

Yang Zhang | CV

✉ zhang@cispa.de • 🏠 yangzhangalmo.github.io

last update: December 4, 2024

Employment

CISPA Helmholtz Center for Information Security <i>Tenured Faculty, equivalent to full professor</i>	Saarbrücken, Germany 5/2023 -
CISPA Helmholtz Center for Information Security <i>Tenure-Track Faculty</i>	Saarbrücken, Germany 2/2020 - 4/2023
CISPA Helmholtz Center for Information Security <i>Research Group Leader</i>	Saarbrücken, Germany 1/2019 - 1/2020
CISPA, Saarland University <i>Postdoctoral Researcher</i>	Saarbrücken, Germany 1/2017 - 12/2018

Education

University of Luxembourg <i>Ph.D. in Computer Science, highest honor</i>	Luxembourg, Luxembourg 12/2012 - 11/2016
Shandong University <i>Master in Computer Science</i>	Jinan, China 9/2009 - 6/2012
University of Luxembourg <i>Master in Informatics, exchange student</i>	Luxembourg, Luxembourg 9/2010 - 10/2011
Shandong University <i>Bachelor in Software Engineering</i>	Jinan, China 9/2005 - 6/2009

Research Interests

- Trustworthy Machine Learning (Safety, Privacy, and Security)
- Misinformation, Hate Speech, and Memes
- Social Network Analysis

Service

- PC Member
 - 2025: USENIX Security, NDSS, ICLR (Area Chair), NAACL (Area Chair), WWW, KDD
 - 2024: IEEE S&P, CCS, ICML, NeurIPS, ICLR, ACL (Area Chair), EMNLP (Area Chair), CVPR, ECCV, WWW, KDD, SaTML
 - 2023: IEEE S&P, CCS, NDSS, ICML, NeurIPS, ICLR, WWW, KDD, SaTML
 - 2022: CCS, USENIX Security, NeurIPS, ICLR, WWW, KDD, AAI, PETS, ASIACCS
 - 2021: CCS, USENIX Security, WWW, AAI, Euro S&P, PETS, ASIACCS
 - 2020: CCS, WWW, ICWSM, RAID, PETS
 - 2019: CCS, ISMB/ECCB

- Editorial Board
 - ACM Transactions on Privacy and Security (TOPS)
- PC Chair
 - International Workshop on Ethics in Computer Security (EthiCS 2022-2023)
- Organizer
 - Privacy and Security in ML Seminars
- Ph.D. Thesis Committee
 - Salijona Dyrnishi, University of Luxembourg, 2024
 - Hailong Hu, University of Luxembourg, 2024
 - Bang Wu, Monash University, 2024
 - Sinem Sav, EPFL, 2023
 - Inken Hagestedt, Saarland University, 2021
 - Benjamin Zhao, University of New South Wales, 2021

Awards

- Best paper finalist at CSAW Europe 2024
- Best paper finalist at CSAW Europe 2023
- Best paper award honorable mention at CCS 2022
- Busy Beaver teaching award nomination for seminar "Privacy of Machine Learning" at Saarland University (2022 Winter)
- Busy Beaver teaching award nomination for advanced lecture "Machine Learning Privacy" at Saarland University (2022 Summer)
- Busy Beaver teaching award for seminar "Privacy of Machine Learning" at Saarland University (2021 Winter)
- Distinguished reviewer award at TrustML Workshop 2020 (co-located with ICLR 2020)
- Distinguished paper award at NDSS 2019
- Best paper award at ARES 2014

Open-Sourced Software

Our lab is fully committed to open science, which led to the establishment of TrustAIRLab.

- Much of the code developed by our lab is accessible through our GitHub organization
- A curated selection of datasets collected by our lab can be found on our Hugging Face organization

Publication

My publication list can also be found at DBLP and Google Scholar; however, they may not be up to date.

