F-BLEAU: Fast Black-box Leakage Estimation

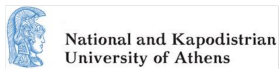
via Machine Learning

Giovanni Cherubin
SPRING lab, EPFL
@gchers

Konstantinos Chatzikokolakis
University of Athens

Catuscia Palamidessi
INRIA, École Polytechnique

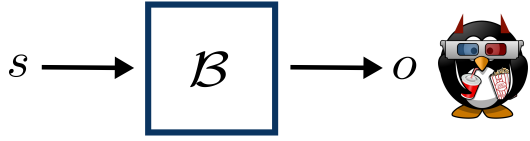
IEEE Symposium on Security&Privacy
21 May, 2019



A black-box

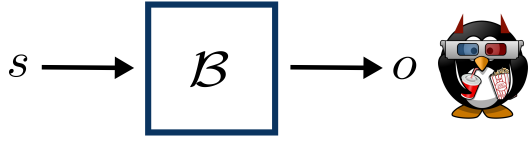
Assume no knowledge of system's internals

A black-box



Assume no knowledge of system's internals

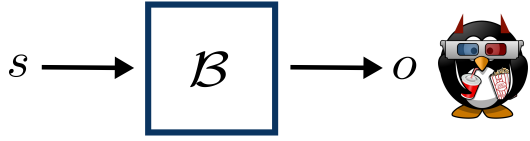
A black-box



Assume no knowledge of system's internals

How can we measure:

A black-box

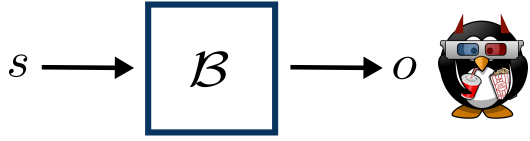


Assume no knowledge of system's internals

How can we measure:

- how much o leaks about s?

A black-box



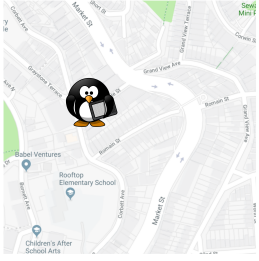
Assume no knowledge of system's internals

How can we measure:

- how much o leaks about s?
- how hard is it to predict s given o?

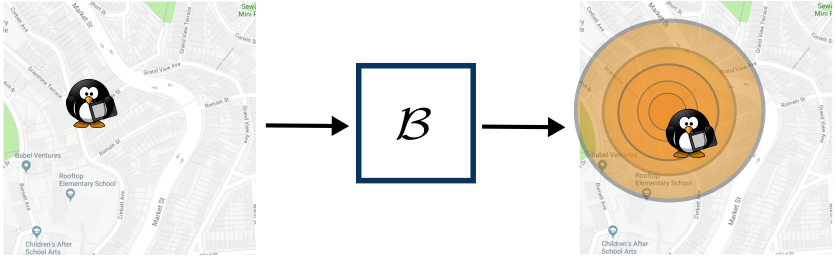
Application Examples

Location privacy



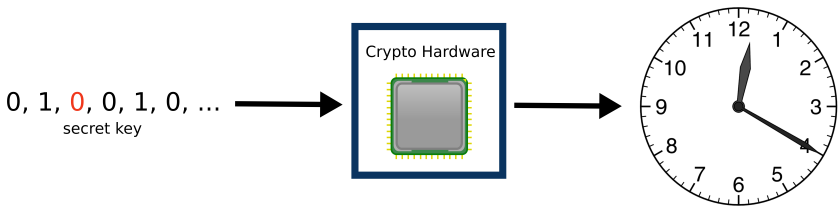
Application Examples

Location privacy



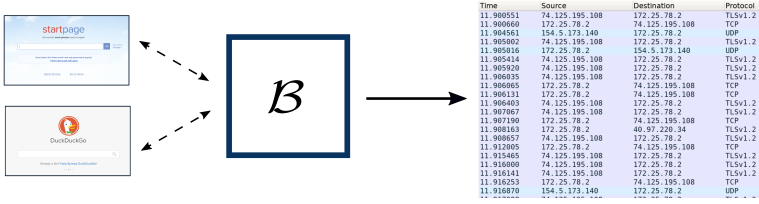
Application Examples

Side channels in crypto primitives' implementation



Application Examples

Network traffic analysis (e.g., Website Fingerprinting)



