Xudong Pan

Affiliation	Fudan University	**Mobile Phone**	+86 137 7611 8281
Research Focus	AI Security	**Email**	xdpan18@fudan.edu.cn
Graduation Time	June, 2023	**Address**	A6008, Interdisciplinary Bld No.2
Language Skills	Chinese (native), English (GRE 321)		2005 Songhu Rd, Yangpu Dist.
	Japanese (JLPT N1, proficient)		Shanghai, 200082, China

Education

■ Ph.D. in Computer Science - Fudan University *09/2018 - 06/2023*

- *Thesis*: *"Neuron State Based Risk Analysis on Data Reconstruction from Deep Learning Systems and the Countermeasures"*
- *Advisor*: Prof. Min Yang (Changjiang Distinguished Professor, Dean of School of Computer Science)
- **Academic Star Award** (Top 10 in Fudan University, across all STEM graduate schools)
- **National Scholarships** (Awarded by the Ministry of Education, **twice**)
- **Youth Outstanding Paper Nomination**, World Artificial Intelligence Conference (WAIC)

(10 selected from near 400 excellent papers in the past five years, authored by famous university and industrial institutes, e.g., *Stanford, Cambridge, Oxford, MSRA*, and published at top-tier venues, e.g., *Nature and Science*)

■ B.S. in Software Engineering - Fudan University *09/2014 - 06/2018*

- *GPA* (CS-Related): 3.87/4.0 (**Top** 1% of the school)
- Independent research published at top-tier AI conference ICML'18

Research Experience

Sept. 2018- Present Research Assistant, System Software and Security Lab
School of Computer Science, Fudan University (Mentors: Prof. Min Yang, Prof. Mi Zhang)

During my doctoral research, I mainly work on the interdisciplinary area of AI and security. I **discover, analyze and cure** security flaws of the cutting-edge AI algorithms when they are deployed in real-world intelligent systems in the wild, including but not limited to face recognition, toxic language detection, intelligent healthcare, and autonomous driving.

Research Outputs: This work resulted in 17 published papers at the top-tier AI and security conference/journals including **TPAMI** (2020, IF=24.314), **TKDE** (2021, IF=9.235), **NeurIPS** (2022), **KDD** (2019, 2022), **AAAI** (2020), **IEEE S&P** (2020, Youth Outstanding Paper Nomination at WAIC) and **USENIX Security** (2020, 2022×2) as the primary researcher. I held 4 invention patents on AI security (1 authorized, 3 pending). Besides, I led a team to **win the 1st place in AutoDriving CTF** at DEFCON 29 and DEFCON 30.

Sept. 2015- Jun. 2018 Undergraduate Researcher, System Software and Security Lab
School of Computer Science, Fudan University (Mentors: Prof. Min Yang, Prof. Mi Zhang)

In 2015-2016, I studied the fundamentals of machine learning, deep learning, statistical learning theory, principles of system security by attending lab seminars, discussion with my advisors and self-teaching. I also learned differential and algebraic geometry from courses in the physics department. In 2017-2018, I conducted research on adversarial machine learning, especially the geometric theory behind DeepFake, and on building recommender systems via deep learning techniques.

Research Outputs: This work resulted in one published paper at the commonly-recognized top-tier AI conference **ICML** (2018) and **WWW** (2018) as the primary researcher. In my ICML paper, a Riemannian geometric theory is developed to explain the enabler of the DeepFake technique, which I later expanded to be accepted at **TPAMI** (2020), where topological invariants on data manifolds are used to determine the existence of optimal image generators. In 2022, researchers at Tencent, ByteDance and Kuaishou successfully applied my theory for better sequence generation.

Publications

Summary

- Four in commonly-recognized top-tier security conferences (IEEE Security & Privacy and USENIX Security, **all first-authored**)
- Eight in commonly-recognized top-tier artificial intelligence conferences/journals (TPAMI, TKDE, NeurIPS, ICML, KDD, AAAI, WWW)
- One in commonly-recognized top-tier database conferences (ICDE)

Conference Publications

(top-tier conferences are highlighted in bold)

C15. [**NeurIPS'22**] **Xudong Pan**, Shengyao Zhang, Mi Zhang, Yifan Yan, Min Yang. House of Cans: Covert Transmission of Internal Datasets via Capacity-Aware Neuron Steganography, the 36th Annual Conference on Neural Information Processing Systems (NeurIPS), 2022.

C14. [**KDD'22**] **Xudong Pan**, Yifan Yan, Mi Zhang, Min Yang. MetaV: A Meta-Verifier Approach to Task-Agnostic Model Fingerprinting, the 28th SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), P1327–1336, 2022.

