Xinlei He — CV

✉ xinleihe@hkust-gz.edu.cn • xinleihe.github.io • Xinlei He

Employment

Hong Kong University of Science and Technology (Guangzhou) **Guangzhou, China**
Assistant Professor, DSA & IoT Thrust, Information Hub *February 2024 – Now*

Education

CISPA Helmholtz Center for Information Security **Saarbrücken, Germany**
Ph.D. in Computer Science *February 2020 – September 2023*
Advisor: Dr. Yang Zhang

Fudan University **Shanghai, China**
Master in Computer Science *September 2017 – January 2020*
Advisor: Prof. Yang Chen

Fudan University **Shanghai, China**
Bachelor in Computer Science *September 2013 – June 2017*
Advisor: Prof. Yang Chen

Research Interests

- Trustworthy Machine Learning
- Misinformation, Hate Speech, and Memes

Services

- TPC Chair/Co-chair
  - 2024: SocialMeta (`https://socialmeta2024.github.io/`)
- Conference PC Member
  - 2025: ACM ASIACCS, IEEE SaTML, IEEE Euro S&P, ICLR, CVPR, IJCAI
  - 2024: IEEE S&P, ACM ASIACCS, IEEE ICDCS, ACSAC
  - 2022: ESORICS
  - 2021: ESORICS (Poster Session)
  - 2020: SocInfo
- Conference Reviewer
  - 2024: ICLR
  - 2023: IEEE CVPR, NeurIPS
  - 2022: LoG
- Journal Reviewer
  - 2023: IEEE TDSC, ACM TOPS
  - 2022: IEEE TDSC

Awards

- The Norton Labs Graduate Fellowship 2022 (20,000 USD)
- Best paper award at LAMPS 2024

Publication

†: Corresponding author.

Conference

[1] Lijin Wang, Jingjing Wang, Tianshuo Cong†, **Xinlei He**†, Zhan Qin, and Xinyi Huang. From Purity to Peril: Backdooring Merged Models From ”Harmless” Benign Components. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2025.

[2] Jingyi Zheng, Tianyi Hu, Tianshuo Cong, and **Xinlei He**†. CL-Attack: Textual Backdoor Attacks via Cross-Lingual Triggers. In *AAAI Conference on Artificial Intelligence (AAAI)*. AAAI, 2025.

[3] Yichen Gong, Delong Ran, **Xinlei He**, Tianshuo Cong, Anyu Wang, and Xiaoyun Wang. Safety Misalignment Against Large Language Models. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2025.

[4] Tianshuo Cong, Delong Ran, Zesen Liu, **Xinlei He**, Jinyuan Liu, Yichen Gong, Qi Li, Anyu Wang, and Xiaoyun Wang. Have You Merged My Model? On The Robustness of Large Language Model IP Protection Methods Against Model Merging. In *1st ACM CCS Workshop on Large AI Systems and Models with Privacy and Safety Analysis (LAMPS)*. ACM, 2024.

[5] Joann Qiongna Chen, **Xinlei He**, Zheng Li, Yang Zhang, and Zhou Li. A Comprehensive Study of Privacy Risks in Curriculum Learning. In *Privacy Enhancing Technologies Symposium (PETS)*, 2025.

[6] **Xinlei He**, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. MGTBench: Benchmarking Machine-Generated Text Detection. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2024.

[7] **Xinlei He**, Savvas Zannettou, Yun Shen, and Yang Zhang. You Only Prompt Once: On the Capabilities of Prompt Learning on Large Language Models to Tackle Toxic Content. In *IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2024.

[8] Tianshuo Cong, **Xinlei He**, Yun Shen, and Yang Zhang. Test-Time Poisoning Attacks Against Test-Time Adaptation Models. In *IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2024.

[9] Boyang Zhang, Zheng Li, Ziqing Yang, **Xinlei He**, Michael Backes, Mario Fritz, and Yang Zhang. SecurityNet: Assessing Machine Learning Vulnerabilities on Public Models. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2024.

[10] 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 *Proceedings on Privacy Enhancing Technologies*, 2024.

[11] Yiting Qu, Xinyue Shen, **Xinlei He**, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-Image

Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2023.

[12] Ziqing Yang, **Xinlei He**, Zheng Li, Michael Backes, Mathias Humbert, Pascal Berrang, and Yang Zhang. Data Poisoning Attacks Against Multimodal Encoders. In *International Conference on Machine Learning (ICML)*. PMLR, 2023.

[13] Yihan Ma, Zhikun Zhang, Ning Yu, **Xinlei He**, Michael Backes, Yun Shen, and Yang Zhang. Generated Graph Detection. In *International Conference on Machine Learning (ICML)*. PMLR, 2023.

[14] Zeyang Sha, **Xinlei He**, Ning Yu, Michael Backes, and Yang Zhang. Can't Steal? Cont-Steal! Contrastive Stealing Attacks Against Image Encoders. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023.

[15] Boyang Zhang, **Xinlei He**, Yun Shen, Tianhao Wang, and Yang Zhang. A Plot is Worth a Thousand Words: Model Information Stealing Attacks via Scientific Plots. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2023.

