Aalto University

UNIVERSITY OF WATERLOO

Conflicts Between ML Security/Privacy Techniques

Sebastian Szyller, N. Asokan

https://sebszyller.com https://asokan.org/asokan/

@sebszyller @nasokan

Model theft is an important concern

Machine learning models: business advantage and intellectual property (IP)

Cost of

- gathering relevant data
- labeling data
- expertise required to choose the right model training method
- resources expended in training

Adversary who steals the model can avoid these costs.

Defending against model theft

We can try to:

- prevent (or slow down) model extraction, or
- detect it

Or deter the attacker by providing the means for ownership demonstration:

- model watermarking
- data watermarking
- fingerprinting

Other ML security & privacy concerns

There are considerations other than model ownership:

- model evasion (defense: adversarial training)
- training data reconstruction (defense: differential privacy)
- membership inference (defense: regularization, early stopping)
- model poisoning (defense: regularization, outlier/anomaly detection)
- …

How does ownership demonstration interact with the other defenses?

We investigate pairwise interactions of:

model watermarking

data watermarking

fingerprinting

WITH

differential privacy

adversarial training

Setup & Baselines

We use the following techniques (and corresponding metrics):

- Out-of-distribution (OOD) backdoor watermarking (test and watermark accuracy)
- Radioactive data (test accuracy and loss difference)
- Dataset Inference (verification confidence)
- DP-SGD (model accuracy for the given epsilon)
- Adversarial training with PGD (test and adv. accuracy for the given epsilon)

Dataset	No defense	Watermarking		Radioactive Data		Dataset Inference	DP-SGD (eps=3)	ADV. TR.	
	TEST	TEST	WM	TEST	Loss. Diff.	Confidence	TEST	TEST	ADV.
MNIST	*0.99*	*0.99*	*0.97*	*0.98*	*0.284*	*<e-30*	*0.98*	*0.99*	*0.95*
FMNIST	*0.91*	*0.87*	*0.99*	*0.88*	*0.19*	*<e-30*	*0.86*	*0.87*	*0.69*
CIFAR10	*0.92*	*0.82*	*0.97*	*0.85*	*0.2*	*<e-30*	*0.38*	*0.82*	*0.82*

Interaction with differential privacy

Differential privacy is a strong per-sample regulariser:

- Watermarking rendered ineffective
- Lower but still sufficient confidence for radioactive data
- No effect on the DI fingerprint

Dataset	DP-SGD (eps=3)
	TEST
MNIST	*0.98*
FMNIST	*0.86*
CIFAR10	*0.38*

Dataset	No defense	Watermarking				Radioactive Data				Dataset Inference	
		Baseline		with DP		Baseline		with DP		Baseline	with DP
	TEST.	TEST	WM	TEST	WM	TEST	Loss. Diff.	TEST	Loss. Diff.	Conf.	Conf.
MNIST	*0.99*	*0.99*	*0.97*	*0.97*	*0.30*	*0.98*	*0.284*	*0.97*	*0.091*	*<e-30*	*<e-30*
FMNIST	*0.91*	*0.87*	*0.99*	*0.86*	*0.28*	*0.85*	*0.19*	*0.84*	*0.11*	*<e-30*	*<e-30*
CIFAR10	*0.92*	*0.82*	*0.97*	*0.38*	*0.12*	*0.85*	*0.2*	*0.35*	*0.19*	*<e-30*	*<e-30*

Interaction with DP (tweaks and relaxations)

Tweaking DP-SGD:

- Naively increasing eps (less noise) does not improve WM accuracy
- Increasing gradient clipping threshold is better (not sufficient)

Tweaking the watermark:

- Bigger trigger set gives better WM accuracy (not sufficient)
- Training longer is better (not sufficient)

With strict DP-SGD, OOD backdoor watermarking does not work.

What if we relax DP-SGD?