Time	Source	Destination	Protocol
11.900551	74.125.195.108	172.25.78.2	TLSv1.2
11.900660	172.25.78.2	74.125.195.108	TCP
11.904561	154.5.173.140	172.25.78.2	UDP
11.905002	74.125.195.108	172.25.78.2	TLSv1.2
11.905016	172.25.78.2	154.5.173.140	UDP
11.905414	74.125.195.108	172.25.78.2	TLSv1.2
11.905920	74.125.195.108	172.25.78.2	TLSv1.2
11.906035	74.125.195.108	172.25.78.2	TLSv1.2
11.906065	172.25.78.2	74.125.195.108	TCP
11.906131	172.25.78.2	74.125.195.108	TCP
11.906403	74.125.195.108	172.25.78.2	TLSv1.2
11.907067	74.125.195.108	172.25.78.2	TLSv1.2
11.907190	172.25.78.2	74.125.195.108	TCP
11.908163	172.25.78.2	40.97.220.34	TLSv1.2
11.908657	74.125.195.108	172.25.78.2	TLSv1.2
11.912005	172.25.78.2	74.125.195.108	TCP
11.915465	74.125.195.108	172.25.78.2	TLSv1.2
11.916000	74.125.195.108	172.25.78.2	TLSv1.2
11.916141	74.125.195.108	172.25.78.2	TLSv1.2
11.916253	172.25.78.2	74.125.195.108	TCP
11.916870	154.5.173.140	172.25.78.2	UDP

Bayes risk

$$Pr\left(\text{  }(o) \neq s\right)$$

Bayes risk

$$R^* = \min_{\text{  }} \{ Pr\left(\text{  }(o) \neq s \right) \}$$

Bayes risk

Bayes risk R^* is the probability of error of the **optimal** adversary:

$$R^* = \min_{\text{  }} \{ Pr\left(\text{  }(o) \neq s \right) \}$$

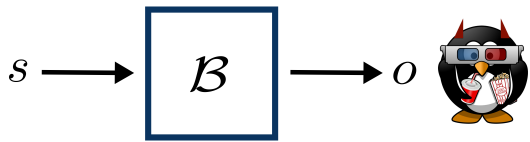
Bayes risk

Bayes risk R^* is the probability of error of the **optimal** adversary:

$$R^* = \min_{\text{  }} \{ Pr\left(\text{  }(o) \neq s \right) \}$$

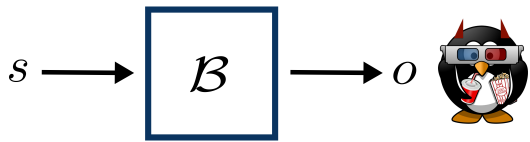
On its basis we can compute **leakage measures** (e.g., Min-entropy, multiplicative/additive leakage)

Black-box estimates



We wish to measure leakage based on queries we make to the system:

Black-box estimates



We wish to measure leakage based on queries we make to the system:

$$(s_1, o_1), (s_2, o_2), \ldots, (s_n, o_n)$$

Frequentist paradigm [C+'10]

(E.g., `leakiEst`, `LeakWatch`)

Suppose adversary observed data:

$$(s, 0.4), (s, 0.7), (s, 1.2), (s, 0.4), (s, 0.4)$$

Frequentist paradigm [C+'10]

(E.g., `leakiEst`, `LeakWatch`)

