SHAPr

An Efficient and Versatile Membership Privacy Risk Metric for Machine Learning

***Vasisht Duddu*, Sebastian Szyller, N. Asokan**

vasisht.duddu@uwaterloo.ca, contact@sebszyller.com, asokan@acm.org

https://crysp.uwaterloo.ca/research/SSG/

Why measure membership privacy risk?

Regulatory requirements for privacy risk assessment

Membership inference attacks (MIAs) risk leaking sensitive data

Need a metric to estimate the likelihood of MIAs' success

Measuring membership privacy risk: desiderata

“Principled”

independent of specific MIAs (“future-proof”)

Fine-grained

measure risk of individual training data records

Effective

assess susceptibility to MIAs

Efficient

reasonable computational overhead

Measuring membership privacy risk: State of the art

	Independent	Fine-grained	Effective	Efficient
MLPrivacyMeter[1] MLDoctor [2]	✗	✗	✓	✓
Song et al. [3]	✗	✓	✓	✓
Long et al. [4]		✓	✓	✗

[1] Murakonda et al. *ML Privacy Meter: Aiding Regulatory Compliance by Quantifying the Privacy Risks of Machine Learning*. HotPETs 2020.
[2] Liu et al. *ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models.* USENIX 2022.
[3] Song et al. *Systematic Evaluation of Privacy Risks in Machine Learning*. USENIX 2021.
[4] Long et al. *Towards Measuring Membership Privacy*. ArXiv 2017.
[5] Feldman. *Does Learning Require Memorization? A Short Tale about a Long Tail*. STOC 2020.

SHAPr: a new metric for membership privacy

Shapley Values

- Game-theoretic approach[1] to equitably assign utility among different players
- Proposed[2,3] for economic data valuation in data marketplaces
- Based on the leave-one-out approach

$$\boldsymbol{\phi} = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{S} \subseteq \boldsymbol{D} \setminus z} [\boldsymbol{U}(\boldsymbol{S} \cup \boldsymbol{z}) - \boldsymbol{U}(\boldsymbol{S})] \frac{\mathbf{1}}{\binom{|\boldsymbol{D}|-\mathbf{1}}{|\boldsymbol{S}|}}$$

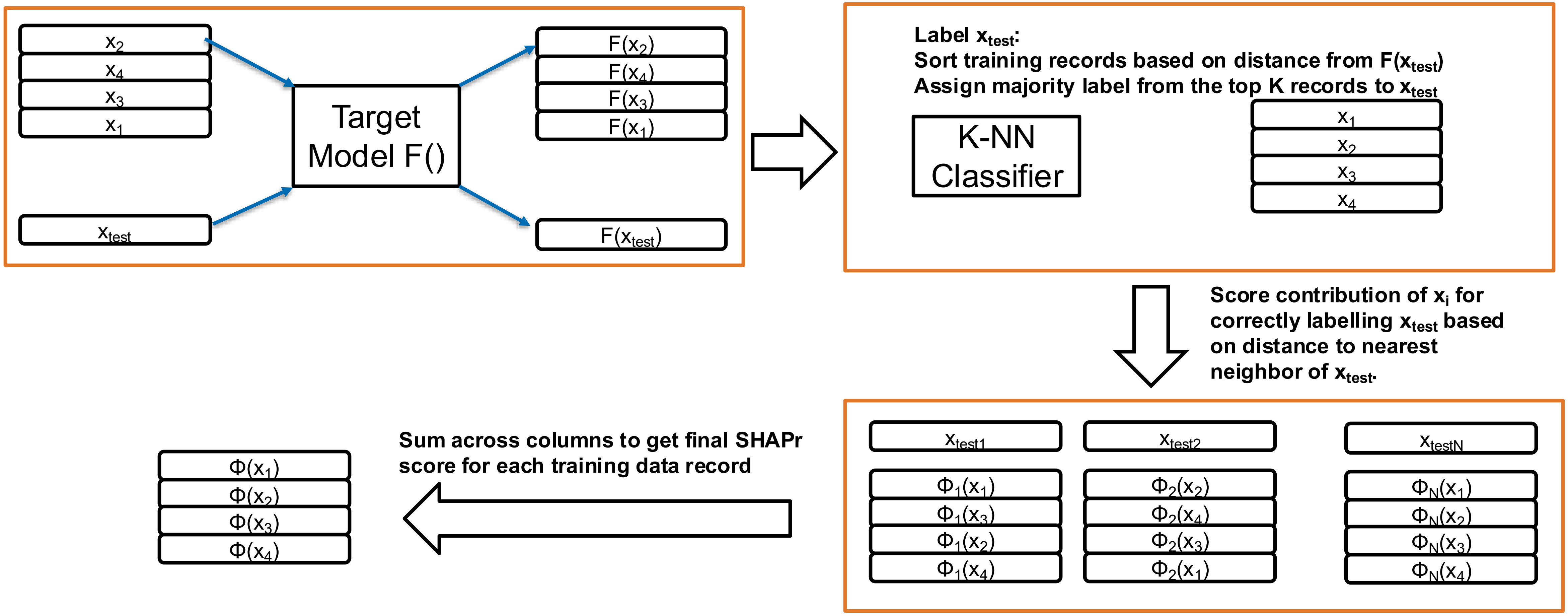
Average over different subsets of training data

