Yixin Wu

✉ yixin.wu@cispa.de • yxoh.github.io • Yixin Wu

Education

Saarland University & CISPA Helmholtz Center — **Saarbrücken, Germany**
Ph.D. in Computer Science — *November 2021 –*
Advisor: Prof. Michael Backes & Dr. Yang Zhang

Sichuan University — **Sichuan, China**
Bachelor in Cyber Science and Engineering — *September 2017 – June 2021*
Advisor: Prof. Cheng Huang

National University of Singapore — **Singapore, Singapore**
Workshop — *July 2019 – August 2019*
Advisor: Prof. Hugh Anderson

Research Interests

- Trustworthy Machine Learning (Data Privacy)
- Misinformation, Hate Speech, and Memes

Service

- Journal reviewer
  - TIFS, TKDE
- External reviewer
  - 2024: ICLR, WWW, SP
  - 2023: CCS, ICML, NeurIPS, KDD
  - 2022: USENIX Security, AAAI, PoPETs
  - 2021: CCS, ICLR

Publication

Conference

[1] **Yixin Wu**, Yun Shen, Michael Backes, and Yang Zhang. Image-perfect imperfections: Safety, bias, and authenticity in the shadow of text-to-image model evolution. In *ACM Conference on Computer and Communications Security (CCS)*. ACM, 2024.

[2] **Yixin Wu**, Rui Wen, Michael Backes, Pascal Berrang, Mathias Humbert, Yun Shen, and Yang Zhang. Quantifying Privacy Risks of Prompts in Visual Prompt Learning. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2024.

[3] **Yixin Wu**, Xinlei He, Pascal Berrang, Mathias Humbert, Michael Backes, Neil Zhenqiang Gong, and Yang Zhang. Link Stealing Attacks Against Inductive Graph Neural Networks. In *Privacy Enhancing Technologies Symposium (PETS)*. PETS, 2024.

[4] Xinyue Shen*, **Yixin Wu***, Michael Backes, and Yang Zhang. Voice Jailbreak Attacks Against GPT-4o. *CoRR abs/2405.19103*, 2024.

[5] Yiting Qu, Xinyue Shen, **Yixin Wu**, Michael Backes, Savvas Zannettou, and Yang Zhang. UnsafeBench: Benchmarking Image Safety Classifiers on Real-World and AI-Generated Images. *CoRR abs/2405.03486*, 2024.

[6] **Yixin Wu**, Ning Yu, Michael Backes, Yun Shen, and Yang Zhang. On the Proactive Generation of Unsafe Images From Text-To-Image Models Using Benign Prompts. *CoRR abs/2310.16613*, 2023.

[7] **Yixin Wu**, Ning Yu, Zheng Li, Michael Backes, and Yang Zhang. Membership Inference Attacks Against Text-to-image Generation Models. *CoRR abs/2210.00968*, 2022.

[8] Xinlei He, Rui Wen, **Yixin Wu**, Michael Backes, Yun Shen, and Yang Zhang. Node-Level Membership Inference Attacks Against Graph Neural Networks. *CoRR abs/2102.05429*, 2021.

Journal

[9] **Yixin Wu**, Cheng Huang, Xing Zhang, and Hongyi Zhou. Grouptracer: Automatic attacker ttp profile extraction and group cluster in internet of things. *Security and Communication Networks (SCN)*, 2020.

[10] **Yixin Wu**, Yuqiang Sun, Cheng Huang, Peng Jia, and Luping Liu. Session-based webshell detection using machine learning in web logs. *Security and Communication Networks (SCN)*, 2019.