Demystifying Membership Inference Attacks in Machine Learning as a Service

Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei

Outline

- General formulation of black-box membership inference attack

  -> similar to the Shokri MIA approach

- Empirical experiments about what makes a model vulnerable against MIA

  -> two main factors: dataset and target model

1. General formulation of black-box membership inference attack

Workflow of a MIA

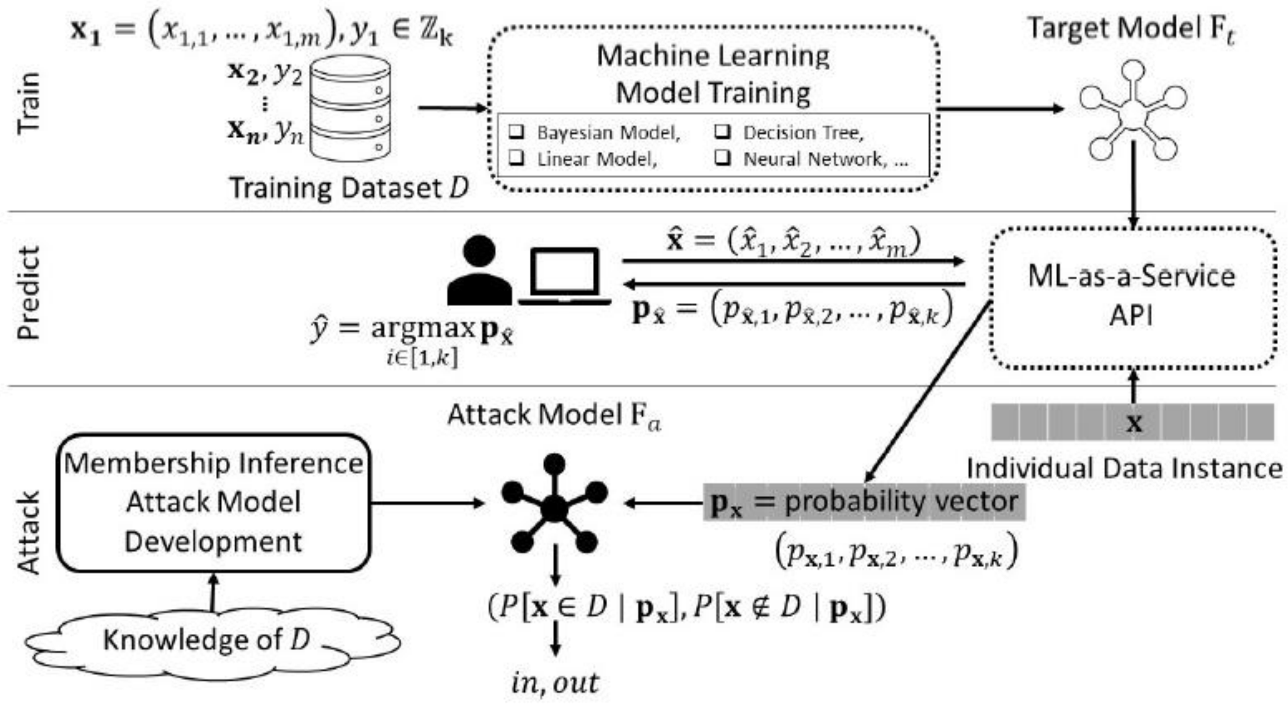

Fig. 1. The workflow of a Membership Inference Attack.

Black-box setting:
- Only black-box access through the prediction API to target model
- No knowledge about the target model (e.g. model type/architecture)