Conference

- [1] Rui Wen and Michael Backes and **Yang Zhang**. Understanding Data Importance in Machine Learning Attacks: Does Valuable Data Pose Greater Harm? In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2025.
- [2] Yihan Ma and Xinyue Shen and Yixin Wu and Boyang Zhang and Michael Backes and **Yang Zhang**. The Death and Life of Great Prompts: Analyzing the Evolution of LLM Prompts from the Structural Perspective. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2024.
- [3] Yukun Jiang and Zheng Li and Xinyue Shen and Yugeng Liu and Michael Backes and **Yang Zhang**.

- ModScan: Measuring Stereotypical Bias in Large Vision-Language Models from Vision and Language Modalities. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2024.
- [4] Junjie Chu and Zeyang Sha and Michael Backes and **Yang Zhang**. Reconstruct Your Previous Conversations! Comprehensively Investigating Privacy Leakage Risks in Conversations with GPT Models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2024.
- [5] Rui Wen and Zheng Li and Michael Backes and **Yang Zhang**. Membership Inference Attacks Against In-Context Learning. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2024.
- [6] Yixin Wu and Yun Shen and Michael Backes and **Yang Zhang**. Image-Perfect Imperfections: Safety, Bias, and Authenticity in the Shadow of Text-To-Image Model Evolution. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2024.
- [7] Jinghuai Zhang and Jianfeng Chi and Zheng Li and Kunlin Cai and **Yang Zhang** and Yuan Tian. BadMerging: Backdoor Attacks Against Model Merging. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2024.
- [8] Zeyang Sha and Yicong Tan and Mingle Li and Michael Backes and **Yang Zhang**. ZeroFake: Zero-Shot Detection of Fake Images Generated and Edited by Text-to-Image Generation Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2024.
- [9] Hao Li and Zheng Li and Siyuan Wu and Chengrui Hu and Yutong Ye and Min Zhang and Dengguo Feng and **Yang Zhang**. SeqMIA: Sequential-Metric Based Membership Inference Attack. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2024.
- [10] Xinlei He and Xinyue Shen and Zeyuan Chen and Michael Backes and **Yang Zhang**. MGTBench: Benchmarking Machine-Generated Text Detection. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2024.
- [11] Xinyue Shen and Zeyuan Chen and Michael Backes and Yun Shen and **Yang Zhang**. “Do Anything Now”: Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2024.
- [12] Rui Zhang and Hongwei Li and Rui Wen and Wenbo Jiang and Yuan Zhang and Michael Backes and Yun Shen and **Yang Zhang**. Instruction Backdoor Attacks Against Customized LLMs. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2024.
- [13] Xinyue Shen and Yiting Qu and Michael Backes and **Yang Zhang**. Prompt Stealing Attacks Against Text-to-Image Generation Models. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2024.
- [14] Boyang Zhang and Zheng Li and Ziqing Yang and Xinlei He and Michael Backes and Mario Fritz and **Yang Zhang**. SecurityNet: Assessing Machine Learning Vulnerabilities on Public Models. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2024.
- [15] Yixin Wu and Rui Wen and Michael Backes and Pascal Berrang and Mathias Humbert and Yun Shen and **Yang Zhang**. Quantifying Privacy Risks of Prompts in Visual Prompt Learning. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2024.
- [16] Hai Huang and Zhengyu Zhao and Michael Backes and Yun Shen and **Yang Zhang**. Composite Backdoor Attacks Against Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL (NAACL Findings)*. ACL, 2024.
- [17] Yukun Jiang and Xinyue Shen and Rui Wen and Zeyang Sha and Junjie Chu and Yugeng Liu and Michael Backes and **Yang Zhang**. Games and Beyond: Analyzing the Bullet Chats of Esports Livestreaming. In *International Conference on Weblogs and Social Media (ICWSM)*. AAAI, 2024.
- [18] Yiting Qu and Zhikun Zhang and Yun Shen and Michael Backes and **Yang Zhang**. FAKEPCD: Fake Point Cloud Detection via Source Attribution. In *ACM Asia Conference on Computer and Communications Security (ASIACCS)*, pages 930–946. ACM, 2024.

