Jing Xu

CISPA, Germany   jingxu.buaa@gmail.com   +491747735811   Jing Xu - GoogleScholar
xujing1994.github.io   Jing Xu - LinkedIn   xujing1994

Profile

Machine Learning researcher specializing in **trustworthy and privacy-preserving machine learning**, with expertise in differential privacy, membership inference, generative models, and federated learning. Strong background in securing foundation models (LLMs, GNNs, IARs) and evaluating, mitigating privacy risks in large-scale AI systems. Author of 20+ papers at top-tier machine learning and security venues, including NeurIPS, ICML, CCS, Euro S&P, ACSAC, and other venues. Experienced in reproducibility-driven research, cross-disciplinary collaboration, and leveraging cutting-edge AI technologies to address real-world challenges.

Education

Delft University of Technology, The Netherlands *Oct 2019 – May 2024*
PhD in Computer Science

- Research Focus: Machine Learning, Graph Neural Networks, AI Security, Explainable AI.
- Thesis: *Exploring backdoor attacks on graph neural networks*; Supervisors: Prof. Inald Lagendijk, Prof. Frans Oliehoek, Prof. Stjepan Picek.
- Achievements: Published 15+ papers at top-tier ML/Security conferences & journals; research collaborations with University of Padua, TU Darmstadt, and industry partners.

Beihang University, China *Sep 2016 – Jan 2019*
MSc in Optical Engineering

- Specialization: Electrical Engineering, Signal Processing, Computer Vision
- GPA: **3.857/4.0**-RANK: **top 5%**

Shanghai University, China *Sep 2012 – July 2016*
BSc in Electrical Engineering

- Specialization: Computer Vision, Signal Processing, Automata
- GPA:**3.86/4.0**-RANK: **1/931**

Experience

Proleptic Lecturer in AI Safety (AP) *York, United Kingdom*
York University *Apr. 2026 – Present*

- Developed and promote reseach on Safety of Large Language Models;
- Collaborate with interdisciplinary teams to address ethical and societal implications of AI technologies;
- Continue publishing and presenting at top-tier machine learning venues, e.g., NeurIPS, ICML.

Postdoc Researcher *Saarbrücken, Germany*
CISPA Helmholtz Center for Information Security *Nov. 2023 – Mar. 2026*

- Developed privacy-preserving machine learning mechanisms for advanced AI models, e.g., large language models (LLMs) and graph prompt learning.
- Led projects on LLM unlearning, active defense against GNN model stealing, and IAR model ownership protection; Built robust ML pipelines using Git, Jupyter, and Docker for seamless development.
- Collaborated with industrial partners (e.g., DeepMind); Mentored junior researchers.
- Continued publishing and presenting at top-tier machine learning venues, e.g., NeurIPS, ICML.

PhD in Computer Science *Delft, Netherlands*
TU Delft *Oct 2019 - May 2024*

- Pioneered novel (e.g., explainability-based) backdoor attacks and defenses on GNNs.
- Developed benchmarks and open-source tools for backdoor attack and detection algorithms with an

emphasis on reproducibility and transferability of findings.
- Published extensively and presented at top-tier venues (CCS, Euro S&P, and ACSAC).

Researcher Intern *Beijing, China*
Momo Technology Company, Deep Learning Lab *Jan 2019 - July 2019*

- Developed data pipelines and automated workflows for processing facial image datasets used in the training and evaluation of the face recognition model.
- Developed GAN-based methods for face recognition and object detection.
- Ensured deep learning models' robustness against adversarial attacks and spoofing.

Research Interests

- Trustworthy and Privacy-preserving Machine Learning
- Differential Privacy and Synthetic Data
- Model Unlearning, and AI Content Watermarking
- Reproducibility, Explainability, and Responsible AI

Technologies

Programming & Frameworks: Python, C++, MATLAB, PyTorch, TensorFlow, Hugging Face Transformers, Scikit-learn, OpenCV

Development & Deployment: Linux, Slurm, Docker, Git, GitLab, tmux, SSH, Jupyter

Mathematics: Optimization, Control Theory, Probability, Linear Algebra

Soft Skills: Scientific Writing, Cross-functional Collaboration, Project Coordination

Publications

[1] **Jing Xu***, Franziska Boenisch, Adam Dziedzic. ADAGE: Active Defenses Against GNN Extraction. *AsiaCCS 2026.*

[2] Marco Arazzi*, Mauro Conti, Stefanos Koffas, Marina Krcek, Antonino Nocera, Stjepan Picek, **Jing Xu**. Label Inference Attacks Against Node-level Vertical Federated GNNs. *Under Major Revision at Transactions on Dependable and Secure Computing (TDSC)*, 2026.

[3] Xun Wang*, Vincent Hanke, **Jing Xu**, Michael Backes, Franziska Boenisch, Adam Dziedzic. Frequency-Domain Model Fingerprinting for Image Autoregressive Models. *AAAI Workshop on AI Governance: Alignment, Morality, Law and Design (AIGOV), 2026.*

[4] Kunhao Li*, Di Wu, Jun Bai, **Jing Xu**, Lei Yang, Ziyi Zhang, Yiliao Song, Wencheng Yang, Taotao Cai, Yan Li. Who Owns This Sample: Cross-Client Membership Inference Attack in Federated Graph Neural Networks. *Under review at ICML 2026.*

[5] Adarsh Jamadandi*, **Jing Xu**, Adam Dziedzic, Franziska Boenisch. Memorization in Graph Neural Networks. *The Annual Conference on Neural Information Processing Systems (NeurIPS), 2025.*