- Splitting the training into the DP part (genuine data) and non-DP (watermark) helps
- Watermark is embedded successfully (accuracy > 0.9)
- Privacy loss analysis is not tight anymore

Interaction with adversarial training

Dataset	ADV. TR.	
	TEST	ADV.
MNIST	*0.99*	*0.95*
FMNIST	*0.87*	*0.69*
CIFAR10	*0.82*	*0.82*

Adversarial training creates a robust L_p bubble:

- Watermarking not affected but adversarial accuracy drops
- Significant drop in the confidence of radioactive data
- No effect on the DI fingerprint

Dataset	No defense	Watermarking					Radioactive Data					DI	
		Baseline		with ADV. TR.			Baseline		with ADV. TR.			Baseline	with ADV. TR.
	TEST	TEST	WM	TEST	WM	ADV	TEST	Loss. Diff.	TEST	Loss. Diff.	ADV	Conf.	Conf.
MNIST	*0.99*	*0.99*	*0.97*	*0.97*	*0.99*	*0.88*	*0.98*	*0.284*	*0.97*	*0.001*	*0.95*	*<e-30*	*<e-30*
FMNIST	*0.91*	*0.87*	*0.99*	*0.86*	*0.99*	*0.51*	*0.85*	*0.19*	*0.84*	*0.0007*	*0.69*	*<e-30*	*<e-30*
CIFAR10	*0.92*	*0.82*	*0.97*	*0.78*	*0.97*	*0.65*	*0.85*	*0.2*	*0.81*	*0.003*	*0.81*	*<e-30*	*<e-30*

False positives in Dataset Inference 1/2

We noticed false positives when DI is combined with other defenses:

- models would trigger confident FPs w.r.t. unrelated models (e.g. MNIST to FMNIST)
- But we saw FPs even in our DI baseline (i.e., without other defenses)

We revisited the original[1] DI itself (CIFAR10):

- use the implementation from the official repo[2]
- Models provided in the repo work as intended
- We trained many independent models:
  - Without any other defense
  - We can reproduce the results from the paper, however...

[1] - Dataset Inference: Ownership Resolution in Machine Learning
[2] - Dataset Inference, GitHub repository

False positives in Dataset Inference 2/2

We revisited the original[1] DI itself (CIFAR10):

- **The original split for CIFAR10 uses:**
  - the training set for the teacher model
  - the test set to train the independent model
  - the test set and the training set are used for the distinguisher (double-dip on the test set)

- **We split CIFAR10 training set into two non-overlapping chunks (A and B):**
  - one for the teacher (A), one for the independent model (B)
  - the test and the A set are used for the distinguisher
  - independent model B triggers a FP with high confidence

Model trained on:	Verification p-value
A (teacher)	*e-23*
Test (original)	*0.1*
B (independent)	*e-12*
A+B	*e-13*

Is dataset-based fingerprinting feasible?