- [19] Xinlei He and Savvas Zannettou and 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.
- [20] Tianshuo Cong and Xinlei He and 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.
- [21] Minxing Zhang and Ning Yu and Rui Wen and Michael Backes and **Yang Zhang**. Generated Distributions Are All You Need for Membership Inference Attacks Against Generative Models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2024.
- [22] Zeyang Sha and Zheng Li and Ning Yu and **Yang Zhang**. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 3418–3432. ACM, 2023.
- [23] Yiting Qu and Xinyue Shen and Xinlei He and Michael Backes and 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)*, pages 3403–3417. ACM, 2023.
- [24] Joann Qiongna Chen and Tianhao Wang and Zhikun Zhang and **Yang Zhang** and Somesh Jha and Zhou Li. Differentially Private Resource Allocation. In *Annual Computer Security Applications Conference (ACSAC)*, pages 772–786. ACM, 2023.
- [25] Boyang Zhang and Xinlei He and Yun Shen and 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)*, pages 5289–5306. USENIX, 2023.
- [26] Wai Man Si and Michael Backes and **Yang Zhang** and Ahmed Salem. Two-in-One: A Model Hijacking Attack Against Text Generation Models. In *USENIX Security Symposium (USENIX Security)*, pages 2223–2240. USENIX, 2023.
- [27] Zheng Li and Ning Yu and Ahmed Salem and Michael Backes and Mario Fritz and **Yang Zhang**. UnGANable: Defending Against GAN-based Face Manipulation. In *USENIX Security Symposium (USENIX Security)*, pages 7213–7230. USENIX, 2023.
- [28] Min Chen and Zhikun Zhang and Michael Backes and Tianhao Wang and **Yang Zhang**. FACE-AUDITOR: Data Auditing in Facial Recognition Systems. In *USENIX Security Symposium (USENIX Security)*, pages 7195–7212. USENIX, 2023.
- [29] Haiming Wang and Zhikun Zhang and Tianhao Wang and Shibo He and Michael Backes and Jiming Chen and **Yang Zhang**. PrivTrace: Differentially Private Trajectory Synthesis by Adaptive Markov Model. In *USENIX Security Symposium (USENIX Security)*, pages 1649–1666. USENIX, 2023.
- [30] Yihan Ma and Zhikun Zhang and Ning Yu and Xinlei He and Michael Backes and Yun Shen and **Yang Zhang**. Generated Graph Detection. In *International Conference on Machine Learning (ICML)*, pages 23412–23428. PMLR, 2023.
- [31] Ziqing Yang and Xinlei He and Zheng Li and Michael Backes and Mathias Humbert and Pascal Berrang and **Yang Zhang**. Data Poisoning Attacks Against Multimodal Encoders. In *International Conference on Machine Learning (ICML)*, pages 39299–39313. PMLR, 2023.
- [32] Kai Mei and Zheng Li and Zhenting Wang and **Yang Zhang** and Shiqing Ma. NOTABLE: Transferable Backdoor Attacks Against Prompt-based NLP Models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15551–15565. ACL, 2023.
- [33] Zeyang Sha and Xinlei He and Ning Yu and 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)*, pages 16373–16383. IEEE, 2023.