General attack formulation

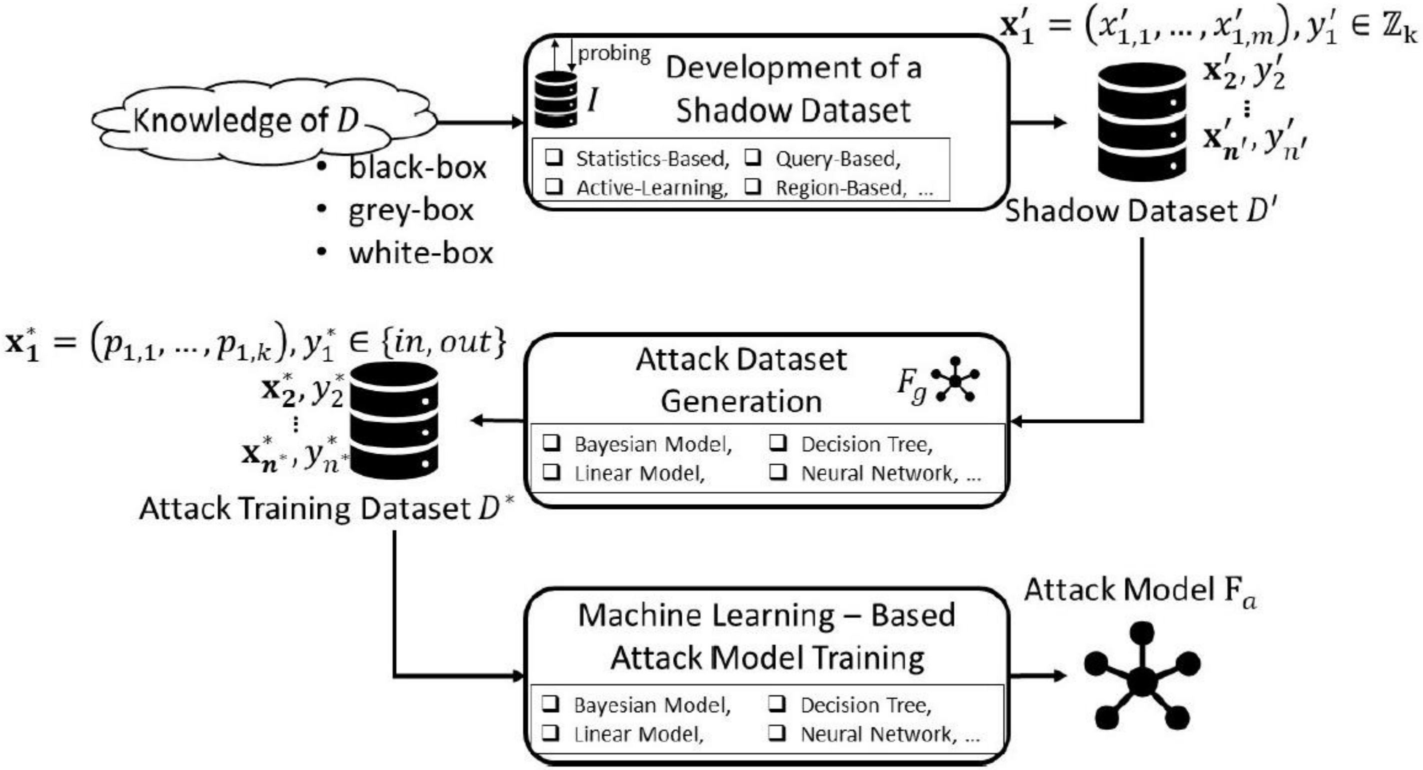

Fig. 2. Membership Attack Model Development.