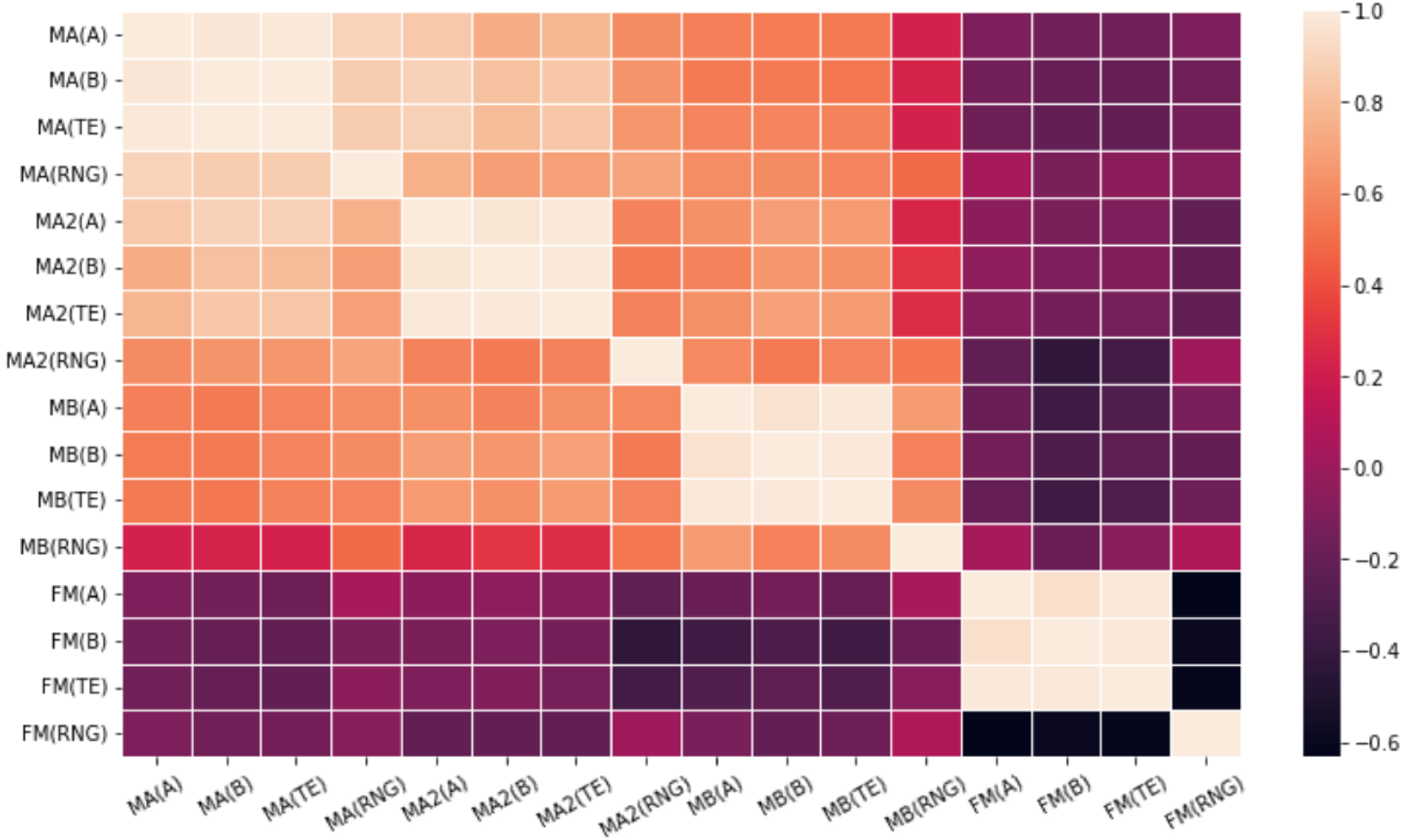
Yes, if model output has enough entropy to distinguish among instances of:

1. same model architecture trained on the same data
2. same model architecture trained on different data from the same distribution
3. other architectures/data distributions

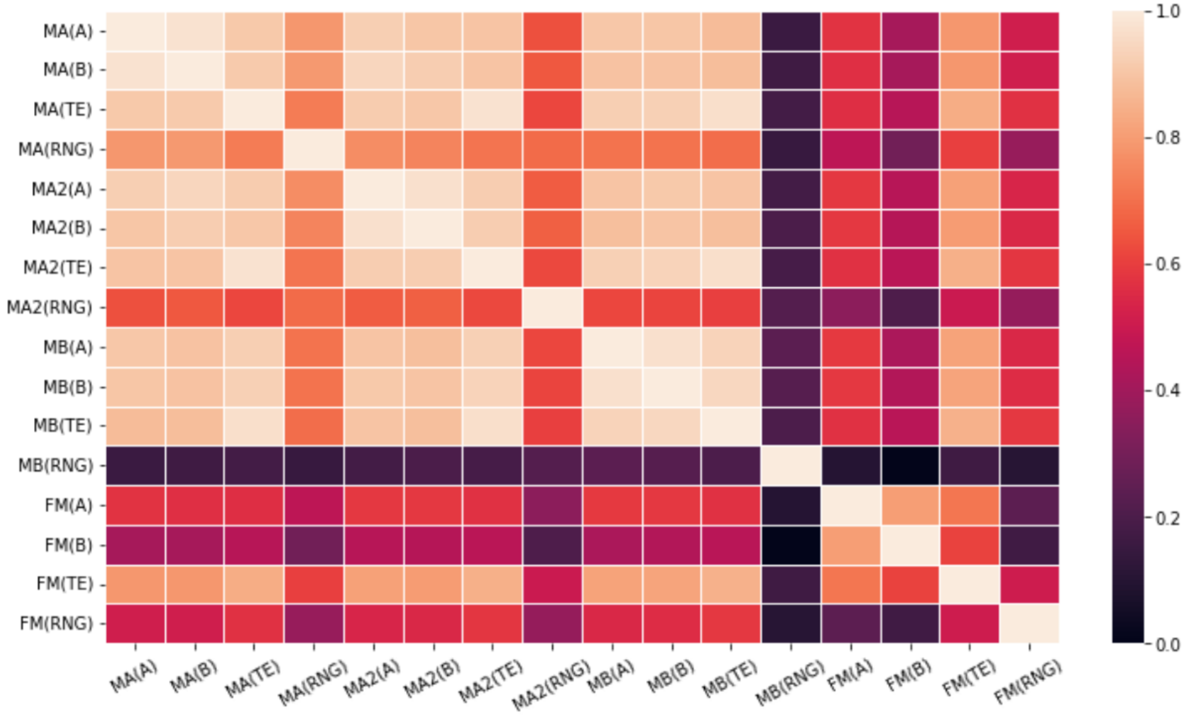
Preliminary experiment – (cumulative) distance between two models' outputs:

