I am a computer security researcher interested in developing robust machine learning techniques to solve security problems with strong theoretical guarantees and high performance. I have combined formal methods, deep learning, and economic cost modeling to train security classifiers with verified robustness properties, that can increase the economic cost for adaptive attackers to evade detection.

Machine learning techniques have shown tremendous promise to solve security problems in production. For example, Google applies deep learning to scan malicious email attachments, antivirus companies use machine learning to detect malware and Advanced Persistent Threat, and social network platforms deploy machine learning techniques to fight spam, online fraud, and hate speech. Despite wide adoption, ML-based detection systems are not robust against attacks. For instance, Gmail's malicious document scanner can be evaded by trivial transformations over the PDFs [1, 2]. In the arms race between defenders and attackers, my research focus has been to increase the cost for attackers to succeed. Unfortunately, state-of-the-art defenses are not robust against adversaries in real-world settings and significantly sacrifice performance to gain robustness.

My key insight is to gain domain knowledge from measurement studies, and then integrate the domain knowledge to build principled defenses. I have combined techniques across different research areas: online abuse measurement, formal methods and algorithm design. First, I have measured the operations, economic cost and profit of various malicious activities, which allowed me to understand what kinds of theoretical guarantees are practical for security classifiers, and where we have the opportunities to increase the attack cost. Second, I have formalized robustness properties that, if achieved, would increase the economic cost of evasion for adaptive attackers. Third, I have designed and implemented new algorithms to train provably robust security classifiers to satisfy all the robustness properties while maintaining good performance.

My research is among the first to introduce and enforce *verified robustness properties for security classifiers*. Robustness properties are security guarantees of the classifier that can provably eliminate arbitrary attacks under some restricted threat models. My key result is, enforcing robustness properties can increase the economic cost of evasion for unrestricted adaptive attackers who know about the defense, with only modest impact on performance. I have applied my work to security datasets including Cryptojacking, Twitter spam, and PDF malware, and demonstrated that it is not only sound but also practical [3, 4, 5]. My work strikes a balance between providing theoretical guarantees and building practical machine learning tools and detection systems. My research has resulted in seven top-tier security conference papers, received a CCS Best Paper Award Runner-Up [3], and also made an impact in the real-world. I have developed new methods for collecting DNS data that have been used by more than 70 researchers from 50 companies and organizations in more than 15 countries [6]. Our dataset has enabled use cases such as detecting phishing webpages, malicious residential IP addresses, and trademark abuse. My ongoing work has resulted in a funding award from Google ASPIRE (Android Security and PrIvacy REsearch) program. In the future, I intend to apply these approaches to solve real-world problems in network security and malware detection.

Completed Research

In order to use machine learning to solve security problems, I need to gain domain knowledge about attack operations and attack costs. In my thesis, I have conducted measurement studies over malicious behaviors [6, 7, 8], the impact and profit of attackers [9, 10] and the evasion cost of attackers [11]. From these studies, I have learned that: *attackers are constantly evading the defense, but they prefer minimizing the economic cost of evasion*. For example, to evade online abuse detection systems that use features related to domain names and their IP addresses, the attackers purchase cheap domain names a lot more frequently than expensive server IPs.

My goal is to increase the economic cost for attackers to evade security classifiers. Unfortunately, existing defenses do not model the behaviors of real-world attackers. Therefore, I proposed new notions of robustness that can better model the security threats. I formulated real-world attackers' threat models as robustness properties. A robustness property specifies that, arbitrary attacks restricted by the property must not change a malicious prediction to benign. Since attackers adapt the strategies using knowledge about the defense, the robustness properties should eliminate low-cost attacks, forcing the adaptive attackers to use more sophisticated and expensive strategies.

I used domain knowledge and economic cost measurement studies to formulate global robustness properties for security classifiers [3], building block robustness properties for PDF malware classifiers [4], and low-cost attacks against tree models [5]. I proposed new training algorithms for security classifiers to obtain these properties. My results demonstrated that we can increase the evasion cost for adaptive attackers unrestricted by the properties.

Verified Global Robustness Properties for Security Classifiers [3]. State-of-the-art robust training methods can achieve local robustness guarantees, which hold for some inputs, but not all inputs. In contrast, global robustness guarantees hold for all inputs, which provide robustness guarantee even if the attacker makes new inputs, and they is highly desirable for security classifiers.

Property: We use domain knowledge and measurement studies to analyze broad classes of evasion strategies that are common in practice. Then, we mathematically formulate six global robustness properties, among which five are new properties. For example, we measure that, the Twitter account price increases as the follower number increases. Thus, we formulate the monotonicity property as, the output of the fake account classifier monotonically decreases as the follower number increases, forcing the attacker to purchase more followers to evade detection.

Proof: We structure the classifier as an ensemble of logic rules. We use an integer linear program (ILP) to encode the classifier and violation of the properties. If the ILP is infeasible, the classifier satisfies the properties.