Similar to the Shokri MIA approach

Generation of shadow dataset

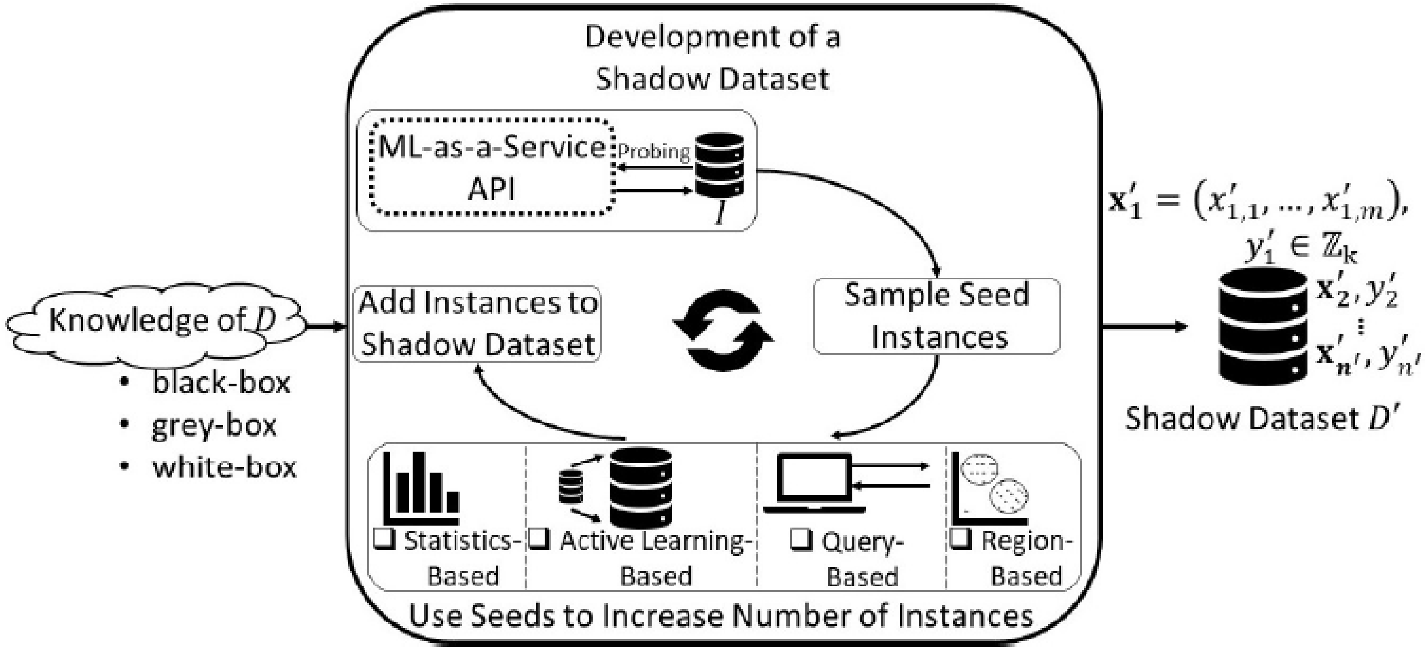

Fig. 3. Development of a Shadow Dataset.

An example: Region-based technique

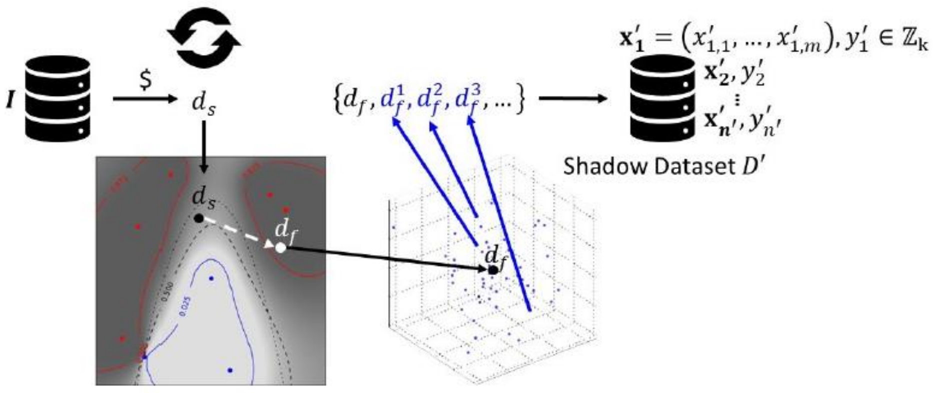

Fig. 4. Shadow dataset development using query-based and region-based techniques with black-box data knowledge. Figures adapted from images in [20] and [21]

