LOGAN: Membership Inference Attacks Against Generative Models

Jamie Hayes*, Luca Melis*, George Danezis, and Emiliano De Cristofaro
Privacy in ML is 🔥🔥🔥
Privacy in ML is 🔥🔥🔥

Most papers on privacy in ML focus on inferring:
Privacy in ML is 🔥🔥🔥

Most papers on privacy in ML focus on inferring:

1. Inclusion of a data point in the training set
Privacy in ML is 🔥🔥🔥

Most papers on privacy in ML focus on inferring:

1. Inclusion of a data point in the training set
2. What class representatives look like
Privacy in ML is 🔥🔥🔥

Most papers on privacy in ML focus on inferring:

1. Inclusion of a data point in the training set
2. What class representatives look like
3. Properties of training data
Privacy in ML is 🔥🔥🔥

Most papers on privacy in ML focus on inferring:

1. Inclusion of a data point in the training set
2. What class representatives look like
3. Properties of training data
Privacy in ML is 🔥🔥🔥

Most papers on privacy in ML focus on inferring:

1. Inclusion of a data point in the training set
2. What class representatives look like
3. Properties of training data
Privacy in ML is 🔥🔥🔥

Most papers on privacy in ML focus on inferring:

1. Inclusion of a data point in the training set
2. What class representatives look like
3. Properties of training data

- Membership Inference
- Model Inversion
- Property Inference
Privacy in ML is 🔥🔥🔥

Most papers on privacy in ML focus on inferring:

1. Inclusion of a data point in the training set

**Membership Inference**

Model Inversion —> Fredrikson et al., Model inversion attacks that exploit confidence information and basic countermeasures. ACM CCS’15.

Property Inference —> Melis et al., Exploiting Unintended Feature Leakage in Collaborative Learning. IEEE S&P’19
Membership Inference
Membership Inference

Adversary wants to test whether data of a target victim has been used to train a model.
Adversary wants to test whether data of a target victim has been used to train a model

Serious problem if inclusion in training set is privacy-sensitive
Adversary wants to **test** whether data of a target **victim** has been used to train a model

Serious problem if inclusion in training set is privacy-sensitive

E.g., main task is: predict whether a smoker gets cancer
Membership Inference

Adversary wants to test whether data of a target victim has been used to train a model

Serious problem if inclusion in training set is privacy-sensitive

E.g., main task is: predict whether a smoker gets cancer

Membership inference is a very active research area, not only in machine learning...
Membership Inference (cnt’d)

Membership inference is a very active research area, not only in machine learning...
Membership Inference (cnt’d)

Membership inference is a very active research area, not only in machine learning...

Given $f(data)$, infer if $x \in data$ (e.g., $f$ is aggregation)
Membership Inference (cnt’d)

Membership inference is a very active research area, not only in machine learning...

Given $f(\text{data})$, infer if $x \in \text{data}$ (e.g., $f$ is aggregation)
[HSR$^+$08, WLW$^+$09] for genomic data
[Pyrgelis et al., NDSS’18] for mobility data
Membership Inference (cnt’d)

Membership inference is a very active research area, not only in machine learning…

Given $f(data)$, infer if $x \in data$ (e.g., $f$ is aggregation)

[HSR$^+$08, WLW$^+$09] for genomic data
[Pyrgelis et al., NDSS’18] for mobility data

Well-understood problem, besides the more obvious leakage
Membership Inference (cnt’d)

Membership inference is a very active research area, not only in machine learning…

\[ \text{Given } f(\text{data}), \text{ infer if } x \in \text{ data (e.g., } f \text{ is aggregation)} \]

[HSR\textsuperscript{+}08, WLW\textsuperscript{+}09] for \textit{genomic} data

[Pyrgelis et al., NDSS’18] for \textit{mobility} data

Well-understood problem, besides the more obvious leakage

Establish wrongdoing

Assess protection, e.g., from differentially private defenses
Machine Learning as a Service
Machine Learning as a Service

Prediction API

Training API
Machine Learning as a Service

Predictions are leaky!