Suppose adversary observed data:

$$(s, 0.4), (s, 0.7), (s, 1.2), (s, 0.4), (s, 0.4)$$

$\text{Freq}\left(0.4\right) =$

Frequentist paradigm [C+'10]

(E.g., `leakiEst`, `LeakWatch`)

Suppose adversary observed data:

$$(\underline{s, 0.4}), (s, 0.7), (s, 1.2), (\underline{s, 0.4}), (\underline{s, 0.4})$$

$$\text{Freq}\left(0.4\right) =$$

Frequentist paradigm [C+'10]

(E.g., `leakiEst`, `LeakWatch`)

Suppose adversary observed data:

$$(s, \underline{0.4}), (s, 0.7), (s, 1.2), (s, \underline{0.4}), (s, \underline{0.4})$$

$\text{Freq}\left(0.4\right) =$ Most frequent among $\{s, s, s\}$

Frequentist paradigm [C+'10]

(E.g., `leakiEst`, `LeakWatch`)

Suppose adversary observed data:

$$(s, 0.4), (s, 0.7), (s, 1.2), (s, 0.4), (s, 0.4)$$

$\text{Freq}\left(0.4\right)$ = Most frequent among $\{s, s, s\} = s$

Frequentist paradigm [C+'10]

(E.g., `leakiEst`, `LeakWatch`)

Suppose adversary observed data:

$$(\underline{s, 0.4}), (s, 0.7), (s, 1.2), (\underline{s, 0.4}), (\underline{s, 0.4})$$

$\text{Freq}\left(0.4\right) =$ Most frequent among $\{s, s, s\} = s$

$\text{Freq}\left(0.5\right) =$

Frequentist paradigm [C+'10]

(E.g., `leakiEst`, `LeakWatch`)

Suppose adversary observed data:

$$(s, 0.4), (s, 0.7), (s, 1.2), (s, 0.4), (s, 0.4)$$

$\text{Freq}\left(0.4\right)$ = Most frequent among $\{s, s, s\} = s$

$\text{Freq}\left(0.5\right)$ = Random guessing (e.g., most frequent label overall)

Frequentist paradigm [C+'10]

(E.g., `leakiEst`, `LeakWatch`)

Suppose adversary observed data:

$$(\underline{s, 0.4}), (s, 0.7), (s, 1.2), (\underline{s, 0.4}), (\underline{s, 0.4})$$

$\text{Freq}\left(0.4\right)$ = Most frequent among $\{s, s, s\} = s$

$\text{Freq}\left(0.5\right)$ = Random guessing (e.g., most frequent label overall)

- Does not work for **continuous** output space

Frequentist paradigm [C+'10]

(E.g., `leakiEst`, `LeakWatch`)

Suppose adversary observed data:

$$(s, 0.4), (s, 0.7), (s, 1.2), (s, 0.4), (s, 0.4)$$

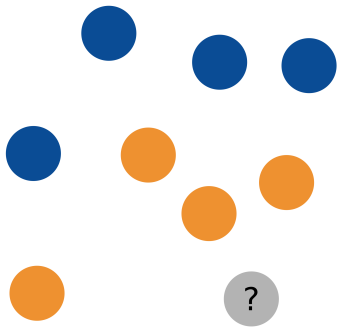
$\text{Freq}\left(0.4\right) =$ Most frequent among $\{s, s, s\} = s$

$\text{Freq}\left(0.5\right) =$ Random guessing (e.g., most frequent label overall)

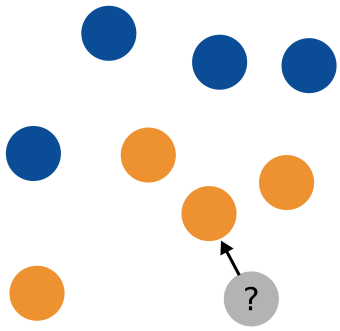
- Does not work for **continuous** output space
- Does not scale to **large** systems (needs at least one example per output value)

