Leakage Certification Made Simple

Aakash Chowdhury [1] Arnab Roy [3] Carlo Brunetta [4] Elisabeth Oswald [1, 2]

[1] University of Klagenfurt; [2]University of Birmingham; [3]University of Innsbruck; [4]Independent researcher

Crypto, 2024

Certification/Evaluation

Security critical products **must undergo certification** before they can enter the market.
Side channel attacks are important and expensive part of such an evaluation:

- May have to aim for "ideal adversary" (i.e. evidence significant time and effort to estimate adversarial leakage model, evidence quality of derived model);
- May have to demonstrate attack based on classical and deep learning models.

Model building, evaluation, and attacks require multiple data sets (and maybe even multiple independent repeats), which makes attack based evaluations extremely expensive in practice.

Are there any alternatives?

"How to Certify the Leakage of a Chip?" (Durvaux et al. 2014)

Among alternatives, the 2014 paper by Durvaux et al. looked at mutual information (MI) "like" quantities to capture the "strength" of an (estimated) adversarial leakage model.

- Absolute strength: how close is an (estimated) model to the "true device leakage"?
- Relative strength: how do multiple (estimated) models compare to each other?

Estimating MI is a hard problem for high dimensional, and/or discrete-continuous mixture models.

Previous work provides solutions for evaluating relative strength for (high dimensional) **classification models**.

Making Things "Simple" (Our Work)

We provide a unified treatment of classification models and predictive models.

We provide novel MI based quantities to express both relative and absolute strength (including a useful equivalence in relation to predictive models).

We leverage a new result for high-dimensional MI estimation and demonstrate experimentally that it accurately estimates the MI for the ideal adversary.

We provide several real world case studies.

Our solution aids "simplicity" because all quantities that we define are MI based and can be efficiently estimated by the same statistical estimator.

Rest of this Talk

Using minimal terminology and formalism we explain quantities to define the "strength" of a model (aka concrete adversary) for classification models.

We briefly show some pitfalls when estimating MI (like) quantities, motivating why an approach that can be carried out with one reliable estimator makes things "simple".

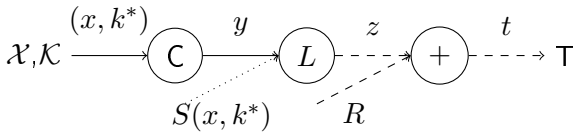
We show/compare how to use our (and previous) quantities to assess the "profiling complexity" in a concrete scenario.

The Side Channel Scenario

Adversary gets, alongside input/output $\mathcal{X}$, observations (typically multivariate traces) T, that depend on $\mathcal{X}$, some unknown instance $k^* \in \mathcal{K}$, some randomness S perhaps dependent on (x, k), and some independent randomness R:

$$T(x, k, S(x, k), R) = L(C(x, k), S) + R$$

Visually expressed as:



The MI, $I((X, K); T)$, intuitively expresses how much information there is in the traces about key K.

Expressing Model Quality (Previous Work)

Assume C is one-to-one, and T are discrete.

$$\mathrm{HI}(Y;T;\hat{L}) = H(Y) + \sum_{y\in\mathcal{Y}} p_Y(y) \cdot \sum_{t\in\mathcal{T}} p_{(Y,\hat{L})}(t|y) \log_2 p_{(Y,\hat{L})}(y|t)$$

$$\mathrm{PI}(Y;T;\hat{L}) = H(Y) + \sum_{y\in\mathcal{Y}} p_Y(y) \cdot \sum_{t\in\mathcal{T}} p_{(Y,T)}(t|y) \log_2 p_{(Y,\hat{L})}(y|t)$$

Bronchain et al. 2019; Masure et al. 2023 proposed:

- Estimate $I(Y;T)$ via the HI estimator.
- The absolute strength of a model M is given via the regret, $\mathrm{HI}(Y;T;M) - \mathrm{PI}(Y;T;M)$.
- The relative strengths of two models is given by the difference of their regrets (or PIs).

Expressing Model Quality (Our Work)

Motivated by the fact that the PI largely captures a property of the conditional distribution $Y|T$, we adopt the link to the conditional cross-entropy, see also McAllester et al. 2020:

$$I(T;Y) \geq H(Y) - H(P_{Y|T}, P_{Y|T_M})$$

Equality holds if and only if $P_{Y|T} = P_{Y|T_M}$.

We define the absolute strength of a model as

$$\delta(T, M) = I(T;Y) - \Big(H(Y) - H(P_{Y|T}, P_{Y|T_M})\Big).$$

Spot the Difference ...