Marginal contribution of z

- Independent, fine-grained, effective, but not efficient?
- Once computed, useful for other applications, e.g. data valuation ("versatile")

[1] Shapley. *A Value of n-person Games*. Contribution to the Theory of Games 1953.
[2] Jia et al. *Efficient Task-Specific Data Valuation for Nearest Neighbour Algorithms*. VLDB 2019.
[3] Jia et al. *Scalability vs. Utility: Do We Have to Sacrifice One for the Other in Data Importance Quantification*? CVPR 2021.

Efficiently computing Shapley values via K-NN



[1] Jia et al. *Efficient Task-Specific Data Valuation for Nearest Neighbor Algorithms*. VLDB 2019.
[2] Jia et al. *Scalability vs. Utility: Do We Have to Sacrifice One for the Other in Data Importance Quantification*? CVPR 2021.

Effectiveness: Susceptibility to MIAs

Ground truth: Success of Modified Entropy MIA[1]

Baseline: Song et al's[1] "privacy risk scores" (SPRS)

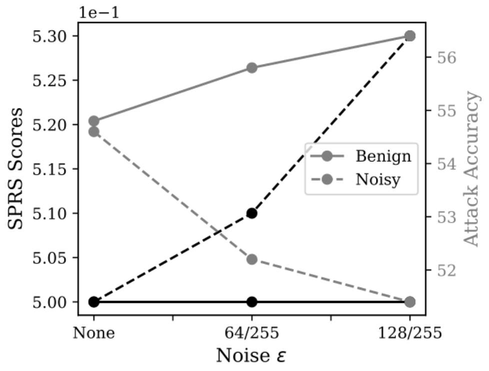
SHAPr and **SPRS** have comparable effectiveness