Generation of attack model

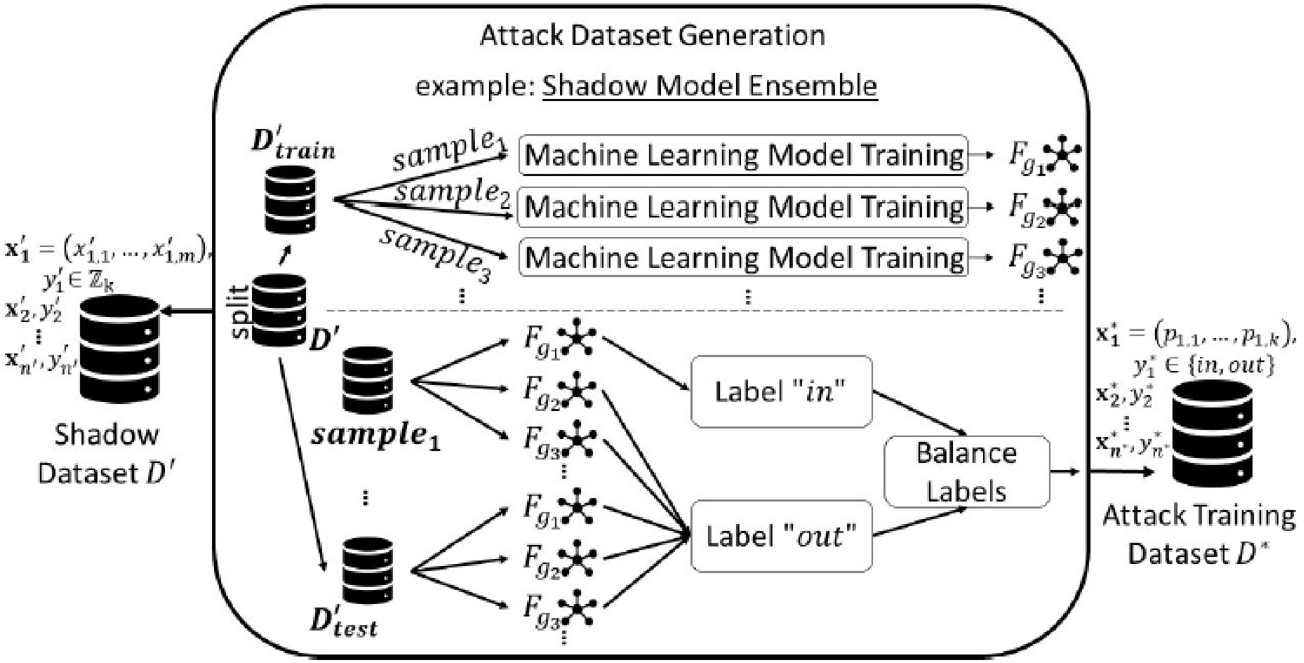

Fig. 5. Generation of an Attack Model Training Set.

3 approaches:

- data partition-based ensembles:
  split shadow dataset to q partitions and train q shadow models with each partition
- model-based ensembles:
  since type of target model is unknown, we don’t know which model is best for shadow model training. Ensemble methods (boosting/ bagging) combine multiple different models and thus reduce the risk of choosing the wrong hypothesis.
- hybrid ensemble models

2. Empirical experiments about what makes a model vulnerable against MIA

Experimental setup

- 4 model types:
  KNN, Decision Tree, Naïve Bayes, logistic regression
  - ➢ note that this paper does not consider neural networks
- 4 datasets:
  - Adult: 14 numeric features, binary classification
  - MNIST
  - CIFAR-10
  - Purchase 10/20/50/100: binary features
- Metrics of MIA:
  - attack accuracy (main metric)
  - attack precision

MIA is data-driven

Dataset	In-Class Standard Deviation	Number of Classes	Accuracy of Membership Inference
Adult	0.1433	2	59.89
MNIST	0.1586	10	61.75
CIFAR-10	0.2301	10	90.44
Purchases-10	0.3820	10	82.29
Purchases-20	0.3873	20	88.98
Purchases-50	0.3873	50	93.71
Purchases-100	0.3832	100	95.74

TABLE 1
Comparison of datasets versus membership inference attack accuracy using a decision tree model.

2 factors of datasets influence attack accuracy

- in-class standard deviation:
  similar/uniform instances makes it hard for the adversary to decide the inclusion
- number of classes:
  more classes imply more information at the target model output, i.e. more leaked information. Also, more classes -> smaller regions of each class -> single instance is more likely to alter the decision boundary -> easier attack

MIA is transferable

Purchases-20	Attack Data Generation Model			
Attack Model	DT	k-NN	LR	NB
DT	**88.98**	**87.49**	72.08	81.84
k-NN	**88.23**	72.57	**84.75**	74.27
LR	**89.02**	**88.11**	**88.99**	83.57
NB	**88.96**	78.60	**89.05**	66.34

TABLE 2
Accuracy of membership inference attack against a decision tree target model trained on the Purchases-20 dataset.

