GIORGIO SEVERI

GitHub
LinkedIn
Google Scholar

Brookline, MA

Research Interests

Adversarial Machine Learning, Artificial Intelligence Security and Software Security.

Education

Ph.D., Northeastern University, Boston, MA — 2018 - 2024
Major: Computer Science.
Advisor: Prof. Alina Oprea.
Research topic: Machine Learning Security
Thesis: On the Robustness of Machine Learning Training in Security Sensitive Environments.

Master of Science, Sapienza University of Rome, Rome, Italy — 2015 - 2018
Major: Computer Science and Engineering.
Final grade: 110/110 cum Laude.
Thesis: Malwords, Malware classification and clustering based on textual memory content.

Bachelor of Science, Sapienza University of Rome, Rome, Italy — 2011 - 2014
Major: Computer Science and Engineering
Final grade: 107/110
Thesis: FreebleApp, Development of a smart, location based, mobile advertisement platform on Android OS.

Experience

Senior AI Safety Researcher — Sept. 2024 - Present
Microsoft AI Red Team, Cambridge, MA.

- Performing safety and security assessment of production scale models.
- Researching novel adversarial techniques for security testing of AI systems.

Applied research intern — Summer 2022
Microsoft AI Red Team, Redmond, WA.

- Performed security assessment of production scale models.
- Designed inference time attacks against text-to-image diffusion models to evaluate the propensity to generate undesirable content.
- Developed attacks to test the robustness of deployed safeguards for generative pipelines.
- Implemented tools to test current and future textual input filters against imperceptible perturbations with genetic optimization algorithms.

Applied research intern — Summer 2021
Microsoft Azure Trustworthy Machine Learning, (Remote) Redmond, WA.

- Performed security assessment of production scale models.
- Tested code generation large language models for memorization, PII emission, and generation of otherwise undersirable content.
- Evaluated denial of service attacks against large-scale code generation models with imperceptible textual perturbations.

- Helped with organizing the Machine Learning Security Evasion Competition by developing baseline attacks for the anti-phishing challenge.

Data Science Intern Summer 2019
FireEye Data Science, Reston, VA

- Developed new model-explanation guided backdoor poisoning attacks to target malware classificaiton models.
- Worked on hardening malware classification models through adversarial training, on domain-feasible adversarial examples.
- Developed effective ways to initialize neural network models for multi-domain (Windows/Linux/MacOS) malware classification through transfer learning.

Graduate Assistantship 2018 - 2024
Northeastern University, Khoury College of Computer Sciences, Boston, MA.

- Teaching assistant for *CY 7790: Special Topics in Security and Privacy: Machine Learning Security and Privacy* taught by professor Alina Oprea, Fall 2021.
- Graduate Fellowship for academic year 2018-2019.
- Research assistant in the Network and Distributed Systems Security Lab (NDS2) with professor Alina Oprea.

Junior Research Scientist, Summer 2017
New York University, Tandon School of Engineering, New York, NY.

- Conducted research on malware analysis and classification, with record and replay sandboxing systems.
- Employed text mining and machine learning techniques to classify and cluster malicious software samples.

Student Internship Summer 2016
European Space Agency ESA, ESRIN, Earth Observation Directorate, Italy.

- Evaluated usability of satellite image resources for Hackathon participants.
- Developed an Android mobile application, in Java, to test a newly deployed web service.

Internal work placement 2014 - 2015
Sapienza University, Department of Computer, Control, and Management Engineering Antonio Ruberti, Rome, Italy.

Publications and Patents

Blake Bullwinkel, Amanda Minnich, Shiven Chawla, Gary Lopez, Martin Pouliot, Whitney Maxwell, Joris de Gruyter, Katherine Pratt, Saphir Qi, Nina Chikanov, Roman Lutz, Raja Sekhar Rao Dheekonda, Bolor-Erdene Jagdagdorj, Eugenia Kim, Justin Song, Keegan Hines, Daniel Jones, Giorgio Severi, Richard Lundeen, Sam Vaughan, Victoria Westerhoff, Pete Bryan, Ram Shankar Siva Kumar, Yonatan Zunger, Chang Kawaguchi, and Mark Russinovich. "Lessons From Red Teaming 100 Generative AI Products". ArXiv, 2025.