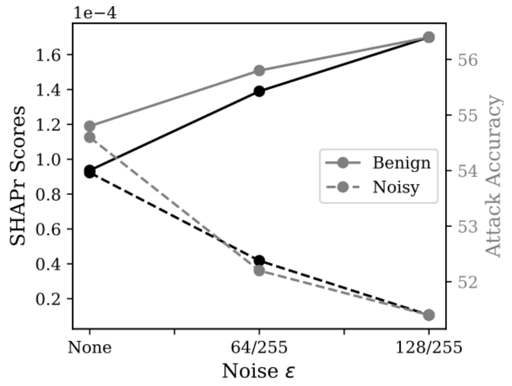
Dataset	Metric	Precision	p-value	Recall	p-value
SPRS Datasets					
LOCATION	SPRS	0.96 ± 1e-16	>0.05	0.93 ± 1e-16	<0.01
	SHAPr	0.96 ± 0.000		0.85 ± 0.000	
PURCHASE	SPRS	0.95 ± 1e-16	>0.05	0.80 ± 0.000	<0.01
	SHAPr	0.95 ± 1e-16		0.81 ± 0.000	
TEXAS	SPRS	0.92 ± 1e-16	<0.01	0.95 ± 0.000	<0.01
	SHAPr	0.96 ± 1e-16		0.74 ± 1e-16	
Additional Datasets					
MNIST	SPRS	0.99 ± 0.002	<0.01	0.57 ± 0.013	<0.01
	SHAPr	0.99 ± 8e-4		0.94 ± 0.001	
FMNIST	SPRS	0.99 ± 0.005	0.05	0.98 ± 0.026	<0.01
	SHAPr	0.99 ± 0.005		0.89 ± 0.026	
USPS	SPRS	0.79 ± 0.201	0.84	0.76 ± 0.074	<0.01
	SHAPr	0.77± 0.230		0.98 ± 0.009	
FLOWER	SPRS	0.98 ± 0.010	0.81	0.81 ± 0.040	<0.01
	SHAPr	0.98 ± 0.010		0.94 ± 0.008	
MEPS	SPRS	0.96 ± 1e-16	<0.01	0.99 ± 0.000	<0.01
	SHAPr	0.97 ± 1e-16		0.91 ± 1e-16	
CREDIT	SPRS	0.94 ± 0.006	<0.01	0.81 ± 2e-4	<0.01
	SHAPr	0.89 ± 0.004		0.92 ± 0.002	
CENSUS	SPRS	0.98 ± 0.000	<0.05	1.00 ± 0.000	<0.05
	SHAPr	0.93 ± 0.000		0.84 ± 0.000	

[1] Song et al. *Systematic Evaluation of Privacy Risks in Machine Learning*. USENIX 2021.

Effectiveness: Effect of Noise Addition

Ground truth: With added noise, MIA accuracy decreases for noisy data but increases for the rest



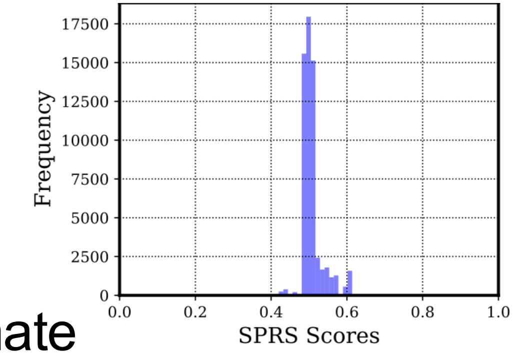


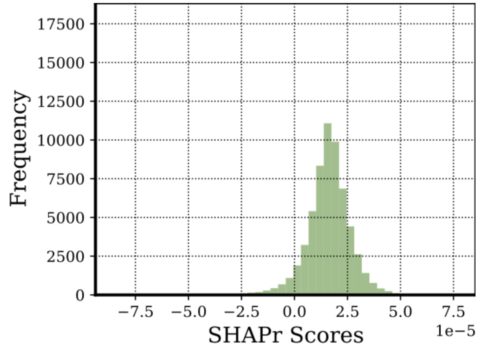
SHAPr mirrors the MIA accuracy trend

SPRS does not

Comparing Distributions:

- Different records have difference influence on model performance →variable privacy risks
- Majority SPRS scores ~0.5 → inconclusive risk estimate





“Principled”: Is SPRS future proof?

Simulated “future”: Modified Entropy MIA[1]

baseline from 

Simulated “past”: Original Entropy MIA

Recall drops drastically in the simulated “past”