- [34] Yiting Qu and Xinlei He and Shannon Pierson and Michael Backes and **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)*, pages 293–310. IEEE, 2023.
- [35] Rui Wen and Zhengyu Zhao and Zhuoran Liu and Michael Backes and Tianhao Wang and **Yang Zhang**. Is Adversarial Training Really a Silver Bullet for Mitigating Data Poisoning? In *International Conference on Learning Representations (ICLR)*, 2023.
- [36] Yugeng Liu and Zheng Li and Michael Backes and Yun Shen and **Yang Zhang**. Backdoor Attacks Against Dataset Distillation. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2023.
- [37] Xiaojian Yuan and Kejiang Chen and Jie Zhang and Weiming Zhang and Nenghai Yu and **Yang Zhang**. Pseudo Label-Guided Model Inversion Attack via Conditional Generative Adversarial Network. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 3349–3357. AAAI, 2023.
- [38] Yufei Chen and Chao Shen and Yun Shen and Cong Wang and **Yang Zhang**. Amplifying Membership Exposure via Data Poisoning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2022.
- [39] Wai Man Si and Michael Backes and Jeremy Blackburn and Emiliano De Cristofaro and Gianluca Stringhini and Savvas Zannettou and **Yang Zhang**. Why So Toxic?: Measuring and Triggering Toxic Behavior in Open-Domain Chatbots. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 2659–2673. ACM, 2022.
- [40] Hai Huang and Zhikun Zhang and Yun Shen and Michael Backes and Qi Li and **Yang Zhang**. On the Privacy Risks of Cell-Based NAS Architectures. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1427–1441. ACM, 2022.
- [41] Yiyong Liu and Zhengyu Zhao and Michael Backes and **Yang Zhang**. Membership Inference Attacks by Exploiting Loss Trajectory. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 2085–2098. ACM, 2022.
- [42] Zheng Li and Yiyong Liu and Xinlei He and Ning Yu and Michael Backes and **Yang Zhang**. Auditing Membership Leakages of Multi-Exit Networks. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1917–1931. ACM, 2022.
- [43] Min Chen and Zhikun Zhang and Tianhao Wang and Michael Backes and Mathias Humbert and **Yang Zhang**. Graph Unlearning. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 499–513. ACM, 2022.
- [44] Tianshuo Cong and 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)*, pages 579–593. ACM, 2022.
- [45] Yun Shen and Yufei Han and Zhikun Zhang and Min Chen and Ting Yu and Michael Backes and **Yang Zhang** and Gianluca Stringhini. Finding MNEMON: Reviving Memories of Node Embeddings. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 2643–2657. ACM, 2022.
- [46] Xinlei He and Hongbin Liu and Neil Zhenqiang Gong and **Yang Zhang**. Semi-Leak: Membership Inference Attacks Against Semi-supervised Learning. In *European Conference on Computer Vision (ECCV)*, pages 365–381. Springer, 2022.
- [47] Yugeng Liu and Rui Wen and Xinlei He and Ahmed Salem and Zhikun Zhang and Michael Backes and Emiliano De Cristofaro and Mario Fritz and **Yang Zhang**. ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models. In *USENIX Security Symposium (USENIX Security)*, pages 4525–4542. USENIX, 2022.
- [48] Yufei Chen and Chao Shen and Cong Wang and **Yang Zhang**. Teacher Model Fingerprinting Attacks Against Transfer Learning. In *USENIX Security Symposium (USENIX Security)*, pages 3593–3610. USENIX, 2022.

