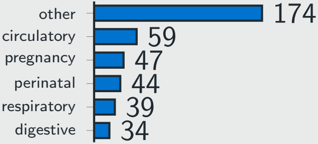
fe

a1

⋮

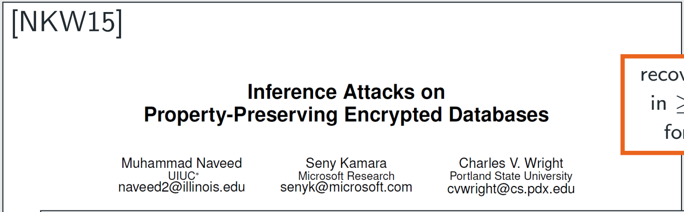



"major diagnostic category" statistics usa



# Inference attacks

[NKW15]



**Inference Attacks on**
**Property-Preserving Encrypted Databases**

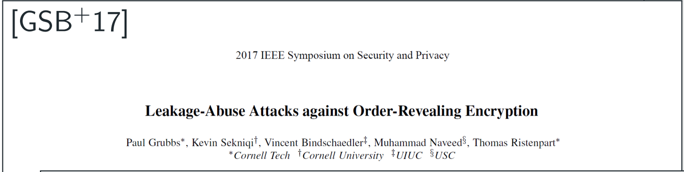
Muhammad Naveed
UIUC*
naveed2@illinois.edu

Seny Kamara
Microsoft Research
senyk@microsoft.com

Charles V. Wright
Portland State University
cvwright@cs.pdx.edu

recovered MDC values in $\geq$ 20% of records for 75% hospitals

[GSB+17]



2017 IEEE Symposium on Security and Privacy

**Leakage-Abuse Attacks against Order-Revealing Encryption**

Paul Grubbs*, Kevin Sekniqi†, Vincent Bindschaedler‡, Muhammad Naveed§, Thomas Ristenpart*
*Cornell Tech †Cornell University ‡UIUC §USC

[PW16]

**The Shadow Nemesis: Inference Attacks on Efficiently Deployable, Efficiently Searchable Encryption**

David Pouliot
Portland State University
Portland, OR 97207

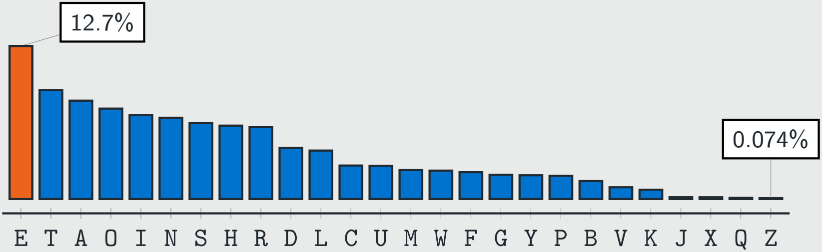
Charles V. Wright
Portland State University
Portland, OR 97207

# Overview of our results

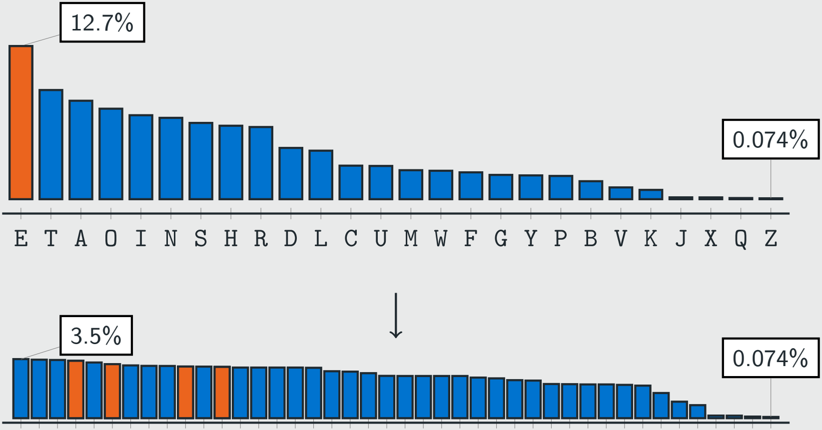
- frequency-smoothing (FS) encryption framework
- construction from homophonic encoding (HE) and deterministic encryption (DE)
- analytical and experimental evaluation of smoothness
- 8-bit FS encoding: recover $\geq$ 20% of MDC values for only 2% of hospitals
    - when *exact* distribution is known

