Limin Yang

liminy2@illinois.edu +1 (540) 998-9158 https://liminyang.web.illinois.edu LinkedIn: liminyang GitHub: whyisyoung

Key Areas

Machine learning security, deep learning explanation, data-driven security, and malware detection and attribution.

Education

University of Illinois at Urbana-Champaign, Ph.D. in Computer Science *Aug.2019 – May 2023*
Virginia Tech, Ph.D. in Computer Science *Aug.2018 – Aug.2019*
East China Normal University, Masters Study in Computer Science *Sep.2015 – Jun.2018*
East China Normal University, BEng in Computer Science *Sep.2011 – Jun.2015*

Internships

TikTok, Security Engineering Intern, Mountain View, Mentors: Dazhuo Li and Hamed Ashouri *May 2021 – Aug.2021*

- **Spam Rule System**: Added ∼25 factors and ∼20 rules to detect ∼50,000 spams/week (extra gain: 5,000 spams/week).
- **Allowlist Domains**: added ∼300 domains to allowlist with semi-automation; protected more than 1 million emails/week.
- **User Action Clustering**: hand-picked 37 features and clustered similar emails based on user actions; it helped to double the size of ground-truth by finding more false positives and false negatives and augment threat intelligence.

XuebaJun Inc., Search & Rank Intern, Shanghai, Mentor: Yao Yu *Sep.2016 – Oct.2016*

- Search Engine: Helped reduce 33% of the response latency by debugging the searching of XuebaJun app.

Projects

University of Illinois at Urbana-Champaign, Research Assistant, Advisor: Gang Wang *Jun.2019 – Present*

- **Out-of-distribution Detection, Explainable AI**: Built a system `CADE` with self-supervised learning and autoencoder to detect and explain out-of-distribution samples. `CADE` is 2 times faster and achieves higher detection rate (F_1 = 96%) than state-of-the-art method (F_1 = 80% or lower) on Android malware and network intrusion datasets. It also worked well on a security company Blue Hexagon's PE malware and identified 161 out of 165 unseen malware families.
- **Malware Detection**: Surveyed 115 papers on how researchers use VirusTotal, measured the label dynamics of 14,000+ PE malware via daily snapshots over one year, and analyzed the correlations and causalities between VirusTotal engines. Identified questionable usage and offered suggestions for better use of VirusTotal.

Virginia Tech, Research Assistant, Advisor: Gang Wang *Aug.2018 – May 2019*

- **Phishing Detection**: Created and controlled 66 phishing sites to measure the label inconsistencies and dynamics between security vendors and VirusTotal engines. Provided insights on the poor detection performance of VirusTotal and vendors' own APIs and suggestions to utilize VirusTotal more properly on URL labelling.
- **IoT Vulnerability**: Built an Amazon Alexa skill and a Google Home action and verified that replay attack and SQL injection are feasible with proof-of-concept experiments on both platforms.

Pennsylvania State University, Research Intern, Advisors: Xinyu Xing and Gang Wang *Sep.2017 – Feb.2018*

- **Bug Reproduction**: Manually reproduced 368 bugs based on 6,000+ crowd-sourced reports. Obtained quantitative evidence on the prevalence of missing information in vulnerability reports and low reproducibility.

Top-Tier Publications

[USENIX Security'21] Limin Yang, Wenbo Guo, Qingying Hao, et al. "*CADE: Detecting and Explaining Concept Drift Samples for Security Applications*". Artifact Evaluated. Acceptance rate: 18.7%.

[USENIX Security'20] Shuofei Zhu, Jianjun Shi, Limin Yang, et al. "*Measuring and Modeling the Label Dynamics of Online Anti-Malware Engines*". Artifact Evaluated. Acceptance rate: 16.1%.

[IMC'19] Peng Peng, Limin Yang, Linghai Song, Gang Wang. "*Opening the Blackbox of VirusTotal: Analyzing Online Phishing Scan Engines*". Acceptance rate: 19.8%.

[USENIX Security'18] Dongliang Mu, Alejandro Cuevas, Limin Yang, et al. "*Understanding the Reproducibility of Crowd-reported Security Vulnerabilities.*". Acceptance rate: 19.1%.

Skills

Language: **Python (10k+ SLOC)**, C++, C. Basics: **Linux, Shell, Git, SQL, Hive**, macOS, Docker, tmux , AWS.
Deep Learning: **Keras, TensorFlow**, PyTorch. Libraries: **Scikit-learn, Numpy, Pandas**, Matplotlib.
Frameworks: Ruby on Rails, wxPython, Hadoop, PySpark, Windows SDK. Database: MySQL, PostgreSQL.