- [49] Zhikun Zhang and Min Chen and Michael Backes and Yun Shen and **Yang Zhang**. Inference Attacks Against Graph Neural Networks. In *USENIX Security Symposium (USENIX Security)*, pages 4543–4560. USENIX, 2022.
- [50] Xinyue Shen and Xinlei He and Michael Backes and Jeremy Blackburn and 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.
- [51] Yun Shen and Xinlei He and Yufei Han and **Yang Zhang**. Model Stealing Attacks Against Inductive Graph Neural Networks. In *IEEE Symposium on Security and Privacy (S&P)*, pages 1175–1192. IEEE, 2022.
- [52] Ahmed Salem and Rui Wen and Michael Backes and Shiqing Ma and **Yang Zhang**. Dynamic Backdoor Attacks Against Machine Learning Models. In *IEEE European Symposium on Security and Privacy (Euro S&P)*, pages 703–718. IEEE, 2022.
- [53] Ahmed Salem and Michael Backes and **Yang Zhang**. Get a Model! Model Hijacking Attack Against Machine Learning Models. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2022.
- [54] Junhao Zhou and Yufei Chen and Chao Shen and **Yang Zhang**. Property Inference Attacks Against GANs. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2022.
- [55] Xinlei He and **Yang Zhang**. Quantifying and Mitigating Privacy Risks of Contrastive Learning. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 845–863. ACM, 2021.
- [56] Min Chen and Zhikun Zhang and Tianhao Wang and Michael Backes and Mathias Humbert and **Yang Zhang**. When Machine Unlearning Jeopardizes Privacy. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 896–911. ACM, 2021.
- [57] Minxing Zhang and Zhaochun Ren and Zihan Wang and Pengjie Ren and Zhumin Chen and Pengfei Hu and **Yang Zhang**. Membership Inference Attacks Against Recommender Systems. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 864–879. ACM, 2021.
- [58] Zheng Li and **Yang Zhang**. Membership Leakage in Label-Only Exposures. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 880–895. ACM, 2021.
- [59] Xiaoyi Chen and Ahmed Salem and Michael Backes and Shiqing Ma and Qingni Shen and Zhonghai Wu and **Yang Zhang**. BadNL: Backdoor Attacks Against NLP Models with Semantic-preserving Improvements. In *Annual Computer Security Applications Conference (ACSAC)*, pages 554–569. ACSAC, 2021.
- [60] Xinlei He and Jinyuan Jia and Michael Backes and Neil Zhenqiang Gong and **Yang Zhang**. Stealing Links from Graph Neural Networks. In *USENIX Security Symposium (USENIX Security)*, pages 2669–2686. USENIX, 2021.
- [61] Zhikun Zhang and Tianhao Wang and Jean Honorio and Ninghui Li and Michael Backes and Shibo He and Jiming Chen and **Yang Zhang**. PrivSyn: Differentially Private Data Synthesis. In *USENIX Security Symposium (USENIX Security)*, pages 929–946. USENIX, 2021.
- [62] Fatemeh Tahmasbi and Leonard Schild and Chen Ling and Jeremy Blackburn and Gianluca Stringhini and **Yang Zhang** and Savvas Zannettou. “Go eat a bat, Chang!”: On the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19. In *The Web Conference (WWW)*. ACM, 2021.
- [63] Rui Wen and Yu Yu and Xiang Xie and **Yang Zhang**. LEAF: A Faster Secure Search Algorithm via Localization, Extraction, and Reconstruction. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1219–1232. ACM, 2020.
- [64] Dingfan Chen and Ning Yu and **Yang Zhang** and Mario Fritz. GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 343–362. ACM, 2020.