We propose the first general algorithm to train security classifiers with verified global robustness properties. Our key idea is to apply Counterexample Guided Inductive Synthesis (CEGIS). We design (1) a verifier to check whether a candidate classifier satisfies the properties and find counterexamples that violate the properties, and (2) a synthesizer to generate new candidate classifiers that eliminate violations of the properties. They continuously interact with each other until we find an accurate classifier that satisfies all robustness properties. We have evaluated our technique over datasets to detect Cryptojacking, Twitter spam accounts, and Twitter spam URLs. In comparison to seven state-of-the-art training methods, we are the first to train classifiers to satisfy all the properties. For example, we can train a Twitter spam account classifier that satisfies five global robustness properties with 89.4% accuracy. Whereas, the best existing training algorithm can achieve 91.9% accuracy with only one property.

Robustness Properties for PDF Malware Classifiers [4]. PDF malware classifiers are ubiquitous in users' lives. Previous method to train robust PDF malware classifiers increased the false positive rate to as high as 85% [12]. We need new techniques to achieve robustness with low false positive rate.

Property: Since the attacker needs to preserve the syntax of PDF to generate real PDF malware variants, adversarial malware variants are different from the seed malware by a number of PDF subtrees. We define a new distance metric in PDF subtrees to formulate robustness properties. For example, given a seed PDF malware, any malware variant that has deleted an arbitrary PDF subtree from the seed malware must be predicted as malicious.

Proof: For a given malicious input, we use symbolic interval analysis [13] to over-approximate the output of the neural network model under potential attacks, and then prove that the over-approximated output of the corresponding malware variants are predicted as malicious.

It is hard to obtain multiple properties simultaneously. Thus, we propose a new certifiable training technique [14] for neural networks to balance the training objective among multiple robustness properties while achieving high test accuracy. Our technique over-approximates the effect of potential attacks over the training data, and minimizes the errors caused by them. Our results show that, we can train a robust classifier to achieve three properties with only 0.56% false positive rate and no impact on the regular test accuracy. Sophisticated attacks are composed of building block operations that insert or delete one subtree in the PDF malware. With building block robustness properties defined at subtree distance one, our robust model maintains 7% higher robust accuracy than all of state-of-the-art baseline models against unrestricted whitebox attacks.

Cost-aware Robust Trees for Security [5]. Tree ensemble models such as random forest and gradient boosted decision trees are widely used in security. Despite their popularity, the robustness of these models, especially against a strong adversary who is aware of the feature manipulation cost, has not been thoroughly studied.

Property: First, we analyze the cost unit to perturb each feature, where a unit could represent one dollar. Then, we give the attacker a budget to perturb each feature value such that a feature with larger cost unit has a smaller perturbation range. We formulate the low-cost attack under a given budget as the property.

Proof: For a given input, we use a mixed integer linear program (MILP) to encode the output of the tree ensemble under potential attacks, with the goal of minimizing the total attack cost. By solving the MILP, we prove the minimal attack cost over each input that can evade the classifier.

The discrete structure of trees brings new challenges to robust training, since it is intractable to enumerate all possible node splits under attacks. We design a new training algorithm to learn cost-aware robust tree ensembles. To find a new node split, our training algorithm greedily computes the worst quality of the tree as if an arbitrary attacker can perturb the training data points under a given budget for the low-cost attacks, and then maximizes the tree quality. We design a new adaptive whitebox attack that minimizes the evasion cost for each input, using knowledge about the trained cost model. For a Twitter spam URL detection dataset, our robust classifier can increase the evasion cost by $10.6\times$ compared to the state-of-the-art method against the new adaptive attack.

Future Research Directions

To solve security problems using machine learning, we inevitably face adversaries who try to bypass the defense. Therefore, we must build ML-based systems and tools with robustness by design, instead of as an after-thought. Since we cannot predict the future attacks, we must involve humans in the process to utilize their expertise. I will explore techniques to decrease the cost for defenders and reduce the benefits of successful attacks.

Automated Repair of Attack Detection Systems. Malicious and benign behaviors change over time, which has been observed in many security problems including spam filtering, malware detection, BGP hijacking detection, etc. This causes the performance of attack detection systems to quickly degrade, so we must repair them to keep up with the change. State-of-the-art methods rely on analysts to manually craft new signatures for rule-based detectors and label new training samples to repair the ML models. This is very expensive and time-consuming, and is not sufficient to rapidly update the detection systems facing a lot of incoming data (e.g., network traffic analysis). To lower the cost for the defenders, I want to automate the repair for detection systems. In the short term, I plan to use semi-supervised learning techniques to automatically label new data and generate rules, and combine automated repair with analyst repair, to alleviate the burden on analysts. In the long term, I will study how to efficiently and effectively repair detection systems using online learning but also be robust against potential poisoning attacks.