Nearest Neighbor

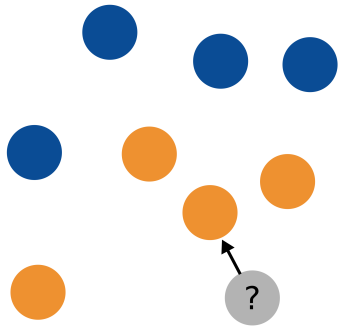
Nearest Neighbor



Nearest Neighbor

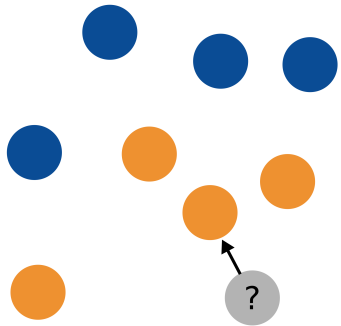


Nearest Neighbor



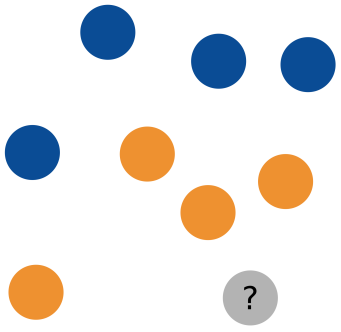
With minor modifications NN is leakage estimator

Nearest Neighbor

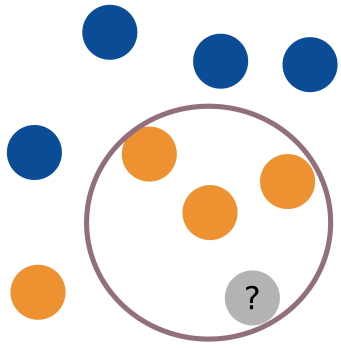


With minor modifications NN is leakage estimator (when o from finite set).

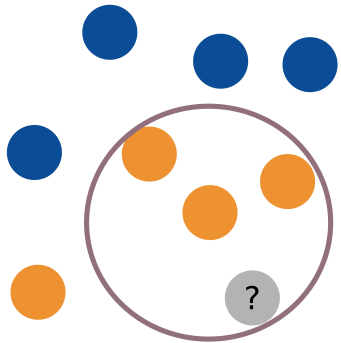
k-Nearest Neighbour [S'77]



k-Nearest Neighbour [S'77]



k-Nearest Neighbour [S'77]



If we choose $k/n \to 0$ and $k \to \infty$ as $n \to \infty$, then k-NN is leakage estimator.

Universally Consistent rules

Main result

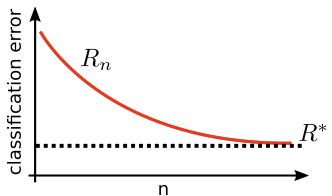
We can use Machine Learning techniques for measuring the leakage.

Universally Consistent rules

Main result

We can use Machine Learning techniques for measuring the leakage.

Definition An ML rule is universally consistent if its error R_n converges to R^* as the size of the training data $n \to \infty$

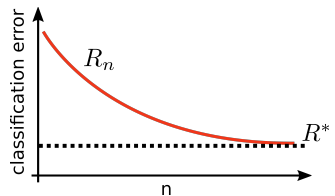


Universally Consistent rules

Main result

We can use Machine Learning techniques for measuring the leakage.

Definition An ML rule is universally consistent if its error R_n converges to R^* as the size of the training data $n \to \infty$



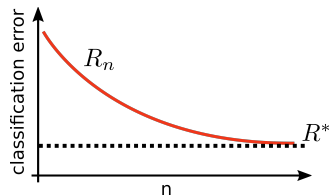
E.g., NN (finite case), k-NN, SVM and Neural Networks (some param. choices)

Universally Consistent rules

Main result

We can use Machine Learning techniques for measuring the leakage.