C13. [**USENIX Security'22**] **Xudong Pan**, Mi Zhang, Beina Sheng, Jiaming Zhu, Min Yang. Hidden Trigger Backdoor Attack on NLP Models via Linguistic Style Manipulation, the 31st USENIX Security Symposium (USENIX Security), P3611-3628, 2022.

C12. [**ICDE'22**] Daizong Ding, Mi Zhang, Yuanmin Huang, **Xudong Pan**, Fuli Feng, Erling Jiang, Min Yang. Towards Backdoor Attack on Deep Learning based Time Series Classification, the 38th IEEE International Conference on Data Engineering (ICDE), P1274-1287, 2022.

C11. [**USENIX Security'22**] **Xudong Pan**, Mi Zhang, Yifan Yan, Jiaming Zhu, Min Yang. Exploring the Security Boundary of Data Reconstruction via Neuron Exclusivity Analysis, the 31st USENIX Security Symposium (USENIX Security), P3989-4006, 2022.

C10. [ACSAC'21] **Xudong Pan**, Mi Zhang, Yifan Yan, Min Yang. Understanding the Threats of Trojaned Quantized Neural Network in Model Supply Chains, the 38th Annual Computer Security Applications Conference (ACSAC), P634–645, 2021.

C9. [CIKM'21] Daizong Ding, Mi Zhang, Hanrui Wang, **Xudong Pan**, Min Yang, Xiangnan He. A Deep Learning Framework for Self-evolving Hierarchical Community Detection, the 30th ACM International Conference on Information and Knowledge Management (CIKM), P372–381, 2021.

C8. [ESORICS'21] **Xudong Pan**, Mi Zhang, Yifan Lu, Min Yang. TAFA: A Task-Agnostic Fingerprinting Algorithm for Neural Networks, the 26th European Symposium on Research in Computer Security (ESORICS), P542-562, 2021.

C7. [ICDM'20] Daizong Ding, Mi Zhang, **Xudong Pan**, Min Yang, Xiangnan He. Modeling Personalized Out-of-Town Distances in Location Recommendation, the 20th IEEE International Conference on Data Mining (ICDM), P112-121, 2020.

C6. [**USENIX Security'20**] **Xudong Pan**, Mi Zhang, Duocai Wu, Qifan Xiao, Min Yang. Justinian's GAAvernor: Robust Distributed Learning with Gradient Aggregation Agent, the 29th USENIX Security Symposium (USENIX Security), P1641-1658, 2020.

C5. [**S&P'20**] **Xudong Pan**, Mi Zhang, Shouling Ji, Min Yang. Privacy Risks of General-Purpose Language Models, the 2020 IEEE Symposium on Security and Privacy (S&P), P1471-1488, 2020. **[Youth Outstanding Paper Nomination, WAIC]**

C4. [**AAAI'20**] Daizong Ding, Mi Zhang, **Xudong Pan**, Min Yang, Xiangnan He. Improving the Robustness of Wasserstein Embedding by Adversarial PAC-Bayesian Learning, the 34th AAAI Conference on Artificial Intelligence (AAAI), P3791-3800, 2020.

C3. [**KDD'19**] Daizong Ding, Mi Zhang, **Xudong Pan**, Min Yang, Xiangnan He. Modeling Extreme Events in Time Series Prediction, the 25th SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), P1114–1122, 2019.

C2. [**ICML'18**] **Xudong Pan**, Mi Zhang, Min Yang. Theoretical Analysis of Image-to-Image Translation with Adversarial Learning, the 35th International Conference on Machine Learning (ICML), P4006-4015, 2018.

C1. [**WWW'18**] Daizong Ding, Mi Zhang, **Xudong Pan**, Pearl Pu. Geographical Feature Extraction for Entities in Location-based Social Networks, the 2018 World Wide Web Conference (**WWW**), P833-842, 2018.

Journal Publications

(top-tier conferences are highlighted in bold)

J4. [JCRD] **Xudong Pan**, Mi Zhang, Min Yang. Fishing Leakage of Deep Learning Training Data via Neuron Activation Pattern Manipulation, Journal of Computer Research and Development (in Chinese), 59(10) P2323-2337, 2022. (*Indexed by EI, Impact Factor 3.363*)

J3. [**TKDE**] Mi Zhang, Daizong Ding, **Xudong Pan**, Min Yang. Enhancing Time Series Predictors with Generalized Extreme Value Loss, IEEE Transactions on Knowledge and Data Engineering (TKDE), Early Access, 2021. *(Indexed by SCI, Impact Factor 9.235)*