- [65] Ahmed Salem and Apratim Bhattacharya and Michael Backes and Mario Fritz and **Yang Zhang**. Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning. In *USENIX Security Symposium (USENIX Security)*, pages 1291–1308. USENIX, 2020.
- [66] Inken Hagestedt and Mathias Humbert and Pascal Berrang and Irina Lehmann and Roland Eils and Michael Backes and **Yang Zhang**. Membership Inference Against DNA Methylation Databases. In *IEEE European Symposium on Security and Privacy (Euro S&P)*, pages 509–520. IEEE, 2020.
- [67] **Yang Zhang** and Mathias Humbert and Bartłomiej Surma and Praveen Manoharan and Jilles Vreeken and Michael Backes. Towards Plausible Graph Anonymization. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2020.
- [68] Jinyuan Jia and Ahmed Salem and Michael Backes and **Yang Zhang** and Neil Zhenqiang Gong. MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 259–274. ACM, 2019.
- [69] Zheng Li and Chengyu Hu and **Yang Zhang** and Shanqing Guo. How to Prove Your Model Belongs to You: A Blind-Watermark based Framework to Protect Intellectual Property of DNN. In *Annual Computer Security Applications Conference (ACSAC)*, pages 126–137. ACSAC, 2019.
- [70] Zhiqiang Zhong and **Yang Zhang** and Jun Pang. A Graph-Based Approach to Explore Relationship Between Hashtags and Images. In *International Conference Web Information Systems Engineering (WISE)*, pages 473–488. Springer, 2019.
- [71] Tahleen Rahman and Bartłomiej Surma and Michael Backes and **Yang Zhang**. Fairwalk: Towards Fair Graph Embedding. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 3289–3295. IJCAI, 2019.
- [72] **Yang Zhang**. Language in Our Time: An Empirical Analysis of Hashtags. In *The Web Conference (WWW)*, pages 2378–2389. ACM, 2019.
- [73] Ahmed Salem and **Yang Zhang** and Mathias Humbert and Pascal Berrang and Mario Fritz and Michael Backes. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2019.
- [74] Inken Hagestedt and **Yang Zhang** and Mathias Humbert and Pascal Berrang and Haixu Tang and XiaoFeng Wang and Michael Backes. MBeacon: Privacy-Preserving Beacons for DNA Methylation Data. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2019.
- [75] Fanghua Zhao and Linan Gao and **Yang Zhang** and Zeyu Wang and Bo Wang and Shanqing Guo. You Are Where You App: An Assessment on Location Privacy of Social Applications. In *International Symposium on Software Reliability Engineering (ISSRE)*, pages 236–247. IEEE, 2018.
- [76] **Yang Zhang** and Mathias Humbert and Tahleen Rahman and Cheng-Te Li and Jun Pang and Michael Backes. Tagvisor: A Privacy Advisor for Sharing Hashtags. In *The Web Conference (WWW)*, pages 287–296. ACM, 2018.
- [77] Pascal Berrang and Mathias Humbert and **Yang Zhang** and Irina Lehmann and Roland Eils and Michael Backes. Dissecting Privacy Risks in Biomedical Data. In *IEEE European Symposium on Security and Privacy (Euro S&P)*, pages 62–76. IEEE, 2018.
- [78] Michael Backes and Mathias Humbert and Jun Pang and **Yang Zhang**. walk2friends: Inferring Social Links from Mobility Profiles. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1943–1957. ACM, 2017.
- [79] Jun Pang and **Yang Zhang**. Quantifying Location Sociality. In *ACM Conference on Hypertext and Social Media (HT)*, pages 145–154. ACM, 2017.
- [80] Jun Pang and **Yang Zhang**. DeepCity: A Feature Learning Framework for Mining Location Check-Ins. In *International Conference on Weblogs and Social Media (ICWSM)*, pages 652–655. AAAI, 2017.