# Frequency-smoothing encryption

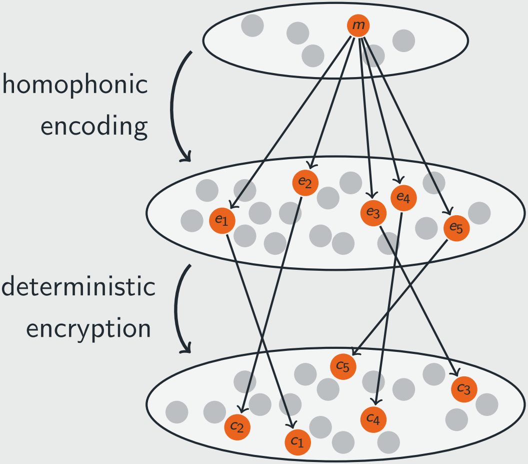
# Inspiration: homophonic encoding (HE)



# Inspiration: homophonic encoding (HE)



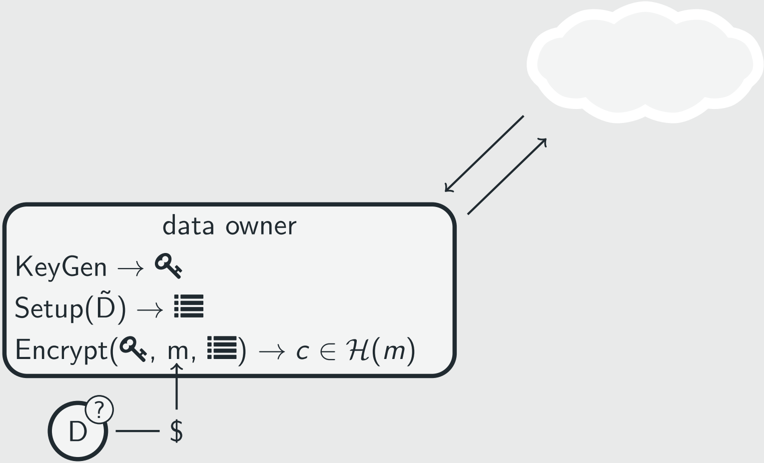
# FS encryption from HE and DE



homophonic encoding

deterministic encryption

# Frequency-smoothing (FS) encryption



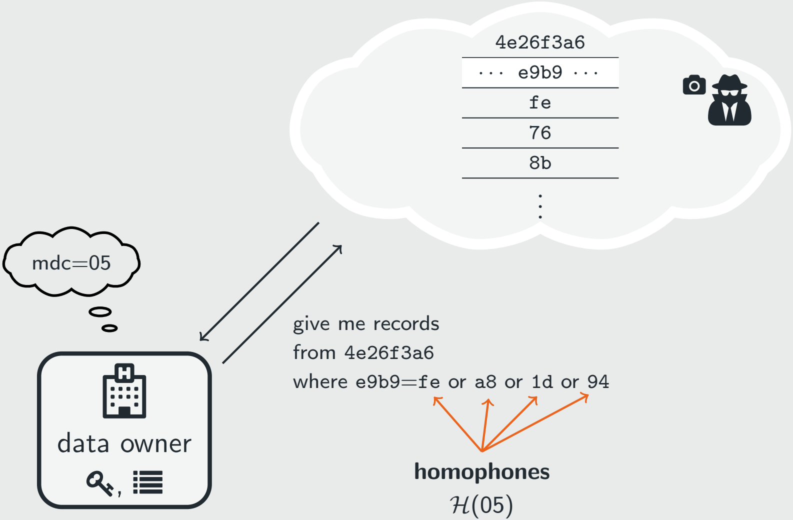
data owner

KeyGen → 🔑

Setup($\tilde{\mathrm{D}}$) → ☰

Encrypt(🔑, m, ☰) → $c \in \mathcal{H}(m)$

# Outsourced database storage with FS encryption



| 4e26f3a6 |
|---|
| ··· e9b9 ··· |
| fe |
| 76 |
| 8b |
| ⋮ |

give me records
from `4e26f3a6`
where `e9b9`=`fe` or `a8` or `1d` or `94`

**homophones**
$\mathcal{H}(05)$

# Frequency-smoothing (FS) encryption security

- adversary has its own estimate $\hat{\mathrm{D}}$ of the data's distribution
- *FS smoothness:* $\mathcal{A}$ gets $\{c_1, \ldots, c_N\}$, $\tilde{\mathrm{D}}$, $\hat{\mathrm{D}}$
  - are the $N$ ciphertexts (i) real – generated by a FS encryption scheme with $\tilde{\mathrm{D}}$, or (ii) fake – sampled from a set of size $|\mathcal{H}|$ uniformly at random?
- *FS message privacy:* $\mathcal{A}$ gets $\{(m_1, c_1), \ldots, (m_N, c_N)\}$, $\tilde{\mathrm{D}}$, $\hat{\mathrm{D}}$
  - are the $N$ ciphertexts (i) real – generated by a FS encryption scheme with $\tilde{\mathrm{D}}$, or (ii) fake – sampled from a set of size $|\mathcal{H}(m_i)|$ uniformly at random?

