Network Security - BA6403
Research focus & interests

- **Security of future networks**
  - AI for security functions in B5G/6G + mobile networks
  - Security automation in constrained distributed environments (edge security)
  - Secure network architecture
  - B5G/6G networks simulation with *cyber range*

- **AI & security + Trustworthy AI systems**
  - AI automation in security operations
  - Defenses against adversarial AI attacks
  - Secure AI system development & deployment
  - Security assessment for AI systems

- **Cyber insurance for emerging technologies**
  - Security testing and security posture management
  - Security risk and compliance management (NIS2, CR Act, AI act)
  - Security training & security scenarios simulation with *cyber range*
  - Targeted applications: AI, cloud, edge network, critical infrastructures
Secure & Trustworthy AI systems
Trustworthy AI

Resilience against attacks
- Evasion attack
- Poisoning attack
- Data inference

Trustworthy AI pillars
- Security
- Fairness
- Transparency
- Accountability
- Privacy
- Reliability
Security of AI
AI systems are vulnerable against new attacks that only targets them: adversarial attacks

- Model poisoning / backdooring
  - Adversarial data to cause skew in model decision

- Model evasion
  - Input maliciously crafted to be misclassified

- Model stealing
  - Uncovering ML models via probing

- Data inference
  - Uncovering private data from ML models

Inference attacks

ML model integrity
Prediction integrity
ML model confidentiality
Training data confidentiality
Adversarial ML: attack surface

Pre-trained model

Data poisoning

Storage platform
Training data

Training platform
Training process

Deployment platform
ML model

Inference process

API boundary

Model poisoning

Model poisoning

Model poisoning

ML Libraries

Data inference

Model stealing

Model evasion

Data poisoning

Data Sources

VTT – beyond the obvious

07/08/2024
Research interests in AI security

*RQ1*: How to ensure and provide evidence that AI systems are secure?

**Security assessment & certification for AI systems**
- Metrics to quantify the security level of AI systems
- Methods and tools for security testing (to compute security metrics)

*RQ2*: How to make AI systems resilient against adversarial attacks?

**Detection of and protection against adversarial attacks**
- Detection approach against evasion attacks
- Protection against poisoning attacks in federated learning

*RQ3*: How to make AI systems resilient against the main cybersecurity threats?

**Mitigation of supply chain attacks against AI systems**
- Identification of AI-specific supply chain attacks
- Definition of conventional and novel mitigation approaches
Security assessment & certification for AI systems
Security assessment for AI Evasion attacks

Aimed functionalities
- Produce quantifiable measures of security/resilience
- Provide an upper bound estimation for security vulnerability
- Implement realistic attacker capabilities
- Applicability against virtually any ML model

Main targeted applications
- Identify and fix vulnerabilities in ML models before deployment
- Select the most secure + reliable (+ explainable + etc.) ML model
  - Evaluate the performance/security/(explainability) trade-off
- Document the performance and the security posture of ML-based systems
  - Support for AI risk management
  - Evidence for security compliance
Empirical security diagnosis for evasion attacks

Process
- Generate synthetic queries: adversarial examples
- Analyze model outputs: correct/incorrect prediction
- Compute resilience metrics based on attacks stats and success
- Generate vulnerability/resilience report

Implements several blackbox evasion attacks

Computes 3 resilience metrics
- Impact
- Complexity
- Detectability
Protection against poisoning attacks in federated learning
Poisoning attacks in federated learning

Attack process
- Malicious client(s) craft poisoned local model
- Send update to aggregator
- Aggregation compromises global model

Impact of attack
- Compromise integrity of global model
  - Decrease in overall accuracy / performance
  - Embedding of backdoors
- Affect all model users
Defenses against FL poisoning


- Privacy-preserving process implemented in aggregator
- Cluster local models to discard obviously malicious updates
- Adaptive clipping to limit negative impact of any single model
- Adaptive noising to mitigate targeted changes to global model

Protection in hierarchical federated learning [3]

- Adapt process with intermediate aggregation layers

---

Mitigation of supply chain attacks against AI systems
Supply chain attack vectors

Attacks through data
Attacks through model
Attacks through software/libraries
Attacks through GPU

Data Sources
- Training Data

Storage platform
- Training data

Training platform
- GPU

Deployment platform
- ML model

Inference process
- API boundary

Inference Data

Base Model
Securing the AI supply chain

Vectors for ML supply chain attacks to secure

- **Training data**
  - Data integrity and quality is difficult to enforce and verify

- **Pre-trained ML models**
  - Complex ML models can be compromised with backdoors or biased
  - ML model integrity is very hard to verify (just weights…)

- **ML software & libraries**
  - ML library compromise is more subtle and difficult to detect
  - E.g., change in objective function can compromise ML algorithm

- **ML hardware, e.g., GPU**
  - Lesser risk, might be harder to compromise