[16] Yiting Qu, **Xinlei He**, Shannon Pierson, Michael Backes, Yang Zhang, and Savvas Zannettou. On the Evolution of (Hateful) Memes by Means of Multimodal Contrastive Learning. In *IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2023.

[17] **Xinlei He**, Hongbin Liu, Neil Zhenqiang Gong, and Yang Zhang. Semi-Leak: Membership Inference Attacks Against Semi-supervised Learning. In *European Conference on Computer Vision (ECCV)*. Springer, 2022.

[18] Tianshuo Cong, **Xinlei He**, and Yang Zhang. SSLGuard: A Watermarking Scheme for Self-supervised Learning Pre-trained Encoders. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2022.

[19] Zheng Li, Yiyong Liu, **Xinlei He**, Ning Yu, Michael Backes, and Yang Zhang. Auditing Membership Leakages of Multi-Exit Networks. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2022.

[20] Xinyue Shen, **Xinlei He**, Michael Backes, Jeremy Blackburn, Savvas Zannettou, and Yang Zhang. On Xing Tian and the Perseverance of Anti-China Sentiment Online. In *International Conference on Weblogs and Social Media (ICWSM)*, pages 944–955. AAAI, 2022.

[21] Yun Shen*, **Xinlei He***, Yufei Han, and Yang Zhang. Model Stealing Attacks Against Inductive Graph Neural Networks. In *IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2022 (* Equal contribution).

[22] Yugeng Liu, Rui Wen, **Xinlei He**, Ahmed Salem, Zhikun Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang. ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2022.

[23] **Xinlei He**, Jinyuan Jia, Michael Backes, Neil Zhenqiang Gong, and Yang Zhang. Stealing Links from Graph Neural Networks. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2021.

[24] **Xinlei He** and Yang Zhang. Quantifying and Mitigating Privacy Risks of Contrastive Learning. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2021.

Journal

[25] Qinge Xie, Qingyuan Gong, **Xinlei He**, Yang Chen, Xin Wang, Haitao Zheng, and Ben Y. Zhao. Trimming mobile applications for bandwidth-challenged networks in developing regions. *IEEE Transactions on Mobile Computing (TMC)*, 22(1):556–573, 2023.

[26] **Xinlei He**, Qingyuan Gong, Yang Chen, Yang Zhang, Xin Wang, and Xiaoming Fu. Datingsec: Detecting malicious accounts in dating apps using a content-based attention network. *IEEE Transactions on Dependable and Secure Computing (TDSC)*, 18(5):2193–2208, 2021.

[27] Qingyuan Gong, Yang Chen, **Xinlei He**, Yu Xiao, Pan Hui, Xin Wang, and Xiaoming Fu. Cross-site prediction on social influence for cold-start users in online social networks. *ACM Transactions on the Web (TWEB)*, 15(2):6:1–6:23, 2021.

[28] Qingyuan Gong, Yang Chen, **Xinlei He**, Zhou Zhuang, Tianyi Wang, Hong Huang, Xin Wang, and Xiaoming Fu. Deepscan: Exploiting deep learning for malicious account detection in location-based social networks. *IEEE Communications Magazine*, 56(11):21–27, 2018.

Teaching

Lecturer

2025 Spring: Discrete Mathematics (Undergraduate Core Course)
2024 Summer: Machine Learning Security & Privacy

Teaching Assistant

- Advanced Lecture: Attacks Against Machine Learning Models (2023 Summer)
- Seminar: Privacy of Machine Learning (2022 Winter)
- Advanced Lecture: Machine Learning Privacy (2022 Summer)
- Seminar: Data-driven Understanding of the Disinformation Epidemic (2022 Summer)
- Seminar: Privacy of Machine Learning (2021 Winter)
- Advanced Lecture: Privacy Enhancing Technologies (2021 Summer)
- Seminar: Data-driven Understanding of the Disinformation Epidemic (2021 Summer)
- Seminar: Data Privacy (2020 Winter)
- Advanced Lecture: Privacy Enhancing Technologies (2020 Summer)
- Seminar: Data-driven Approaches on Understanding Disinformation (2020 Summer)

Students

Ph.D. Students

Zhen Sun **2024.2 - Now**
Hong Kong University of Science and Technology (Guangzhou)

Yule Liu **2024.2 - Now**
Hong Kong University of Science and Technology (Guangzhou)

Jingyi Zheng **2024.2 - Now**
Hong Kong University of Science and Technology (Guangzhou)

Yuemeng Zhao **2024.9 - Now**
Hong Kong University of Science and Technology (Guangzhou)

Wenhan Dong **2024.9 - Now**
Hong Kong University of Science and Technology (Guangzhou)

Zifan Peng **2024.9 - Now**
Hong Kong University of Science and Technology (Guangzhou)

Master Students

Yiming Huang **2024.9 - Now**
Hong Kong University of Science and Technology (Guangzhou)

Zeren Luo **2024.9 - Now**
Hong Kong University of Science and Technology (Guangzhou)

Chenlong Yin **2024.9 - Now**
Hong Kong University of Science and Technology (Guangzhou)

Ziyi Deng **2024.9 - Now**
Hong Kong University of Science and Technology (Guangzhou)