Previous Work

$R(T, M) = \mathrm{HI}(Y; T; M) - \mathrm{PI}(Y; T; M)$

Require HI and PI estimators:

- Needs to discretize traces
- Issues with multivariate traces

Our Work

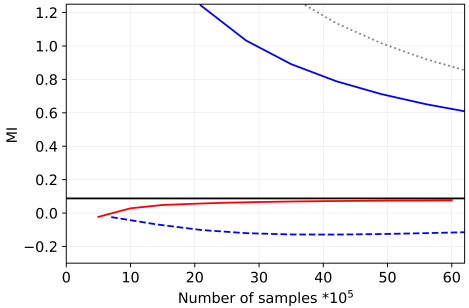
$\delta(T, M) =$

$I(T; Y) - \Big(H(Y) - H(P_{Y|T}, P_{Y|T_M})\Big)$

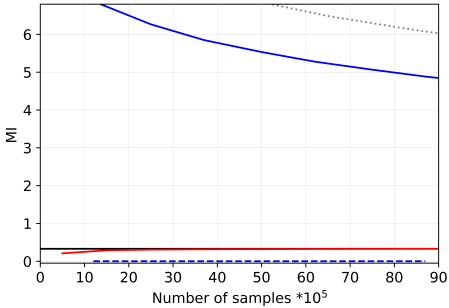
- Simplified into two MI quantities
- Use single "consistent" estimator

Using recent work of Gao et al. 2017 for MI estimation, we produce a performant estimator implementation (GKOV) for the side channel use case (public repo. is referenced in the paper).

The Challenge of Multivariate MI Estimation



(a) $L : (HW, HD)$, $R \sim \mathcal{N}(0, 4)$



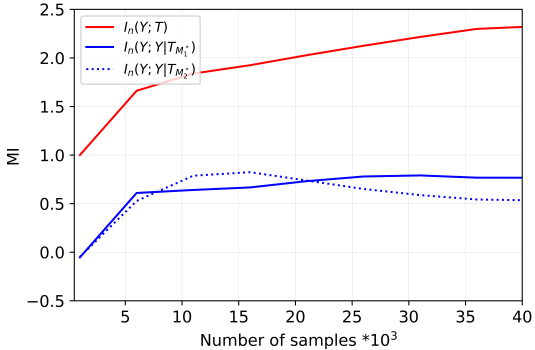
(b) $L : (HW, HW, HD)$, $R \sim \mathcal{N}(0, 2)$

I_n^{hist}: ····· , eHI: — , $I((X, K); T)$: — , I_n^{GKOV}: — , ePI: - - -

The GKOV estimator is superior in comparison to previous estimators (black line is ground truth, red is GKOV).

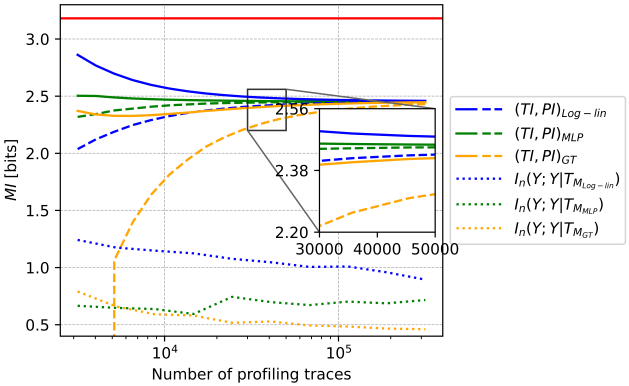
Practical Use Case: How good are given classification models in absolute terms?

Data set: Xilinx Virtex-5 FPGA implementation of unprotected AES-128, 30 points are selected using GKOV. We train deep nets (M_1^*, M_2^*) so that M_1^* is the better model per design.



MI estimates show that both models are far from optimal.

Practical Use Case: How many traces are enough for training?



- All quantities "rank" the classification models consistently.
- The non-MI quantities cannot be "directly compared" with the true MI (red line).
- The MI quantities demonstrate how far the models are from being optimal.

Limitations and Future Directions

Our work is the first to clearly differentiate between theoretical quantities to judge model quality and their estimation: we need more research towards what is the “right quantity”.

GKOV enables multivariate MI estimation, but in practice it is also limited because of the knn search. Better implementations would be helpful.

Can we integrate the quantities directly into training (e.g. could a loss function directly take advantage of some simultaneous MI estimation)?

Thank You
For Your Attention

This project is supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 725042), the Austrian Science Fund (FWF) 10.55776/F85 (SFB SpyCode), and by the EU Horizon project (enCrypton, grant agreement no. 101079319)

References I

[1] Olivier Bronchain, Julien M Hendrickx, ClĂŠment Massart, Alex Olshevsky, and FranĂ§ois-Xavier Standaert. "Leakage certification revisited: Bounding model errors in side-channel security evaluations". In: *Advances in Cryptology–CRYPTO 2019: 39th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 18–22, 2019, Proceedings, Part I 39*. Springer. 2019, pp. 713–737.

[2] FranĂ§ois Durvaux, FranĂ§ois-Xavier Standaert, and Nicolas Veyrat-Charvillon. "How to certify the leakage of a chip?" In: *Advances in Cryptology–EUROCRYPT 2014: 33rd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Copenhagen, Denmark, May 11-15, 2014. Proceedings 33*. Springer. 2014, pp. 459–476.

[3] Weihao Gao, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. "Estimating mutual information for discrete-continuous mixtures". In: *Advances in neural information processing systems* 30 (2017).

References II

[4] LoÄąĚc Masure, GaĂŤtan Cassiers, Julien Hendrickx, and FranĂ§ois-Xavier Standaert. "Information bounds and convergence rates for side-channel security evaluators". In: *IACR Transactions on Cryptographic Hardware and Embedded Systems* 2023.3 (2023), pp. 522–569.

[5] David McAllester and Karl Stratos. "Formal limitations on the measurement of mutual information". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 875–884.