SPRS likely ineffective in assessing risk of future MIAs

Dataset	Metric	Precision	Recall
SPRS Datasets			
LOCATION	Baseline	0.96 ± 1e-16	0.93 ± 1e-16
	Simulated	0.95 ± 1e-16	0.97 ± 1e-16
PURCHASE	Baseline	0.95 ± 1e-16	0.80 ± 0.000
	Simulated	0.99 ± 1e-16	0.50 ± 1e-16
TEXAS	Baseline	0.92 ± 1e-16	0.95 ± 0.000
	Simulated	0.94 ± 6e-4	0.79 ± 0.002
Additional Datasets			
MNIST	Baseline	0.99 ± 0.002	0.57 ± 0.013
	Simulated	0.99 ± 0.001	0.56 ± 0.028
FMNIST	Baseline	0.99 ± 0.005	0.98 ± 0.026
	Simulated	1.0 ± 0.000	0.64 ± 0.035
USPS	Baseline	0.79 ± 0.201	0.76 ± 0.074
	Simulated	0.86 ± 0.160	0.64 ± 0.050
FLOWER	Baseline	0.98 ± 0.010	0.81 ± 0.040
	Simulated	0.99 ± 0.006	0.66 ± 0.094
MEPS	Baseline	0.96 ± 1e-16	0.99 ± 0.000
	Simulated	0.94 ± 0.001	0.67 ± 6e-4
CREDIT	Baseline	0.94 ± 0.006	0.81 ± 2e-4
	Simulated	0.79 ± 0.032	0.39 ± 0.038
CENSUS	Baseline	0.98 ± 0.000	1.00 ± 0.000
	Simulated	0.99 ± 1e-16	0.28 ± 0.000

[1] Song et al. *Systematic Evaluation of Privacy Risks in Machine Learning*. USENIX 2021.

Efficiency: Computational Overhead

Execution time: ~2 mins to ~90 mins (one-time cost)

100x faster than naïve leave-one-out approach

Dataset	# Records	# Features	Execution Time (s)
SPRS Datasets			
LOCATION	1000	446	130.77 ± 3.90
PURCHASE	19732	600	3065.58 ± 19.24
TEXAS	10000	6170	5506.79 ± 17.47
Additional Datasets			
MNIST	60000	784	2747.41 ± 22.65
FMNIST	60000	784	3425.90 ± 34.03
USPS	3000	256	238.67 ± 1.74
FLOWER	1500	2048	174.27 ± 11.74
MEPS	7500	42	732.43 ± 4.95
CREDIT	15000	24	1852.66 ± 30.92
CENSUS	24000	103	3718.26 ± 18.25

Versatility

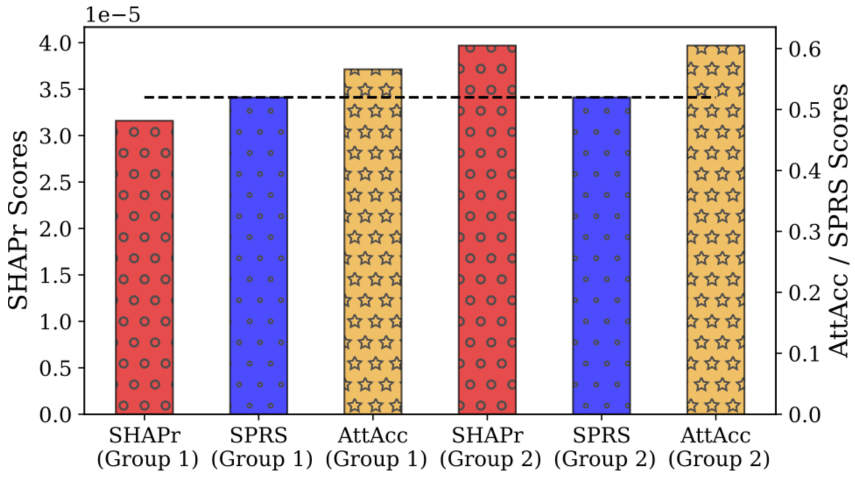
Data Valuation

- SHAPr inherits applicability to data valuation
- Other metrics without heterogeneity and additivity properties likely not applicable for data valuation