Definition An ML rule is universally consistent if its error R_n converges to R^* as the size of the training data $n \to \infty$



E.g., **NN** (finite case), **k-NN**, SVM and Neural Networks (some param. choices)

Results
Location privacy

Defense mechanisms:

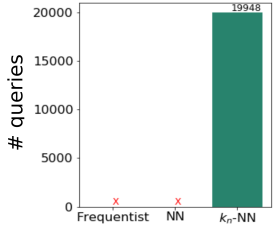
- **Geometric**
- Laplacian
- Blahut-Arimoto [O+'17]

Gowalla dataset: users' location data, $100K$ examples.

Results
Location privacy

Defense mechanisms:

- **Geometric**
- Laplacian
- Blahut-Arimoto [O+'17]



Gowalla dataset: users' location data, $100K$ examples.

Results
Location privacy

Defense mechanisms:

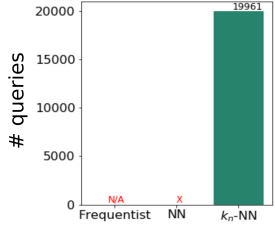
- Geometric
- **Laplacian**
- Blahut-Arimoto [O+'17]

Gowalla dataset: users' location data, $100K$ examples.

Results
Location privacy

Defense mechanisms:

- Geometric
- **Laplacian**
- Blahut-Arimoto [O+'17]

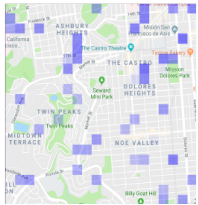


Gowalla dataset: users' location data, $100K$ examples.

Results
Location privacy

Defense mechanisms:

- Geometric
- Laplacian
- **Blahut-Arimoto [O+'17]**



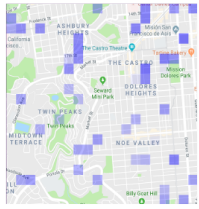
Gowalla dataset: users' location data, $100K$ examples.

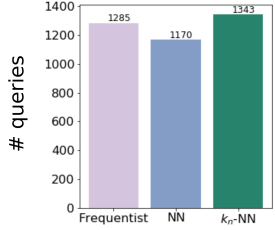
Results

Location privacy

Defense mechanisms:

- Geometric
- Laplacian
- **Blahut-Arimoto [O+'17]**





Gowalla dataset: users' location data, $100K$ examples.

NN-based estimators
Synthetic experiments

- Scale to large systems & excel when there's a metric on the output

NN-based estimators
Synthetic experiments

- Scale to large systems & excel when there's a metric on the output
- When no metric: equivalent to Frequentist

NN-based estimators
Synthetic experiments

- Scale to large systems & excel when there's a metric on the output
- When no metric: equivalent to Frequentist
- However, may converge slowly for maliciously crafted systems

Quick convergence?
No Free Lunch [W'96]

Is there an "optimal" estimator (i.e., which converges faster than all the others)?

Quick convergence?
No Free Lunch [W'96]

Is there an "optimal" estimator (i.e., which converges faster than all the others)?

No Free Lunch Theorem No. Uniformly averaged among all the possible systems, all the estimators are equivalent in terms of convergence.

Quick convergence?
No Free Lunch [W'96]

Is there an "optimal" estimator (i.e., which converges faster than all the others)?

No Free Lunch Theorem No. Uniformly averaged among all the possible systems, all the estimators are equivalent in terms of convergence.

Takeaway Always try several estimators and select the one converging faster.

Quick convergence?
No Free Lunch [W'96]

Is there an "optimal" estimator (i.e., which converges faster than all the others)?

No Free Lunch Theorem No. Uniformly averaged among all the possible systems, all the estimators are equivalent in terms of convergence.

Takeaway Always try several estimators and select the one converging faster.