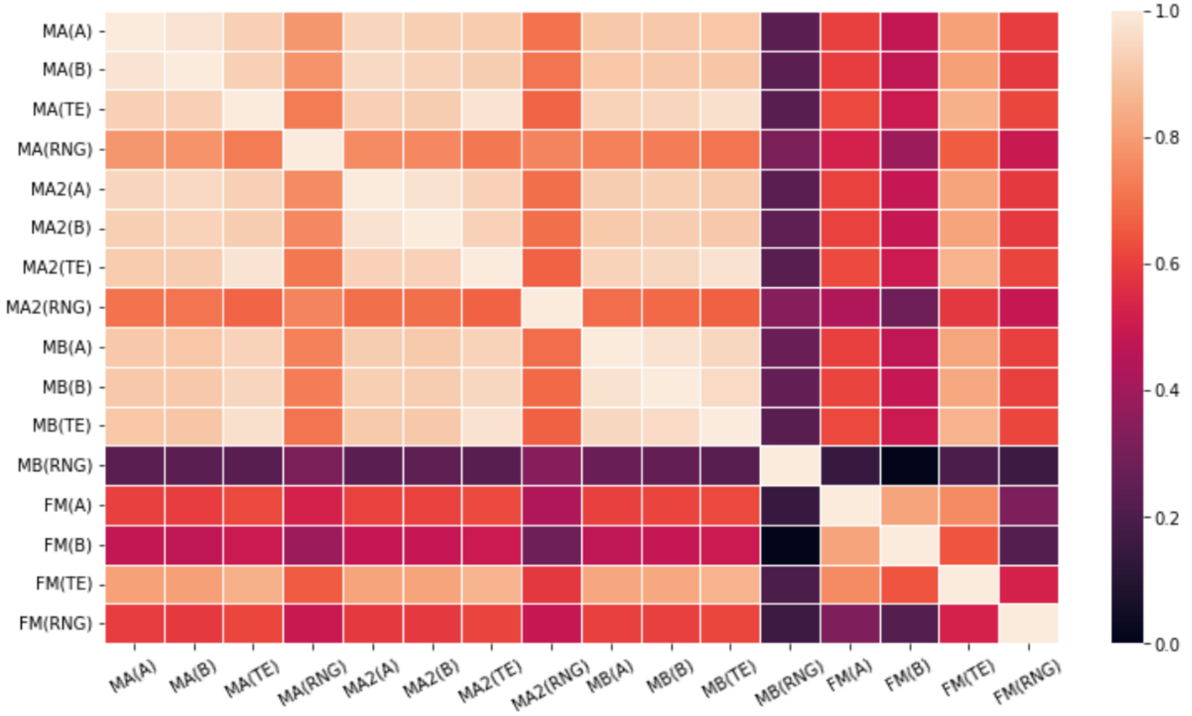
- three models trained on MNIST chunks A and B
  - MA and MA2 trained on the chunk A (type 1)
  - MB trained on the chunk B (type 2)
- a model trained on the full FMNIST (FM) (type 3)
- record outputs of all models for both chunks, the MNIST test set (TE) and random data (RNG)
  - notation example: output on A of a model trained using B – MB(A)

Distinguishing models: cumulative cosine similarity



Distinguishing models: L_1 & L_2 distance*





* Actually $(1 - L_P)$ to be visually consistent with cosine similarity.