# FS encryption from HE and DE: security

$$\left\{\begin{array}{c}\text{HE smoothness}\\ +\\ \text{DE message privacy}\end{array}\right\} \Longrightarrow \left\{\begin{array}{c}\text{FS smoothness}\\ +\\ \text{FS message privacy}\end{array}\right\}$$

- *HE smoothness:* $\mathcal{A}$ gets $\{e_1, \ldots, e_N\}$, $\tilde{\mathsf{D}}$, $\hat{\mathsf{D}}$
    - are the $N$ encodings (i) real – generated by an HE scheme with $\tilde{\mathsf{D}}$, or (ii) fake – sampled from the set $\mathcal{H}$ uniformly at random?
- *DE message privacy:* similar to IND$ [Rog04]
    - could instantiate with small-domain PRP, format-preserving encryption, or synthetic IV mode [RS06]

# HE smoothness when $\mathsf{D}$ is known

- distribution known by all: $\mathsf{D} = \tilde{\mathsf{D}} = \hat{\mathsf{D}}$
  - so distribution $\mathsf{D}_e$ of encoded data depends only on $\mathsf{D}$
- $\mathcal{A}$ must distinguish $\mathsf{D}_e$ from uniform given $N$ samples
- apply optimal distinguisher analysis from [BJV04]

**Theorem**

*For any* HE−SMOOTH *adversary* $\mathcal{A}$ *and sufficiently large* $N$*,*

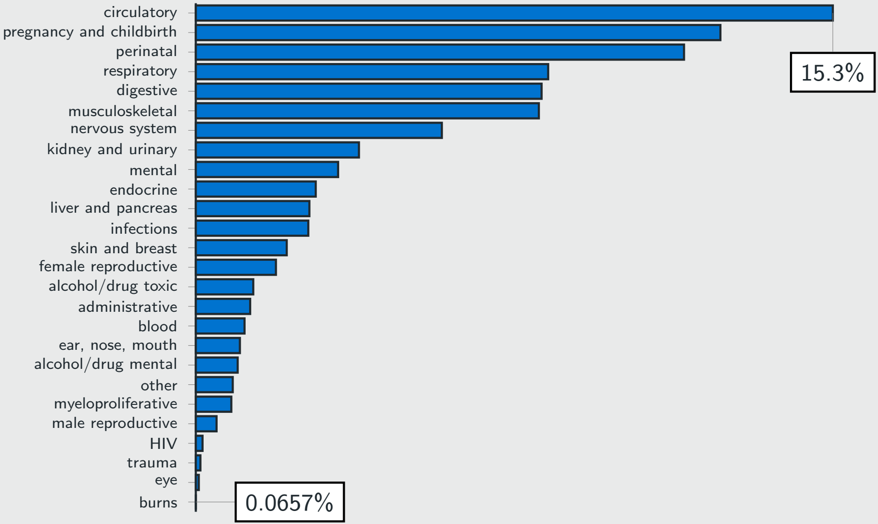
$$\mathrm{Adv}(\mathcal{A}, \mathsf{D}, N) \leq \left| \frac{1}{2} - \Phi\left( -\sqrt{\frac{N \cdot (\log|\mathcal{H}| - H_0(\mathsf{D}_e))}{2}} \right) \right|$$