Chauhari, Harsh*, Giorgio Severi*, John Abascal, Matthew Jagielski, Christopher A. Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. "Phantom: General Trigger Attacks on Retrieval Augmented Language Generation." (2024) Under submission.

*Equal contribution

Severi, Giorgio, Simona Boboila, John Holodnak, Kendra Kratkiewicz, Rauf Izmailov, and Alina Oprea. "Model-agnostic clean-label backdoor mitigation in cybersecurity environments." (2024) Under submission.

Debenedetti, Edoardo, Giorgio Severi, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Eric Wallace, Nicholas Carlini, and Florian Tramèr. "Privacy Side Channels in Machine Learning Systems." In 33rd USENIX Security Symposium (USENIX Security 24). 2024.

Chauhari, Harsh, Giorgio Severi, Alina Oprea, and Jonathan Ullman. "Chameleon: Increasing Label-Only Membership Leakage with Adaptive Poisoning." In International Conference on Learning Representations. ICLR, 2024.

Severi, Giorgio, Simona Boboila, Alina Oprea, John Holodnak, Kendra Kratkiewicz, and Jason Matterer. "Poisoning Network Flow Classifiers." In Proceedings of the 39th Annual Computer Security Applications Conference 2023.

Di Bartolomeo, Sara, Giorgio Severi, Victor Schetinger, and Cody Dunne. "Ask and you shall receive (a graph drawing): Testing ChatGPT's potential to apply graph layout algorithms." In Proc. EuroVis Conference on Visualization. 2023.

Severi, Giorgio, Will Pearce, and Alina Oprea. "Bad Citrus: Reducing Adversarial Costs with Model Distances." In 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 307-312. IEEE, 2022.

Coull, Scott Eric, David Krisiloff, and Giorgio Severi. "System and method for heterogeneous transferred learning for enhanced cybersecurity threat detection." U.S. Patent 11,475,128, issued October 18, 2022.

Severi, Giorgio*, Matthew Jagielski*, Gökberk Yar, Yuxuan Wang, Alina Oprea, and Cristina Nita-Rotaru. "Network-level adversaries in federated learning." In 2022 IEEE Conference on Communications and Network Security (CNS), pp. 19-27. IEEE, 2022.

Jagielski, Matthew, Giorgio Severi, Niklas Pousette Harger, and Alina Oprea. "Subpopulation data poisoning attacks." In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, pp. 3104-3122. 2021.

Severi, Giorgio, Jim Meyer, Scott Coull, and Alina Oprea. "Explanation-Guided Backdoor Poisoning Attacks Against Malware Classifiers." In 30th USENIX Security Symposium (USENIX Security 21). 2021.

Severi, Giorgio, Tim Leek, and Brendan Dolan-Gavitt. "Malrec: compact full-trace malware recording for retrospective deep analysis." In International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, pp. 3-23. Springer, Cham, 2018.

Talks

"Zen and the Art of Adversarial Machine Learning". Will Pearce, Giorgio Severi. Black Hat Europe 2021, London, UK.

"Exploring Backdoor Poisoning Attacks Against Malware Classifiers". Giorgio Severi, Jim Meyer, Scott Coull. Conference on Applied Machine Learning in Information Security, CAMLIS, 2019, Washington, DC.

Academic Service

Program Committee member for the 17th ACM Workshop on Artificial Intelligence and Security 2024.

Program Committee member for the Workshop on Artificial Intelligence System with Confidential Computing (AISCC) 2024.

Program Committee member for the 16th ACM Workshop on Artificial Intelligence and Security 2023.

Program Committee member for the DSN Workshop on Dependable and Secure Machine Learning 2023.

Shadow Program Committee member for the IEEE Symposium on Security and Privacy 2021.

Additional Experience

Staff member at Codemotion Rome, 2017 and 2015.
Mentor at "Tech My Cosplay", Arduino Hackathon Rome, 2017.
Staff member at Data Driven Innovation Rome 2017.
Staff member at Maker Faire Rome 2014.

Languages

Italian, native speaker.
English, European level CEFR C2. IELTS score: 8.5/9. ESOL CPE certificate.

Awards

Winner Accenture Digital Hackathon Rome 2016.
NASA International SpaceApps Challenge 2015.

- Project CROPP, Global winner for category Galactic Impact and Rome local competition.