Attack Model	Target Model	Attack Data Generation Model			
		DT	k-NN	LR	NB
DT	DT	**90.44%**	85.64%	60.48%	65.78%
	k-NN	54.92%	69.32%	55.01%	51.38%
	LR	53.84%	61.06%	61.10%	50.02%
	NB	50.46%	50.58%	49.98%	50.20%
k-NN	DT	**89.96%**	81.55%	89.07%	61.10%
	k-NN	55.33%	68.32%	62.45%	50.89%
	LR	51.34%	59.58%	64.78%	50.09%
	NB	50.12%	50.61%	50.46%	50.11%
LR	DT	**90.37%**	**90.11%**	88.81%	**66.98%**
	k-NN	51.72%	69.90%	65.29%	55.64%
	LR	50.01%	64.34%	67.40%	54.49%
	NB	50.54%	50.63%	50.60%	50.29%
NB	DT	**90.42%**	89.86%	**90.52%**	63.71%
	k-NN	50.33%	68.31%	57.65%	53.08%
	LR	50.00%	64.22%	67.63%	53.54%
	NB	50.58%	50.44%	50.58%	50.01%

TABLE 3
Accuracy for CIFAR-10 dataset across experiments with various attack, data generation, and target models.

Dataset	Standard Deviation in Accuracy Results		
	Fixed F_t	Fixed F_g	Fixed F_a
Adult	**0.0093**	0.0335	0.0328
MNIST	**0.0126**	0.0347	0.0351
CIFAR-10	**0.0643**	0.1233	0.1366
Purchases-10	**0.0396**	0.1069	0.1074
Purchases-20	**0.0545**	0.1336	0.1352
Purchases-50	**0.0705**	0.1468	0.1482
Purchases-100	**0.0849**	0.1468	0.1452

TABLE 4
Standard deviation between accuracy results with (1) fixed F_t type and varying F_g and F_a types, (2) fixed F_g type and varying F_t and F_a types, and (3) fixed F_a type and varying F_t and F_g types.

An adversary may be able to develop an attack model without knowing “best attack model or the “best attack data generation model.

Variation in Generation Model

Dataset	Model Types for $(F_t^{max}, F_g^{max}, F_a^{max})$	Accuracy $(F_t^{max}, F_g^{max}, F_a^{max})$	$type(F_t^{max})$	Accuracy All $type(F_t^{max})$	$type(F_g^{max})$	Accuracy All $type(F_g^{max})$	$type(F_a^{max})$	Accuracy All $type(F_a^{max})$
Adult	(DT, DT, NB)	59.91%	DT	59.89%	DT	59.89%	NB	50.18%
MNIST	(DT, DT, LR)	61.80%	DT	61.75%	DT	61.75%	LR	54.38%
CIFAR-10	(**DT**, **LR**, NB)	90.52%	**DT**	90.44%	**LR**	67.40%	NB	50.01%
Purchases-10	(**DT**, **k-NN**, DT)	82.45%	DT	82.29%	**k-NN**	53.78%	DT	82.29%
Purchases-20	(**DT**, **LR**, NB)	89.05%	DT	88.98%	**LR**	80.50%	NB	51.29%
Purchases-50	(**DT**, **LR**, LR)	93.77%	DT	93.71%	**LR**	88.60%	LR	88.60%
Purchases-100	(**k-NN**, **LR**, DT)	95.86%	**k-NN**	95.74%	**LR**	90.23%	DT	62.19%

TABLE 7

Model set up with maximum accuracy averaged across 10 runs using 10-fold cross validation. Maximum configuration is then compared to configurations where model type is consistent across the target, generation, and attack models using each model type represented in the maximum configuration.

- A counter-intuitive conclusion:

Shadow model need not to be of the same type as the target model.

