REN PANG

College of Information Sciences and Technology, the Pennsylvania State University
Email: rbp5354@psu.edu *Tel*: (484) 747-2401 *Web*: `https://ain-soph.github.io`

A. Research Interests

My current research focuses on understanding and tackling the security and privacy challenges arising in the advances of deep learning and artificial intelligence in general.

B. Education Background

Ph.D., Information Sciences and Technology, Pennsylvania State University 2019–2023

Ph.D., Computer Science and Engineering, Lehigh University (transferred) 2018–2019

BSc., Mathematics, Nankai University 2014-2018

C. Working Experience

Software Engineer (Intern), Meta 2022.05–2022.08

D. Publications

1. The Dark Side of AutoML: Towards Architectural Backdoor Search,
R. Pang, C. Li, Z. Xi, S. Ji, T. Wang,
Arxiv Preprint, 2022.

2. Demystifying Self-supervised Trojan Attacks,
C. Li, **R. Pang**, Z. Xi, T. Du, S. Ji, Y. Yao, T. Wang,
Arxiv Preprint, 2022.

3. Reasoning over Multi-view Knowledge Graphs,
Z. Xi, **R. Pang**, C. Li, T. Du, S. Ji, F. Ma, T. Wang,
Arxiv Preprint, 2022.

4. TrojanZoo: Towards Unified, Holistic, and Practical Evaluation of Neural Backdoors,
R. Pang, Z. Zhang, X. Gao, Z. Xi, S. Ji, P. Cheng, and T. Wang,
Proceedings of *the IEEE European Symposium on Security and Privacy* (EuroS&P), 2022.

5. On the Security Risks of AutoML,
R. Pang, Z. Xi, S. Ji, X. Luo, and T. Wang,
Proceedings of *the USENIX Security Symposium* (USENIX), 2022.

6. Graph Backdoor,
Z. Xi, **R. Pang**, S. Ji, and T. Wang,
Proceedings of *the USENIX Security Symposium* (USENIX), 2021.

7. i-Algebra: Towards Interactive Interpretability of Deep Neural Networks,
X. Zhang, **R. Pang**, S. Ji, F. Ma, and T. Wang,
Proceedings of *the AAAI Conference on Artificial Intelligence* (AAAI), 2021.

8. AdvMind: Inferring Adversary Intent of Black-Box Attacks,
R. Pang, X. Zhang, S. Ji, X. Luo, and T. Wang,
Proceedings of *the ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (KDD), 2020.

9. A Tale of Evil Twins: Adversarial Inputs versus Poisoned Models,
R. Pang, H. Shen, X. Zhang, S. Ji, Y. Vorobeychik, X. Luo, A. Liu, and T. Wang,
Proceedings of *the ACM Conference on Computer and Communications Security* (CCS), 2020.

E. Research Projects

1. *Embedding Vulnerabilities into Neural-Architecture-Search*
This project explores the feasibility to conduct backdoor attack during the NAS process as a new attack vector.

2. *Bridging Auto-Augment and Neural-Architecture-Search*
This ongoing project aims to bridge Auto-Augment and NAS, two major AutoML tasks, and also explore the potential security risks in their integration.

3. *Encapsulating Attack Robustness*
This ongoing project aims to achieve efficient and robust learning via synthesizing compressed versions of given datasets, on which the training bestows models with both accuracy and robustness.

4. *Evaluating Neural Backdoors*
This project builds a universal platform (TrojanZoo) that incorporate the state-of-the-art work on neural backdoors, conducts a comprehensive evaluation of existing attacks and models, and exposes the intricate design spectrum for both attackers and defenders. The findings are published in IEEE EuroS&P'22.

5. *Understanding the Security Risks of AutoML*
This project reveals that Neural-Architecture-Search (NAS) introduces new security risks: compared with manually-designed models, NAS-generated models tend to suffer greater vulnerabilities to a variety of malicious attacks. The findings are published in USENIX Security'22.

6. *Inferring the Adversary's Intent in Black-box Adversarial Attacks*
This project develops a novel estimation model to infer the adversary's intent (e.g., the target class desired by the adversary) during the early stages of black-box adversarial attacks. The findings are published in ACM KDD'20.

7. *Unifying Adversarial Inputs and Trojan Models*
This project unifies two major attack vectors, adversarial inputs and trojan models, and reveals the intriguing "mutual reinforcement" effects: by leveraging one attack vector, the adversary is able to disproportionally amplify the effectiveness of the other. The findings are published in ACM CCS'20.

F. Selected Open-Sourced Artifacts

TrojanZoo: A Unified, Holistic, and Practical Security Evaluation Platform
`https://github.com/ain-soph/trojanzoo`
A universal, flexible PyTorch platform to conduct security analysis of attacks and defenses (e.g., adversarial evasion, backdoor injection, model poisoning) on deep neural network models.

AlpsPlot
`https://github.com/ain-soph/alpsplot`
A customizable Python plotting library based on matplotlib.

TrojanZoo Sphinx Theme
`https://github.com/ain-soph/trojanzoo_sphinx_theme`
A light-weight, customizable theme that generalizes pytorch_sphinx_theme.

G. Selected Awards

Poling Scholarship, Nankai University 2014

H. Teaching Experience

CYBER 497: Machine Learning Security (Penn State), Guest Lecturer, 2020 Spring

CSE 017: Structured Programming and Data Structures (Lehigh), Teaching Assistant, 2018 Fall