J2. [JCRD] **Xudong Pan**, Mi Zhang, Yifan Yan, Yifan Lu, Min Yang. Evaluating Privacy Risks of Deep Learning Based General-Purpose Language Models, Journal of Computer Research and Development (in Chinese), 58(5), P1092-1105, 2021. (*Indexed by EI, Impact Factor 3.363*)

J1. [**TPAMI**] **Xudong Pan**, Mi Zhang, Daizong Ding, Min Yang. A Geometrical Perspective on Image Style Transfer with Adversarial Learning, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 44(1), P63-75, 2020. (*Indexed by SCI, Impact Factor 24.314*)

Submitted Manuscripts

(students mentored by me are underlined)

- **Xudong Pan**, Mi Zhang, Shengyao Zhang, Min Yang. Exploiting the Over-Parametrization of Deep Learning Models for Covert Data Transmission, under review at IEEE Transactions on Pattern Analysis and Machine Intelligence.
- **Xudong Pan**, Mi Zhang, Beina Sheng, Min Yang. TAFA-X: Towards General-Purpose Task-Agnostic Fingerprinting of Deep Neural Networks, under review at IEEE Transactions on Information Forensics and Security.
- Qifan Xiao*, **Xudong Pan***, Yifan Lu, Mi Zhang, Min Yang. Exorcising "Wraith": Protecting LiDAR-based Object Detector in Automated Driving System from Appearing Attacks, **conditionally accepted** at the 32nd USENIX Security Symposium, 2023 (*co-first authors).
- Yifan Yan*, **Xudong Pan***, Mi Zhang, Min Yang. The Hidden Vulnerability of Mainstream White-Box Deep Neural Network Watermarks under Neural Structural Obfuscation, major revision at the 32nd USENIX Security Symposium, 2023 (*co-first authors).
- Yifan Yan*, **Xudong Pan***, Yining Wang, Mi Zhang, Min Yang. "And Then There Were None": Cracking White-box DNN Watermarks via Invariant Neuron Transforms, under review at IEEE Transactions on Information Forensics and Security. (arXiv:2205.00199, *co-first authors)

Other Preprints

(students mentored by me are underlined)

- **Xudong Pan**, Qifan Xiao, Mi Zhang, Min Yang. A Certifiable Security Patch for Object Tracking in Self-Driving Systems via Historical Deviation Modeling, arXiv:2207.08556.
- Ruozi Huang, Mi Zhang, **Xudong Pan**, Beina Sheng. How Sequence-to-Sequence Models Perceive Language Styles? arXiv:1908.05947.

Impact

- My work discovered the **privacy risks of 8 commercial pretrained language models created by Google, OpenAI and Meta (Facebook)**, which were recognized as "*makes them [pretrained language models] significantly more dangerous...*"[1] in one technical report jointly written by the IT giants, and fostered a number of follow-up research on privacy-preserving language modeling.
- My work demonstrated the **stealthy trojan attacks on quantized AI models in real-world edge devices** including Raspberry Pi and NVIDIA Jetson.
- My work analyzed the copyright issues of AI models in third-party model sharing platforms, and built novel forensic techniques for tracing illegal model reuse in the wild. The techniques are now integrated in **our copyright protection platform for AI models** (leader), which is compiled as one of the best practices on AI security (along with cases from IT giants including Ant Group and IBM) in the "*AI Security Standardization Whitebook*" by China Academy of Information and Communications Technology (CAICT).
- My work successfully revealed **the feasibility of physical attacks on Baidu's Apollo self-driving system** in a closed road at campus, and proposed the defense solution which is now integrated into the Apollo open-source platform.

[1] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T.B., Song, D.X., Erlingsson, U., Oprea, A., Raffel, C. Extracting Training Data from Large Language Models. USENIX Security Symposium, 2633-2650, 2021.

Honors and Awards

- **National Scholarships**, Ministry of Education, China (2022, $4,250)
- **Academic Star**, Fudan University (2022, 10 winners across all the STEM graduate schools)
- **Student Grant**, The USENIX Association (2022, $1,000)
- **Youth Outstanding Paper Nomination**, World Artificial Intelligence Conference (2022)
- **1st Place of AutoDriving CTF**, DEFCON 30 (2022, Team Leader)
- **1st Place of AutoDriving CTF**, DEFCON 29 (2021, Team Leader)
- **Student Grant**, The USENIX Association (2020, $500)
- **National Scholarships**, Ministry of Education, China (2020, $4,250)
- **Outstanding Student Pacesetter**, Fudan University (2020, 10 winners across all the graduate schools)