Possible explanation: The generation model's role is to characterize how the target model may be impacted by the inclusion of a particular instance. That is, how the decision boundary of the target model may reveal the inclusion of an instance. So what really matters is whether the shadow model learns similar decision boundary as the target model

- Target model type is more important than generation model or attack model.

Attacks Across Target Model Types

Dataset	LR	k-NN	DT	NB	NN
Adult	*50.13*	51.39	**55.49**	50.22	50.30
MNIST	53.25	*50.44*	**56.66**	50.48	51.70
CIFAR-10	70.25	65.99	**83.94**	*50.03*	78.00
Purchases-10	64.56	53.53	**73.85**	*50.61*	55.00
Purchases-20	75.85	55.36	**81.94**	*50.79*	59.00
Purchases-50	81.61	58.19	**88.88**	*52.08*	86.00
Purchases-100	83.78	60.11	92.19	*54.93*	**93.50**

TABLE 5
Precision of membership inference attack across 5 model types.

Dataset	LR	k-NN	DT	NB
Adult	*50.17*	51.22	**59.89**	50.18
MNIST	54.38	50.59	**61.75**	*50.81*
CIFAR-10	67.40	68.32	**90.37**	*50.01*
Purchases-10	66.82	53.78	**82.29**	*51.00*
Purchases-20	80.50	55.92	**88.98**	*51.29*
Purchases-50	88.60	59.57	**93.71**	*53.49*
Purchases-100	90.23	62.19	**95.74**	*57.61*

TABLE 6
Accuracy of membership inference attack across 4 model types.

Different model types display different vulnerabilities against MIA. Decision Tree are the most vulnerable and Naïve Bayes is the least vulnerable.

Possible explanation:
Target model whose decision boundary is unlikely to be drastically impacted by a particular instance will be more resilient to MIA. The more sensitive the target model to a single instance, the more vulnerable the model to MIA. For DT, a single instance can change the tree branches. For NB, the naïve Bayes assumption indicate a low sensitivity of the NB model to single instances.

Impact of attacker knowledge

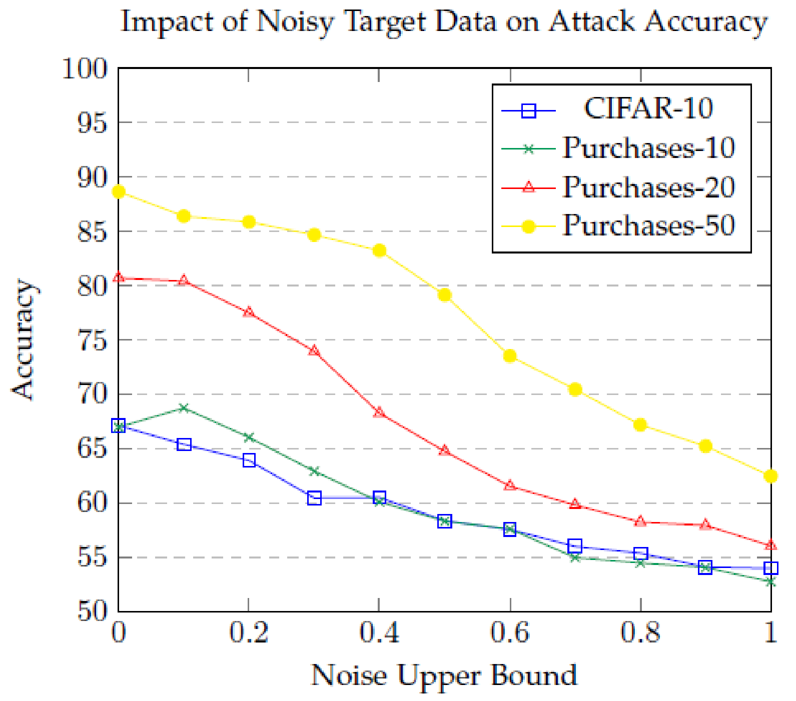

Fig. 7. Impact of Noisy Target Data on Attack Accuracy with Logistic Regression models.