- [81] Yan Wang and Zongxu Qin and Jun Pang and **Yang Zhang** and Xin Jin. Semantic Annotation for Places in LBSN Using Graph Embedding. In *ACM International Conference on Information and Knowledge Management (CIKM)*, page 2343–2346. ACM, 2017.
- [82] **Yang Zhang** and Minyue Ni and Weili Han and Jun Pang. Does #like4like Indeed Provoke More Likes? In *International Conference on Web Intelligence (WI)*, pages 179–186. ACM, 2017.
- [83] Minyue Ni and **Yang Zhang** and Weili Han and Jun Pang. An Empirical Study on User Access Control in Online Social Networks. In *ACM Symposium on Access Control Models and Technologies (SACMAT)*, pages 12–23. ACM, 2016.
- [84] Jun Pang and Polina Zablotskaia and **Yang Zhang**. On Impact of Weather on Human Mobility in Cities. In *International Conference Web Information Systems Engineering (WISE)*, pages 247–256. Springer, 2016.
- [85] Jun Pang and **Yang Zhang**. Location Prediction: Communities Speak Louder than Friends. In *ACM Conference on Online Social Networks (COSN)*, pages 161–171. ACM, 2015.
- [86] **Yang Zhang** and Jun Pang. Distance and Friendship: A Distance-based Model for Link Prediction in Social Networks. In *Asia-Pacific Web Conference (APWeb)*, pages 55–66. Springer, 2015.
- [87] Jun Pang and **Yang Zhang**. Event Prediction with Community Leaders. In *Conference on Availability, Reliability and Security (ARES)*, pages 238–243. IEEE, 2015.
- [88] Marcos Cramer and Jun Pang and **Yang Zhang**. A Logical Approach to Restricting Access in Online Social Networks. In *ACM Symposium on Access Control Models and Technologies (SACMAT)*, pages 75–86. ACM, 2015.
- [89] Jun Pang and **Yang Zhang**. Cryptographic Protocols for Enforcing Relationship-based Access Control Policies. In *Annual IEEE Computers, Software and Applications Conference (COMPSAC)*, pages 484–493. IEEE, 2015.
- [90] Jun Pang and **Yang Zhang**. Exploring Communities for Effective Location Prediction. In *International Conference on World Wide Web (WWW)*, pages 87–88. ACM, 2015.
- [91] **Yang Zhang** and Jun Pang. Community-driven Social Influence Analysis and Applications. In *International Conference on Web Engineering (ICWE)*. Springer, 2015.
- [92] Jun Pang and **Yang Zhang**. A New Access Control Scheme for Facebook-style Social Networks. In *Conference on Availability, Reliability and Security (ARES)*, pages 1–10. IEEE, 2014.
- Journal**.....
- [93] Yan Pang and Tianhao Wang and Xuhui Kang and Mengdi Huai and **Yang Zhang**. White-box Membership Inference Attacks against Diffusion Models. *Proceedings on Privacy Enhancing Technologies*, 2025.
- [94] Joann Qionгна Chen and Xinlei He and Zheng Li and **Yang Zhang** and Zhou Li. A Comprehensive Study of Privacy Risks in Curriculum Learning. *Proceedings on Privacy Enhancing Technologies*, 2025.
- [95] Xiaokuan Zhang and **Yang Zhang** and Yinqian Zhang. VERITRAIN: Validating MLaaS Training Efforts via Anomaly Detection. *IEEE Transactions on Dependable and Secure Computing*, 2024.
- [96] Yixin Wu and Xinlei He and Pascal Berrang and Mathias Humbert and Michael Backes and Neil Zhenqiang Gong and **Yang Zhang**. Link Stealing Attacks Against Inductive Graph Neural Networks. *Proceedings on Privacy Enhancing Technologies*, 2024.
- [97] Cheng-Te Li and Cheng Hsu and **Yang Zhang**. FairSR: Fairness-aware Sequential Recommendation through Multi-Task Learning with Preference Graph Embeddings. *ACM Transactions on Intelligent Systems and Technology*, 2022.
- [98] Xinlei He and Qingyuan Gong and Yang Chen and **Yang Zhang** and 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*, 2021.

[99] Bo-Heng Chen and Cheng-Te Li and Kun-Ta Chuang and Jun Pang and **Yang Zhang**. An Active Learning-based Approach for Location-aware Acquaintance Inference. *Knowledge and Information Systems*, 2018.

[100] Jun Pang and **Yang Zhang**. A New Access Control Scheme for Facebook-style Social Networks. *Computers & Security*, 2015.