`fbleau` based on this idea

TL;DL + future ideas

- Several applications: location privacy, side channels, traffic analysis, ...
- We can use ML (UC) rules: they **scale** to large problems, and tend to converge faster (or equivalently to) frequentist approach

\+ More applications (e.g., side channels, attacks to ML models)

fbleau: `https://github.com/gchers/fbleau`

Takeaway Black-box security and ML are solving similar problems: let's bridge them.

References

Applications

[CS'10] *T. Chothia, V. Smirnov*, A traceability attack against e-passports, 2010.

[BK'08] *M. Backes, B. Köpf*, Formally bounding the side-channel leakage in unknown-message attacks, 2008.

[O+'17] *S. Oya, C. Troncoso, F. Pérez-González*, Back to the drawing board: Revisiting the design of optimal location privacy-preserving mechanisms, 2017.

Black-box security

[C+'10] *K. Chatzikokolakis, T. Chothia, A. Guha*, Statistical measurement of information leakage, 2010.

[C'17] *G. Cherubin*, Bayes, not naïve: Security bounds on website fingerprinting defenses, 2017.

ML

[CH'67] *T. Cover, P. Hart*, Nearest neighbor pattern classification, 1967.

[S'77] *J. Stone*, Consistent nonparametric regression, 1977.

[W'96] *D. H. Wolpert*, The lack of a priori distinctions between learning algorithms, 1996.

F-BLEAU: Fast Black-box Leakage Estimation

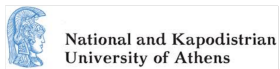
via Machine Learning

Giovanni Cherubin
SPRING lab, EPFL
@gchers

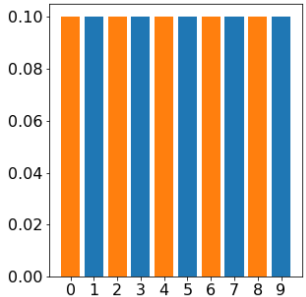
Konstantinos Chatzikokolakis
University of Athens

Catuscia Palamidessi
INRIA, École Polytechnique

IEEE Symposium on Security&Privacy
21 May, 2019

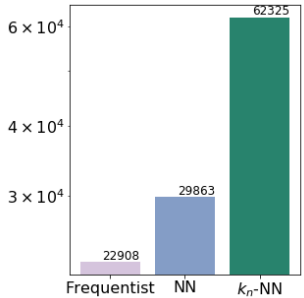


Spiky channel



2 secrets, 10K observable values

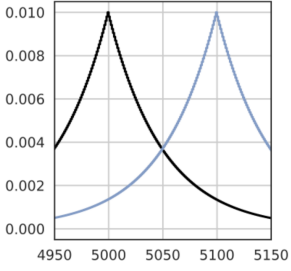
Spiky channel



2 secrets, 10K observable values

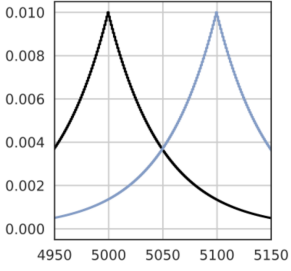
Results

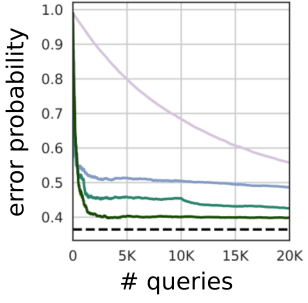
Geometric distribution

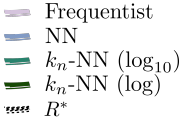


Results

Geometric distribution

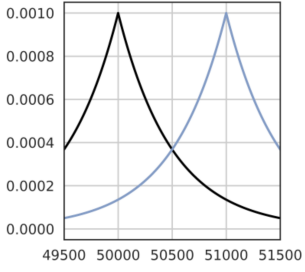


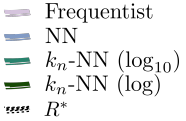




Results

Geometric distribution





Results

Geometric distribution