[6] Xun Wang*, **Jing Xu**, Franziska Boenisch, Michael Backes, Adam Dziedzic. Efficient and Privacy-Preserving Soft Prompt Transfer for LLMs. *ICML 2025.*

[7] **Jing Xu***, Franziska Boenisch, Iyiola Emmanuel Olatunji, Adam Dziedzic. DP-GPL: Differentially Private Graph Prompt Learning. *ICLR Workshop on Foundation Models in the Wild, 2025.*

[8] Iyiola Emmanuel Olatunji*, Franziska Boenisch, **Jing Xu**, Adam Dziedzic. Adversarial Attacks on Graph-aware LLMs.*arXiv 2025.*

[9] Bart Pleiter, Behrad Tajalli, Stefanos Koffas, Gorka Abad, **Jing Xu**, Martha Larson, Stjepan Picek. Backdoor Attacks on Transformers for Tabular Data: An Empirical Study. *arXiv, 2024.*

[10] **Jing Xu***, Stjepan Picek. Poster: Multi-target & Multi-trigger Backdoor Attacks on Graph Neural Networks. *The ACM Conference on Computer and Communications Security (CCS), 2023.*

[11] **Jing Xu***, Stefanos Koffas, Oguzhan Ersoy, Stjepan Picek. Watermarking Graph Neural Networks based on Backdoor Attacks. *IEEE European Symposium on Security and Privacy (Euro S&P), 2023.*

[12] **Jing Xu***, Gorka Abad, Stjepan Picek. Rethinking the Trigger-injecting Position in Graph Backdoor Attack. *Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), 2023.*

[13] Gorka Abad*, **Jing Xu**, Stefanos Koffas, Behrad Tajalli, Stjepan Picek, Mauro Conti. SoK: A Systematic Evaluation of Backdoor Trigger Characteristics in Image Classification. *arXiv, 2023.*

[14] **Jing Xu***, Stefanos Koffas, Stjepan Picek. Unveiling the Threat: Investigating Distributed and Centralized Backdoor Attacks in Federated Graph Neural Networks. *Digital Threats: Research and Practice (DTRAP), 2023.*

[15] Stefanos Koffas, Behrad Tajalli, **Jing Xu**, Mauro Conti and Stjepan Picek*. A Systematic Evaluation of Backdoor Attacks in Various Domains. *Embedded Machine Learning for Cyber-Physical, IoT, and Edge Computing: Use Cases and Emerging Challenges, pages 519 - 552, 2023.*

[16] **Jing Xu***, Rui Wang, Kaitai Liang, Stjepan Picek. More is Better (Mostly): On the Backdoor Attacks in Federated Graph Neural Networks. *Annual Computer Security Applications Conference (ACSAC), 2022.*

[17] **Jing Xu***, Stefanos Koffas, Stjepan Picek. On Exploring Backdoor Attacks in Federated Graph Neural Networks. *Learning from Authoritative Security Experiment Results (LASER) Workshop, 2022.*

[18] **Jing Xu***, Stjepan Picek. Poster: Clean-label Backdoor Attack on Graph Neural Networks. *The ACM Conference on Computer and Communications Security (CCS), 2022.*

[19] Mauro Conti, Jiaxin Li*, Stjepan Picek, **Jing Xu**. Label-Only Membership Inference Attack against Node-Level Graph Neural Networks. *AISec, CCS Workshop, 2022.*

[20] Stefanos Koffas*, **Jing Xu**, Mauro Conti, Stjepan Picek. Can You Hear It? Backdoor Attacks via Ultrasonic Triggers. *The ACM Workshop on Wireless Security and Machine Learning (WiseML), 2022.*

[21] **Jing Xu***, Minhui(Jason) Xue, Stjepan Picek. Explainability-based backdoor attacks against graph neural networks. *The ACM Workshop on Wireless Security and Machine Learning (WiseML), 2021.*

Teaching & Services

- **Guest Lecturer:**
  - Security and Privacy of Machine Learning, Radboud University, The Netherlands, 2022;
  - Advanced Topics in Computer and Network Security, University of Padua, Italy, 2022.
- **Supervision:**
  - MSc and BSc dissertation projects on trustworthy machine learning (TU Delft, 2022 - 2024);
  - Mentoring junior PhD students on multiple projects (CISPA, 2023 - 2025).
- **Organiser:** IJCNN 2025/2026 Special Session on Trustworthy and Explainable Federated Learning: Towards Security and Privacy Future.
- **Conference PC/Reviewer:**
  - AAAI 2026, PETS 2026, AAAI-AIGOV 2026, CCS-LAMPS 2025;
- **Journal Reviewer:**
  - TIFS, TOPS, TKDE, TDSC;

Research Visits & Collaborations

- **DeepMind**, Project collaboration, 2024
- **University of Padua**, Visiting researcher, Italy, 2022 - 2023
- **TU Darmstadt**, Visiting researcher, Germany, 2021

List of Referees

- Referee 1: Dr. Franziska Boenisch, CISPA Helmholtz Center for Information Security, Germany, boenisch@cispa.de
- Referee 2: Dr. Stjepan Picek, Radboud University, The Netherlands, stjepan.picek@ru.nl
- Referee 3: Prof. Timothy Hospedales, University of Edinburgh, The United Kingdom, t.hospedales@ed.ac.uk
- Referee 4: Prof. Dr. George Smaragdakis, Delft University of Technology, The Netherlands, G.Smaragdakis@tudelft.nl