Patents

- **Xudong Pan**, et al. "A Defense Agent for Enhancing the Robustness of Distributed Learning Systems", Invention Patent, China, 2020 (**authorized**)
- **Xudong Pan**, et al. "A Meta-Verifier Approach to Deep Neural Network Fingerprinting", Invention Patent, China, 2022 (pending)
- **Xudong Pan**, et al. "A Training Data Reconstruction Algorithm based on Neuron Exclusivity", Invention Patent, China, 2022 (pending)
- **Xudong Pan**, et al. "A Task-Agnostic Fingerprinting Scheme for Deep Neural Networks", Invention Patent, China, 2021 (pending)

Teaching Experience

- **Academic English**, Undergraduate Course, Teacher Assistant, Spring, 2021/2019
  Responsibility: Teaching how to do presentation on technical topics in English.
- **Introduction to Computing**, Undergraduate Course, Teacher, Spring, 2018
  Responsibility: Teaching introductory python programming and basic computation notions (e.g., loop, condition, recursion, Turing machine) to 100 **overseas students (from Asia, Europe, America, Africa, and Australia) with medical backgrounds**.

Mentoring Experience

I assisted the lab faculties in mentoring junior students for independent research (including 3 junior Ph.D., 5 MS and 2 undergraduate students).

- Ph.D. Students: *Qifan Xiao* (09/2020-present), *Yifan Lu* (09/2021-present), *Yining Wang* (09/2022-present)
- MS Students: *Beina Sheng* (09/2020-present, ByteDance), *Yifan Yan* (09/2020-present, Ant Group), *Shengyao Zhang* (09/2021-present), *Junjie Sun* (09/2022-present), *Wenxuan Li* (09/2022-present)
- Undergraduate: *Feifei Li* (09/2020-present), *Xinnuo Chen* (09/2020-present)

Service & Outreach

- Invited reviewer for CVPR (2023), AAAI (2022, 2021), ICML (2022), IJCAI (2021).
- Invited speaker and panel speaker for IJCAI China (2022).
- In-depth participation in a new revision on the "*AI Security Standardization White-book*", China Academy of Information and Communications Technology (CAICT).
- Co-founder of the *Theoretical Tools for AI Security* study group (weekly meeting, 2022/03-current) and the *Introduction to System Security for AI Researcher* reading group (covering program analysis, fuzzing, symbolic execution, network intrusion detection and how AI helps them, 2022/03-2022/06) at Fudan University.
- Assisting the lab faculties in managing, strengthening and expanding the research group. Specifically, I interviewed applicants who want to enter our research group, and participated in the collaboration negotiation with IT corporations including Baidu and Alibaba.

Presentations

- *House of Cans: Covert Transmission of Internal Datasets via Capacity-Aware Neuron Steganography*
  - The 36th Annual Conference on Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 11/28/2022.
- *General-Purpose AI Model Forensics and Its Practices*
  - The 31th International Joint Conferences on Artificial Intelligence (IJCAI), China Chapter, Shengzheng, China, 11/07/2022.
- *MetaV: A Meta-Verifier Approach to Task-Agnostic Model Fingerprinting*
  - The 28th SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Washington, DC, USA, 08/14/2022.
- *Hidden Trigger Backdoor Attack on NLP Models via Linguistic Style Manipulation*
  - The 31st USENIX Security Symposium (USENIX Security), Boston, MA, USA, 08/11/2022.
- *Exploring the Security Boundary of Data Reconstruction via Neuron Exclusivity Analysis*
  - The 31st USENIX Security Symposium (USENIX Security), Boston, MA, USA, 08/11/2022.
  - Doctoral panel at Fudan University, 04/13/2022.
- *Understanding the Threats of Trojaned Quantized Neural Network in Model Supply Chains*
  - The 38th Annual Computer Security Applications Conference (ACSAC), Austin, TX, USA, 12/06/2021.
- *TAFA: A Task-Agnostic Fingerprinting Algorithm for Neural Networks*
  - The 26th European Symposium on Research in Computer Security (ESORICS), Darmstadt, Germany, 10/04/2021.
- *Security and Privacy of Distributed Learning Systems*
  - AI Security Seminar, hosted by Zhejiang University, Hanzhou, China, 09/08/2020.
- *Justinian's GAAvernor: Robust Distributed Learning with Gradient Aggregation Agent*
  - The 29th USENIX Security Symposium (USENIX Security), Boston, MA, USA, 08/12/2020.
- *Privacy Risks of General-Purpose Language Models*
  - The 2020 IEEE Symposium on Security and Privacy (S&P), San Francisco, CA, USA, 05/18/2020.
- *Theoretical Analysis of Image-to-Image Translation with Adversarial Learning*
  - The 35th International Conference on Machine Learning (ICML), Stockholm, Sweden, 07/10/2018.