Teaching

- 2024 Winter, Seminar: Privacy of Machine Learning
- 2024 Summer, Advanced Lecture: Attacks Against Machine Learning Models
- 2024 Summer, Seminar: Data-driven Understanding of the Disinformation Epidemic
- 2023 Winter, Seminar: Privacy of Machine Learning
- 2023 Summer, Advanced Lecture: Attacks Against Machine Learning Models
- 2023 Summer, Seminar: Data-driven Understanding of the Disinformation Epidemic
- 2022 Winter, Seminar: Privacy of Machine Learning
- 2022 Summer, Advanced Lecture: Machine Learning Privacy
- 2022 Summer, Seminar: Data-driven Understanding of the Disinformation Epidemic
- 2021 Winter, Seminar: Privacy of Machine Learning
- 2021 Summer, Advanced Lecture: Privacy Enhancing Technologies
- 2021 Summer, Seminar: Data-driven Understanding of the Disinformation Epidemic
- 2020 Winter, Seminar: Data Privacy
- 2020 Summer, Advanced Lecture: Privacy Enhancing Technologies
- 2020 Summer, Seminar: Data-driven Approaches on Understanding Disinformation
- 2019 Winter, Seminar: Data Privacy
- 2019 Summer, Advanced Lecture: Privacy Enhancing Technologies
- 2019 Summer, Seminar: Biomedical Privacy
- 2018 Winter, Seminar: Data Privacy
- 2018 Summer, Advanced Lecture: Privacy Enhancing Technologies
- 2018 Summer, Seminar: Adversarial Machine Learning

Students

Postdoc

Zheng Li (*ERCIM WG STM Best Ph.D. Thesis Award 2024*) 11/2023 -

Mingjie Li 10/2023 -

Ph.D. Students

Ye Leng 10/2024 -

Yukun Jiang 6/2024 -

Chi Cui 4/2024 -

Yicong Tan	4/2024 -
Zeyang Sha	3/2023 -
Junjie Chu	11/2022 -
Yixin Wu	11/2022 -
Ziqing Yang	11/2022 -
Xinyue Shen (<i>Abbe Grant 2024</i>)	10/2022 -
Yugeng Liu	1/2022 -
Boyang Zhang	12/2021 -
Hai Huang	11/2021 -
Yiting Qu	11/2021 -
Wai Man Si	11/2021 -
Rui Wen	10/2021 -
Yihan Ma	7/2021 -
Visiting Ph.D. Students	
Chia-Yi Hsu (<i>National Yang Ming Chiao Tung University</i>)	12/2024 -
Yuke Hu (<i>Zhejiang University</i>)	5/2024 -
Ph.D. Preparatory Phase	
Xinyu Zhang	10/2024 -
Zeyuan Chen	10/2023 -
Alumni	
Xinlei He (<i>Norton Labs Graduate Fellowship 2022</i>) <i>now assistant professor at HKUST (GZ)</i>	02/2020 - 9/2023

Tianshuo Cong (*visiting Ph.D. student from Tsinghua University*)
now Shui Mu Scholar at Tsinghua University

08/2021 - 12/2022

Ahmed Salem

2/2017 - 1/2022

now senior researcher at Microsoft Security Response Center

Talks

Keynote.....

- 2024, Australasian Conference on Information Security and Privacy (ACISP)
- 2024, ACNS Workshop on Security in Machine Learning and its Applications (SiMLA)
- 2023, Information Security Conference (ISC)
- 2023, The AsiaCCS Workshop on Secure and Trustworthy Deep Learning Systems
- 2023, Backdoor Attacks and Defenses in Machine Learning (BANDS)
- 2022, PAIR2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data

Seminar and School.....

- 2024, Talk at Nanyang Technological University
- 2023, Lecturer at Summer School on Privacy-Preserving Machine Learning
- 2023, Talk at EPFL
- 2022, Lecturer at Summer School on Privacy-Preserving Machine Learning
- 2022, Distinguished Lecture in ViSP (Vienna Cybersecurity and Privacy Research Center) Distinguished Lecture Series
- 2021, Vector Visitor Talk at Vector Institute
- 2021, Talk at Privacy and Security in ML Seminars
- 2021, Talk at Inria
- 2020, Talk at University College London

In the Press

- 8/2023, Tricks for making AI chatbots break rules are freely available online, *New Scientist*
- 8/2023, Wie Chatbots die eigenen Regeln vergessen, *Deutschlandfunk Nova*
- 12/2022, The internet loves ChatGPT, but there's a dark side to the tech, *Fast Company*
- 4/2020, As the coronavirus spreads, so does online racism targeting Asians, new research shows, *The Washington Post*