*where* $\Phi(\cdot)$ *is cdf of the standard normal distribution and* $H_0(\cdot)$ *is Shannon entropy.*

# Interval-based homophonic encoding (IBHE)

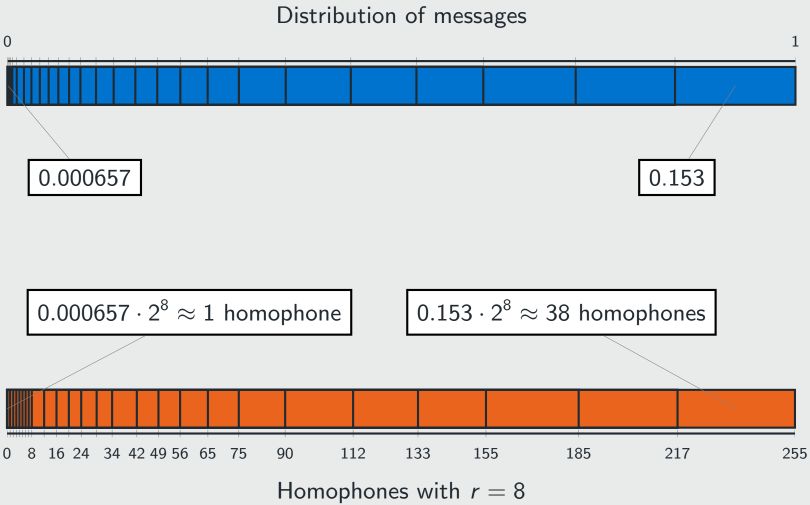
- encodings are $r$-bit strings
- assign message $m$ an interval of length $f_{\mathrm{D}}(m) \cdot 2^r$
- choose homophones uniformly at random from this set
- maintain table of assigned intervals for decoding

# IBHE example: MDC



data source: [Age09]

# IBHE example: MDC



# IBHE example: MDC

- hospital has $N = 130\,000$ records
- probability of least frequent item is $2^{-11} \approx 0.00657$
- to limit smoothness advantage to $2^{-\epsilon}$, need encoding bitlength $r \approx 17 + \epsilon$
- main problem: query expansion

# Practical security

# Experimental evaluation

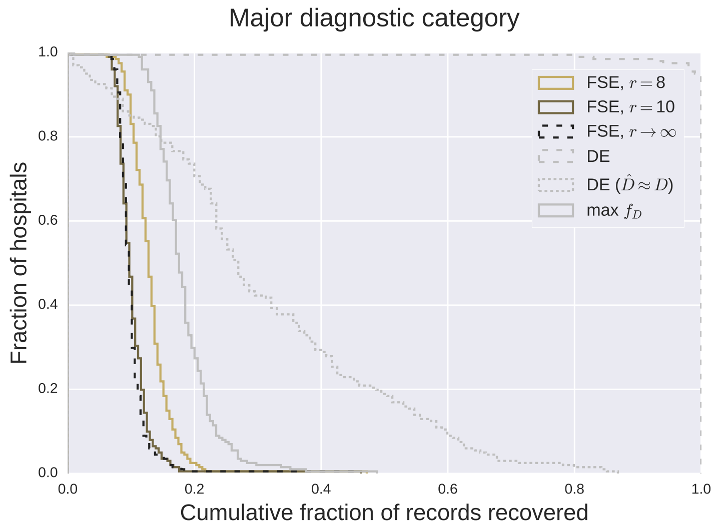
- cryptographic security levels could require unacceptably large encoding lengths
    - and hence blow-up in query expansion
- empirically evaluate smoothness:

