Secure Systems Group, Aalto University

Mika Juuti, Sebastian Szyller, Alexey Dmitrenko, Samuel Marchal, N. Asokan

**PRADA: Protecting Against DNN Model Stealing Attacks**

Why model confidentiality? Avoid whitebox attacks & retain business advantage.

But attackers can use prediction APIs to extract models (build a substitute model).

Stateful analysis of client queries can prevent model extraction.

---

**Model Extraction Attack**
- **Capabilities**: only query-access to prediction API
- **Goal**: build a substitute model using few queries
  - Reproduce predictive behaviour
  - Forge transferable adversarial examples

[Diagram showing DNN model, Prediction API, Client, and PRADA]

---

**Novel Extraction Attack**
- **Jb-topk**: directions of k closest classes
- **Jb-self**: directions of class centroid
- Increased performance over state-of-the-art:
  - +15-30% transferability of adversarial examples
  - +15-20% prediction accuracy
- Synthetic samples improve transferability
- Natural samples improve predictive behaviour

---

**PRADA: Stateful detection of model extraction**
- Analyses the evolution in the distribution of client queries
- Models the user behaviour as a function of novel queries
- Parameterised with window size $W$ and threshold of derivative ratio $\Delta$
- Compares the ratio of subsequent derivatives

[Graph showing Minimum distance trend vs Number of samples seen (from each class)]

- Detects all known attacks quickly
- Low overhead (<25 MB) on MNIST and GTRSB

---

**MNIST: Synthetic query impact**
- Agreement vs Total queries
- Targeted transferability
- Papernot
- Tramer

[Graph showing Samples in the growing set vs Number of samples seen so far]