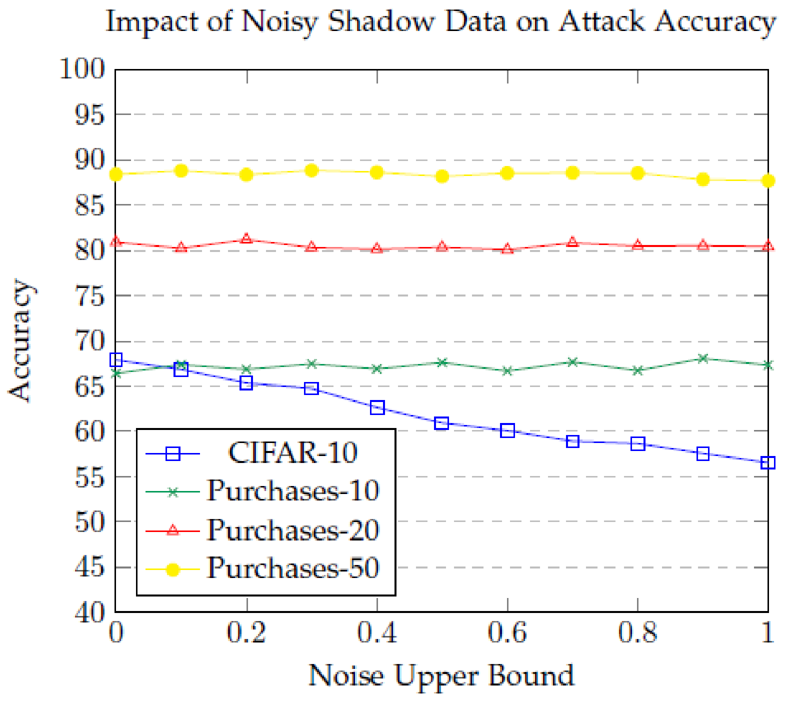

Fig. 8. Impact of Noisy Shadow Data on Attack Accuracy with Linear Regression Models.

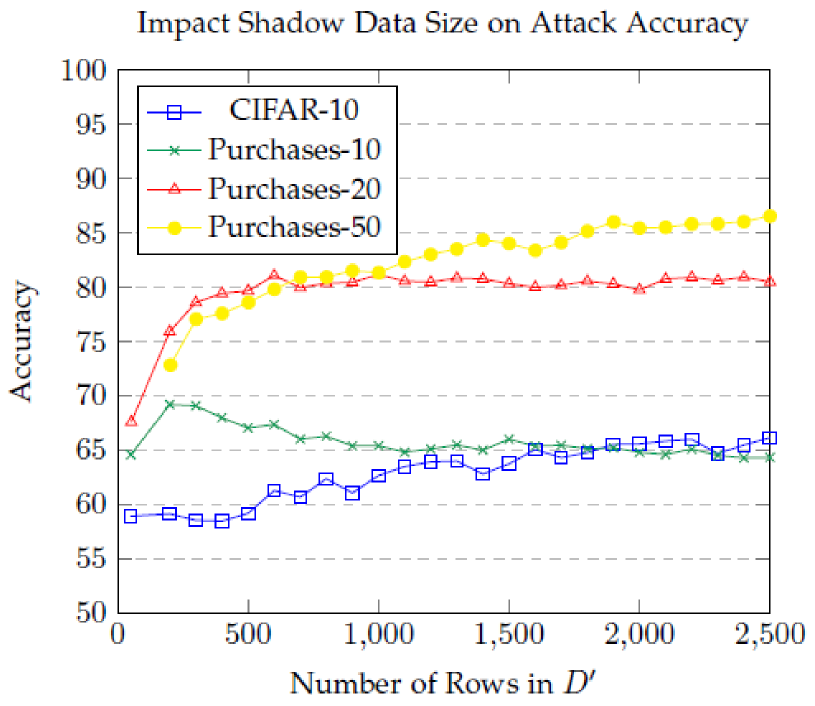

Fig. 9. Impact Shadow Data Size on Attack Accuracy with Logistic Regression models.

- Figure 7 vs Figure 8: the attacker is more likely to be successful if resources are allocated to developing strong, accurate target instances compared to perfectly representative shadow data.
- Figure 9: Larger shadow dataset improves MIA, but only to an extent.