We present a novel method which improves the state-of-the-art and was developed in collaboration with industry. We implemented it as an improvement of a fraud detection system at O2 Czech Republic and supported its deployment into the production environment.

**Problem**

- Challenge in many security applications of machine learning (fraud detection, computer network intrusion detection, spam filtering): attackers adapt in order to avoid detection. But classical machine learning assumes: future observations follow the same distribution as training data.
- Emerging methods to limit the adaptability are not yet applicable in the most practical settings.

**Our Contributions**

- Unlike in the majority of existing methods, threshold randomization to complicate detection avoidance.
- First method with data-driven modeling of attackers (frauds) with no restrictions on a used machine learning algorithm, no restrictions on training data.
- First game-theoretic optimization enabling control over false alarms rate (crucial in security).
- Enhances a general state-of-the-art model:
  - Rigorous theoretical analysis of the general model (12 discovered and formally proven facts)
  - New scalable algorithms, improved worst-case complexity compared to existing baselines:
    - Computation of Nash equilibrium generally belongs to PPAD complexity class, for the model we developed a linear time algorithm.
    - Strong Stackelberg equilibrium: instead of solving a linear number of linear programs using general solvers → a quadratic time algorithm; for randomization under false alarm restriction: a cubic time algorithm.
- Method uses data-driven modeling of adversaries from another state-of-the-art model
  - Introducing applicability to domains with continuous features
  - Preserving possibility to model bounded adaptability of adversaries, variability of adversaries
  - Preparing a journal publication based on the thesis

**Experimental Evaluation and Deployment**

**Evaluation of the General Model Enhancements**

- 170 ROC curves of real-world computer network intrusion detection systems, O2 CZ fraud detection module ROC curve, parameters of the general model generated at random.
- Dramatic improvement in running time and scalability, thus the method remains applicable to larger datasets.
- Nash equilibrium algorithm: three orders of magnitude faster on instances twice as large.
- Strong Stackelberg equilibrium algorithm: an order of magnitude faster on instances twice as large.
- Strong Stackelberg equilibrium is the best fit for security, false alarm rate restriction is crucial in practice. Enforcing the restriction leads to performance sacrifice (>3 times worse value of the defender’s utility function compared to the unconstrained setting).

**Deployment: module for a fraud detection system at O2 Czech Republic**

- Method used to improve robustness of a machine learning module implemented by the author as the last line of defense for the fraud detection system at O2 Czech Republic.
- Training dataset used to create a classifier.
- Validation dataset used to derive randomization of a classification threshold.
- Test dataset used to estimate performance against both static and adaptive populations.
- Results of evaluation:
  - Application of the method did not result in a notable decrease of performance against static population.
  - The method can improve robustness of the classifier against adaptive opponents.
  - Management decided to deploy it to production.