Shokri et al. Membership inference attacks against machine learning models. S&P’17
Membership Inference/Discriminative

Prediction API

![Graph showing predictions for Cat, Penguin, House, Ocean]
Discriminative Model

cat | dog
Discriminative Model

Generative Model

cat | dog
Membership Inference in Generative Models?
Membership Inference in Generative Models?

Generative model

Generative API

Query

Car

Training API

Houses

Cars

Database
Inference without predictions?

Use generative models!

Train GANs to learn the distribution and a prediction model at the same time
Inference without predictions?

Use generative models!

Train GANs to learn the distribution and a prediction model at the same time
White-Box Attack

1) Predict

\[
\begin{align*}
D_{bb}(x_1) &= 0.30 \\
D_{bb}(x_2) &= 0.02 \\
D_{bb}(x_3) &= 0.79 \\
& \quad \vdots \\
D_{bb}(x_{m+n}) &= 0.64
\end{align*}
\]

2) Sort scores

\[
\begin{align*}
D_{bb}(x_{i_1}) &= 0.99 \\
D_{bb}(x_{i_2}) &= 0.98 \\
D_{bb}(x_{i_3}) &= 0.95 \\
& \quad \vdots \\
D_{bb}(x_{i_{m+n}}) &= 0.01
\end{align*}
\]

3) Take top scores

Adversary steals \( G_{\text{target}} \)
\( \quad D_{\text{target}} \)
\( \quad G \)
\( \quad D \)
\( \quad D_{wb} \)

Dataset
Black-Box Attack

1) Predict
2) Sort scores
3) Take top scores

\[
\begin{align*}
D_{bb}(x_1) &= 0.30 \\
D_{bb}(x_2) &= 0.02 \\
D_{bb}(x_3) &= 0.79 \\
& \quad \vdots \\
D_{bb}(x_{m+n}) &= 0.64
\end{align*}
\]

\[
\begin{align*}
D_{bb}(x_{i_1}) &= 0.99 \\
D_{bb}(x_{i_2}) &= 0.98 \\
D_{bb}(x_{i_3}) &= 0.95 \\
& \quad \vdots \\
D_{bb}(x_{i_{m+n}}) &= 0.01
\end{align*}
\]
Datasets

LFW

CIFAR-10

DR

Models

Attacker Model:
DCGAN

Target Model:
DCGAN, DCGAN+VAE, BEGAN
White-Box Results

LFW, top ten classes

CIFAR-10, random 10% subset
Black-Box Results

LFW, top ten classes

CIFAR-10, random 10% subset
DR Dataset

![Accuracy vs Epochs graph showing white-box and random performance.](image-url)
DR Dataset

Accuracy vs Epochs

Accuracy vs Steps

- white-box
- random
- black-box
- black-box with auxiliary knowledge
- random
(a) White-box attack

(b) Black-box attack
In a nutshell...

<table>
<thead>
<tr>
<th>Attack</th>
<th>LFW</th>
<th>CIFAR-10</th>
<th>DR</th>
</tr>
</thead>
<tbody>
<tr>
<td>White-box</td>
<td>100%</td>
<td>100%</td>
<td>95%</td>
</tr>
<tr>
<td>Black-box</td>
<td>40%</td>
<td>37%</td>
<td>22%</td>
</tr>
<tr>
<td>Black-box with aux knowledge</td>
<td>60%</td>
<td>58%</td>
<td>81%</td>
</tr>
<tr>
<td>Random guess</td>
<td>10%</td>
<td>10%</td>
<td>20%</td>
</tr>
</tbody>
</table>
Defense? Differentially Private GAN?

*Triastcyn et al. “Generating differentially private datasets using GANs.” arXiv 1803.03148
Thank you!