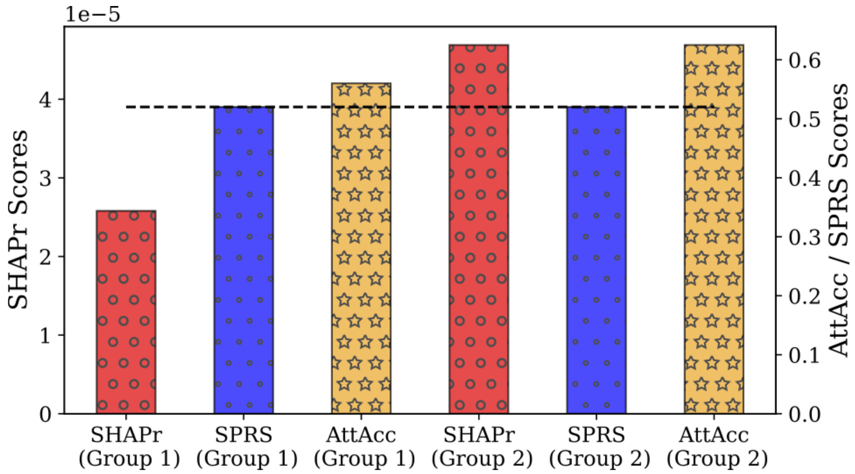
Fairness

- Different subgroups have different privacy risk

- SHAPr scores reflect trend in ground truth
  - Additivity property allows aggregation over subgroups



Race



Gender

[1] Kulynych et al. *Disparate Vulnerability to Membership Inference Attacks*. PETS 2022.

Pitfalls of Data Removal

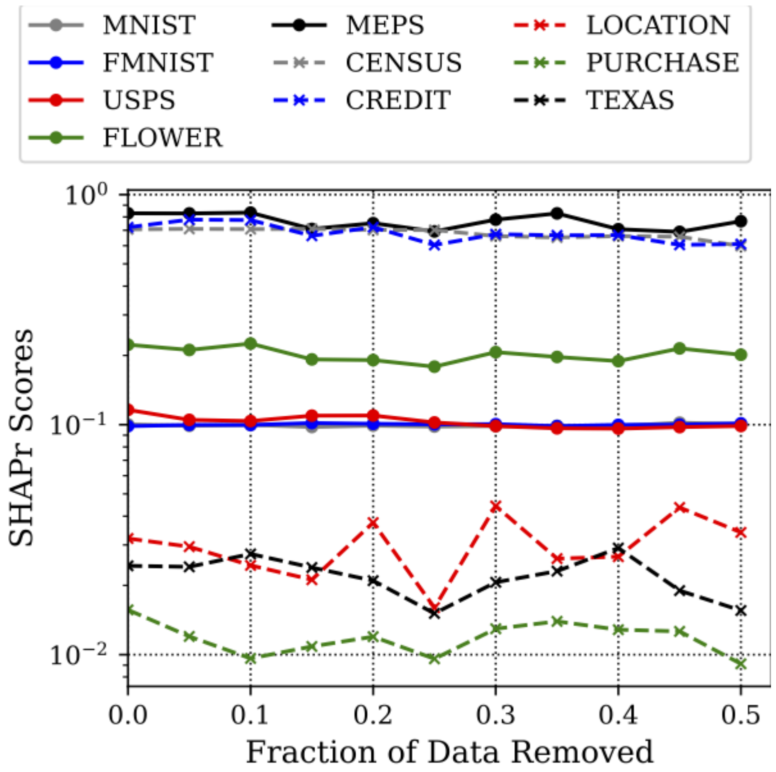
No consistent trend for SHAPr scores

- Influence of other records varies, resulting in fluctuating privacy risk scores

Removing high risk records does not improve privacy

We confirm Long et al.'s[1] observation, and have

- more datasets (10 vs. 1)
- more extensive removal of data records (50% vs 2%)



[1] Long et al. *Towards Measuring Membership Privacy*. ArXiv 2017.

Summary

SHAPr lets model builders assess membership privacy risks of individual data records

SHAPr is:

- Independent of specific MIAs
- Effective in assessing susceptibility to MIAs
- Efficient in terms of computational overhead
- Versatile (other applications like fairness, data valuation)

arXiv:2112.02230
Under review.