12/18/2025

ML PRIVACY ATTACKS

A Comprehensive Analysis of 14 Foundational Research Papers

– MUHAMMAD JUNAID

– BITF22M035

PRESENTATION CONTEXT

1. Context & Foundation (Slides 2-3)
2. The Attack Deep-Dives (Slides 4-8)
3. Summary & Action (Slides 9-10)
4. Q&A

THE CORE PROBLEM

WHY IS ML PRIVACY AN ISSUE?

Data Memorization: ML models don't just learn patterns; they often "memorize" specific training samples.

Sensitivity: Training sets often contain PHI (Protected Health Information) or PII (Personally Identifiable Information).

The Goal: This research identifies how attackers exploit models to leak this sensitive data

ANALYSIS OF 14 FOUNDATIONAL PAPERS

Selection Criteria: High-impact, peer-reviewed papers (2012–2023).

Attack Taxonomy:

1. Membership Inference (3 Papers)
2. Model Inversion (3 Papers)
3. Poisoning & Backdoors (2 Papers)
4. Model Extraction (2 Papers)
5. Property Inference & Generative Attacks (2 Papers)
6. Meta-Surveys (2 Papers)

MEMBERSHIP INFERENCE ATTACKS (MIA)

WAS THIS PERSON IN YOUR DATASET?

Key Paper: Shokri et al. (2017) – Introduced "Shadow Models" to learn how a model reacts to seen vs. unseen data.

Evolution: Choquette-Choo (2021) proved attacks work with Labels only (no confidence scores needed).

Main Driver: Overfitting. The more a model overfits, the easier it is to detect training members.

RECONSTRUCTING PRIVATE DATA

Key Paper: Fredrikson et al. (2015) – Reconstructed recognizable faces from a facial recognition API.

Advance: Zhang et al. (2020) used GANs (Generative Adversarial Networks) to create high-fidelity reconstructions.

Impact: Proves that model parameters implicitly store raw data features.

POISONING & BACKDOOR ATTACKS

SABOTAGING THE TRAINING PROCESS

Poisoning: Biggio et al. (2012) showed that injecting malicious samples can manipulate the model's decision boundaries.

Backdoors (BadNets): Gu et al. (2017) demonstrated "Trojan" models.

- **The Trigger:** A specific pattern (like a sticker on a stop sign) causes the model to fail, while it behaves normally on all other data.

CLONING PROPRIETARY MODELS

The Threat: Tramèr et al. (2016) – Attackers query a Black-box API to "steal" the internal logic of a model.

Knockoff Nets: Orekondy et al. (2019) – Using Reinforcement Learning to optimize queries and create a functional clone of a high-value model for free.

PROPERTY INFERENCE & GENERATIVE RISKS

LEAKING HIDDEN STATISTICS

- **Property Inference:** Ganju et al. (2018) – Extracting aggregate dataset features (e.g., gender ratios) not intended for release.
- **Generative Models:** Song et al. (2022) – Proved that GANs and VAEs are highly susceptible to membership and attribute leakage due to high memorization.

COMPARATIVE SUMMARY TABLE

01 ATTACK TYPE	02 MAJOR PAPER	03 TARGETED ASSET	04 IMPACT
Membership	Shokri (2017)	Individuals	Privacy Breach
Inversion	Fredrikson (2015)	Raw Samples	Identity Theft
Poisoning	Biggio (2012)	Model Logic	System Failure
Extraction	Tramèr (2016)	Intellectual Property	Financial Loss

CONCLUSION & RECOMMENDATIONS

HOW DO WE SECURE THE FUTURE OF ML?

Differential Privacy (DP): The only mathematical guarantee to limit individual data leakage.

Robust Training: Using adversarial training and data sanitization to prevent poisoning.

Rate Limiting: Restricting API queries to prevent model extraction.

Final Word: Privacy is an ongoing battle; as models get smarter, attacks get more sophisticated.

Q&A

– MUHAMMAD JUNAID

– BITF22M035

THANK YOU

– MUHAMMAD JUNAID

– BITF22M035