

# On the Protection of Private Information in Machine Learning Systems: Two Recent Approaches

Abadi et. al.  
9.23.17

Summary by Justin Chen

## Summary:

### 1. Introduction

- Applicability of security principles outlined by Saltzer and Schroeder in the 1970s to modern computer security for machine learning
- Privacy = “ability of an individual to determine whether, when, and to whom personal (or organizational) information is to be released”

### 2. Framing

- Mainly focused on classification for this review
- Identify **threat model**:
  - Extraction of training data from a model
  - Membership test on training set
- This review focuses only on defending against membership tests with differential privacy
- Two kinds of threat models:
  - **White-box attacks**:
    - Attack can **inspect the internals** of the model
  - **Black-box attacks**:
    - Attackers can **query model** on new input arbitrary number of times
- Attackers may also **poison** the dataset during training and read intermediate states during training, but this paper does not focus on those attacks
- Privacy for Inference Inputs
  - This work focuses on training data only
  - Training and inference have different problems when it comes to privacy
    - Training:
      - Learning algorithms memorize data points
      - Could violate data-retention policies
      - Individuals may be more comfortable with submitting their data for training instead of inference
        - Data point can’t be directly analyzed by trained model, assuming data points are not stored
    - Inference:
      - Trained model can be used to learn about individuals
        - This is beyond the scope of this review
- A Systems Perspectives:
  - Concerns with learning pipeline itself
    - Sanitizing data, anonymization, pseudonymization, aggregation, generalization, removal of outliers

- Access control on collected data
- Data-retention policies and data deletion mechanisms

### 3. Noisy SGD

- SGD is common technique for learning parameters
- Using noise is also common and has been heavily explored in several other works
- Noise is also used for privacy
- Balance privacy and accuracy with noisy gradients

### 4. PATE

- Use an ensemble of trained models, each trained on a subset of the data, to train a single model (teacher-student training)
  - If posed with a white-box attack, attacker could infer things about datasets based on activations of model
  - Student queries teachers about unlabeled examples
  - Teachers can be deleted after training student to preserve privacy of original data

### 5. Principles, Revisited

- Saltzer and Schroeder principles applied to Noisy SGD and PATE
- “Training of a model from data is loosely analogous to applying a cryptographic transformation to the data”

#### 1. Economy of mechanism

- Protection mechanism should be designed as simple as possible
- PATE and Noisy SGD don't abide by this principle. Both are too complicated.

#### 2. Fail-safe defaults

- Deny access by default and refuse permissions
- Difficult to apply to PATE and Noisy SGD

#### 3. Complete mediation

- Must always pass protection mechanism to access data
- Internals of model should not be sensitive in cases of white-box attacks
- Complete mediation requires system-wide perspective

#### 4. Open design

- Design of protection mechanism should not depend on secrets and design should not be kept secret
  - Similar to Kerckhoffs's principle
- Noisy SGD and PATE conform to this

#### 5. Separation of privilege

- Must use multiple independent keys for unlocking access
- Difficult to apply to both algorithms considered here

#### 6. Least privilege

- Every program and every user should operate with the least set of privileges necessary to complete the task
- Not obvious how to apply to algorithms considered here

#### 7. Least common mechanism

- Difficult of providing mechanisms shared by more than one user
- May introduced unintended communication channels and behavior
- Hard to satisfy

- Parameters of algorithms considered here could comply with this principle
- 8. **Psychological acceptability**
  - Advocates ease of use - aka principle of least astonishment
  - If users mental model of protection goals matches presented mechanisms, then can help minimize errors
  - Difficult for Noisy SGD because it's proof is heavily involved, but intuition is easier to express for PATE
- 9. **Work factor**
  - Measuring resources of attacker against cost of circumventing protection mechanism
    - e.g. NP-Completeness and prime factorization
- 10. **Compromise recording**
  - Detecting and reporting failures of protection is advantageous
  - Noisy SGD and PATE reveal failures and training data
    - This principle was not considered when designing these algorithms
  - Open debate in secure machine learning field and is a shortcoming of the theory

#### **Conclusion**

- Important to understand how learning algorithms will behave in practice
- Fundamental security principles should guide the design and analysis of future secure learning algorithms