  **how many data items can adversary correctly decrypt?**
- assume distribution D known by all
    - adversary knows how many homophones each message has
- what is optimal attack assuming only frequency information is meaningful?
    - message privacy easily achieved with a PRP

# Maximum likelihood estimation (MLE)

- apply MLE to find most likely decryption function
- **MLE applied to deterministic encryption:** decrypt most frequent ciphertext to most frequent plaintext, and so on [LP15]
- **MLE applied to FS encryption:** decrypt $|\mathcal{H}(m_1)|$ most frequent ciphertexts to most frequent plaintext $m_1$, and so on
- considers only “proper” decryption functions

# Frequency-smoothing (FS) vs. deterministic (DE) encryption



# Summary of contributions

- FS encryption thwarts snapshot inference attacks
- price to pay: query expansion, client storage
- see paper for
  - framework for *dynamic* FS schemes
  - FS construction from HE, PRF, and IV-based encryption
  - banded homophonic encoding scheme
- limited adversarial model, but part of **all** others

# Summary of contributions

- FS encryption thwarts snapshot inference attacks
- price to pay: query expansion, client storage
- see paper for
    - framework for *dynamic* FS schemes
    - FS construction from HE, PRF, and IV-based encryption
    - banded homophonic encoding scheme
- limited adversarial model, but part of **all** others

`marie-sarah.lacharite.2015@rhul.ac.uk`

# References I

- Agency for Healthcare Research and Quality, Rockville, MD.
  **HCUP Nationwide Inpatient Sample (NIS), Healthcare Cost and Utilization Project (HCUP), 2009.**
  `http://www.hcup-us.ahrq.gov/nisoverview.jsp`.
- Thomas Baignères, Pascal Junod, and Serge Vaudenay.
  **How far can we go beyond linear cryptanalysis?**
  In *Advances in Cryptology - ASIACRYPT 2004*, pages 432–450, 2004.
  `https://www.iacr.org/archive/asiacrypt2004/33290427/33290427.pdf`.

# References II

P. Grubbs, K. Sekniqi, V. Bindschaedler, M. Naveed, and T. Ristenpart.
**Leakage-abuse attacks against order-revealing encryption.**
In *IEEE Symposium on Security and Privacy (SP)*, pages 655–672, 2017.
`https://eprint.iacr.org/2016/895`.

Marie-Sarah Lacharité and Kenneth G. Paterson.
**A note on the optimality of frequency analysis vs. $\ell_p$-optimization.**
Cryptology ePrint Archive, Report 2015/1158, 2015.
`https://eprint.iacr.org/2015/1158`.

# References III

- Muhammad Naveed, Seny Kamara, and Charles V. Wright.
  **Inference attacks on property-preserving encrypted databases.**
  In *ACM CCS '15*, pages 644–655, 2015.
  `https://cs.brown.edu/~seny/pubs/edb.pdf`.

- David Pouliot and Charles V. Wright.
  **The shadow nemesis: Inference attacks on efficiently deployable, efficiently searchable encryption.**
  In *ACM CCS '16*, pages 1341–1352, 2016.
  `http://web.cecs.pdx.edu/~dpouliot/p1341-pouliot.pdf`.

# References IV

📄 Phillip Rogaway.
**Nonce-based symmetric encryption.**
In *Fast Software Encryption 2004*, pages 348–358, 2004.
`https://link.springer.com/content/pdf/10.1007/978-3-540-25937-4_22.pdf`.

📄 Phillip Rogaway and Thomas Shrimpton.
**A provable-security treatment of the key-wrap problem.**
In *Advances in Cryptology - EUROCRYPT 2006*, pages 373–390, 2006.
`https://www.iacr.org/archive/eurocrypt2006/40040377/40040377.pdf`.