Interaction between ML security/privacy techniques

Property	Adversarial Training	Differential Privacy	Membership Inference	Oblivious Training	Model/Gradient Inversion	Model Poisoning	Model Watermarking	Model Fingerprinting	Data Watermarking	Explainability	Fairness
Adversarial Training	X	[5]	[9]	?	?	[7]	OURS	OURS	OURS	[11]	?
Differential Privacy		X	[3, 6]	?	?	?	OURS	OURS	OURS	?	[1, 2, 8]
Membership Inference			X	?	?	[10]	?	?	?	?	?
Oblivious Training				X	?	?	?	?	?	?	?
Model/Gradient Inversion					X	?	?	?	?	?	?
Model Poisoning						X	?	?	?	?	?
Model Watermarking							X	?	?	?	?
Model Fingerprinting								X	?	[4]	?
Data Watermarking									X	?	?
Fairness										X	?
Explainability											X

REFERENCES

[1] Hongyan Chang and Reza Shokri. 2021. On the Privacy Risks of Algorithmic Fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS P)*. 292–303. https://doi.org/10.1109/EuroSP51992.2021.00028

[2] Victoria Cheng, Vinith M. Suriyakumar, Natalie Dullerud, Shalmali Joshi, and Marzyeh Ghassemi. 2021. Can You Fake It Until You Make It? Impacts of Differentially Private Synthetic Data on Downstream Classification Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 149–160. https://doi.org/10.1145/3442188.3445879

[3] Thomas Humphries, Simon Oya, Lindsey Tulloch, Matthew Rafuse, Ian Goldberg, Urs Hengartner, and Florian Kerschbaum. 2020. Investigating Membership Inference Attacks under Data Dependencies. https://doi.org/10.48550/ARXIV.2010.12112

[4] Hengrui Jia, Hongyu Chen, Jonas Guan, Ali Shahin Shamsabadi, and Nicolas Papernot. 2022. A Zest of LIME: Towards Architecture-Independent Model Distances. In *International Conference on Learning Representations*. https://openreview.net/forum?id=OUz_9TiTv9j

[5] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. Certified Robustness to Adversarial Examples with Differential Privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*. 656–672. https://doi.org/10.1109/SP.2019.00044

[6] Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlin. 2021. Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning. In *2021 IEEE Symposium on Security and Privacy (SP)*. 866–882. https://doi.org/10.1109/SP40001.2021.00069

[7] Ren Pang, Hua Shen, Xinyang Zhang, Shouling Ji, Yevgeniy Vorobeychik, Xiapu Luo, Alex Liu, and Ting Wang. 2020. *A Tale of Evil Twins: Adversarial Inputs versus Poisoned Models.* Association for Computing Machinery, New York, NY, USA, 85–99. https://doi.org/10.1145/3372297.3417253

[8] Adam Pearce. 2022. Can a Model Be Differentially Private and Fair? https://pair.withgoogle.com/explorables/private-and-fair/. Online; accessed 7 April 2022.

[9] Liwei Song, Reza Shokri, and Prateek Mittal. 2019. Privacy Risks of Securing Machine Learning Models against Adversarial Examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19)*. Association for Computing Machinery, 241–257. https://doi.org/10.1145/3319535.3354211

[10] Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. 2022. Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets. https://doi.org/10.48550/ARXIV.2204.00032

[11] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness May Be at Odds with Accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. https://openreview.net/forum?id=SyxAb30cY7

Conclusion and next steps

In combination with other defenses, ownership verification is brittle:

- Strong regularizers patch weaknesses that WM/Radioactive data exploit
- Difficult to predict the interaction of a given pair of defenses

Thorough exploration vs. combinatorial explosion:

- We present just three pairs but there are more combinations
- What about triplets, quadruplets...?
- Within-type variation also a problem, e.g.
  - We focused on the most popular DP-SGD
  - SCATTER-DP or PATE behave differently

More on our security + ML research at https://ssg.aalto.fi/research/projects/mlsec/model-extraction/

This work: Conflicting Interactions Among Protection Mechanisms for Machine Learning Models