Security Analyst Assistants. Security analysts perform jobs such as threat hunting, enterprise log analysis, binary reverse engineering, etc. They have limited bandwidth and the vast majority of data end up being overlooked. I want to design an automated analyst assistant to provide the most relevant information to the analysts and improve their productivity. For example, in the reverse engineering task, an analyst mentally compares a function's behavior with known implementations, such as common encryption algorithms. If the automated assistant provides potential matches, the analyst can choose the best match and focus on analyzing the big picture. I want to learn common sense models to represent vast amounts of domain knowledge in security. In the short term, I will learn function similarity measures using self-supervised learning over new common knowledge data for reverse engineers, and more broadly explore other common self-supervised tasks for security. In the long term, I will explore how to integrate different assistants into an end-to-end recommendation system to help analysts with different tasks.

Transfer Security Knowledge. Attackers always adapt to target a new platform as new technologies emerge. For example, Cryptojacking malware initially mostly existed in the browsers, and then moved on to target IoT devices and Docker engines that were less protected. As defenders, why do we always start from scratch when the same threat slightly changes the infection or execution mechanisms? I want to investigate techniques to transfer the security domain knowledge when we build new detection models. For example, although the implementation of Cyptojacking malware has changed, how the malware interacts with the outside world may remain largely similar, such as getting commands from the attacker to mine cryptocurrencies, or calling some APIs to persist the execution. In the short term, I will explore methods to learn machine learning modules to detect different suspicious behaviors (that may not be necessarily malicious), such that we can use them as building blocks to train a larger model in a new task. In the longer term, I will study how to ensure robustness when transferring the security knowledge.

References

[1] Weilin Xu, Yanjun Qi, and David Evans. NDSS Talk: Automatically Evading Classifiers (including Gmail's). https://jeffersonswheel.org/2016/ndss-talk-automatically-evading-classifiers-including-gmails.

[2] **Yizheng Chen**. Gmail's Malicious Document Classifier Can Still Be Trivially Evaded. https://surrealyz.medium.com/gmails-malicious-document-classifier-can-still-be-trivially-evaded-93e625745c9d.

[3] **Yizheng Chen**, Shiqi Wang, Yue Qin, Xiaojing Liao, Suman Jana, and David Wagner. Learning Security Classifiers with Verified Global Robustness Properties. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2021.

[4] **Yizheng Chen**, Shiqi Wang, Dongdong She, and Suman Jana. On Training Robust PDF Malware Classifiers. In *29th USENIX Security Symposium (USENIX Security)*, 2020.

[5] **Yizheng Chen**, Shiqi Wang, Weifan Jiang, Asaf Cidon, and Suman Jana. Cost-Aware Robust Tree Ensembles for Security Applications. In *30th USENIX Security Symposium (USENIX Security)*, 2021.

[6] Athanasios Kountouras, Panagiotis Kintis, Chaz Lever, **Yizheng Chen**, Yacin Nadji, David Dagon, Manos Antonakakis, and Rodney Joffe. Enabling Network Security Through Active DNS Datasets. In *International Symposium on Research in Attacks, Intrusions, and Defenses (RAID)*. Springer, 2016.

[7] Panagiotis Kintis, Najmeh Miramirkhani, Charles Lever, **Yizheng Chen**, Rosa Romero-Gómez, Nikolaos Pitropakis, Nick Nikiforakis, and Manos Antonakakis. Hiding in Plain Sight: A Longitudinal Study of Combosquatting Abuse. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2017.

[8] **Yizheng Chen**, Yacin Nadji, Rosa Romero-Gómez, Manos Antonakakis, and David Dagon. Measuring Network Reputation in the Ad-Bidding Process. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA)*. Springer, 2017.

[9] Tielei Wang, Yeongjin Jang, **Yizheng Chen**, Simon P Chung, Billy Lau, and Wenke Lee. On the Feasibility of Large-Scale Infections of iOS Devices. In *23rd SENIX Security Symposium (USENIX Security)*, 2014.

[10] **Yizheng Chen**, Panagiotis Kintis, Manos Antonakakis, Yacin Nadji, David Dagon, Wenke Lee, and Michael Farrell. Financial Lower Bounds of Online Advertising Abuse. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA)*. Springer, 2016.

[11] **Yizheng Chen**, Yacin Nadji, Athanasios Kountouras, Fabian Monrose, Roberto Perdisci, Manos Antonakakis, and Nikolaos Vasiloglou. Practical Attacks Against Graph-based Clustering. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2017.

[12] David Evans. Adversarial Machine Learning: Are We Playing the Wrong Game? https://speakerdeck.com/evansuva/adversarial-machine-learning-are-we-playing-the-wrong-game.

[13] Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana. Efficient formal safety analysis of neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[14] Shiqi Wang, **Yizheng Chen**, Ahmed Abdou, and Suman Jana. MixTrain: Scalable Training of Formally Robust Neural Networks. *arXiv preprint arXiv:1811.02625*, 2018.