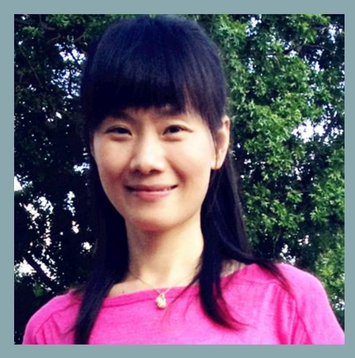
Intelligent Malware Detection Utilizing Novel File Relation-Based Features and Resilient Techniques for Adversarial Attacks

PI: Yanfang Ye (WVU), Co-PI: Katerina Goseva-Popstojanova (WVU)

http://www.csee.wvu.edu/~yaye/

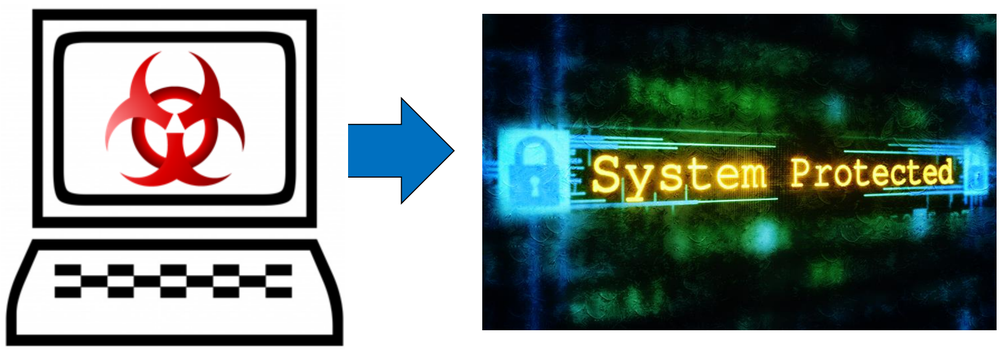


Project Description

Driven by the considerable economic benefits, both the sophistication and the quantity of malware have significantly increased. To protect legitimate users from the evolutional malware attacks, we **aim to** develop much more powerful methods which are capable of protecting the users against new threats, and are more difficult to evade.

Over the last seven years, the PI has been working in collaborations with the industry partners (e.g., Kingsoft and Comodo) for intelligent malware analysis and detection. Built on the PI's long-term and strong collaboration with the anti-malware companies in the preliminary work, several key challenges for modern malware detection have been identified:

How secure is your computer?



Our goal is to design and develop intelligent and resilient solutions against malware attacks.

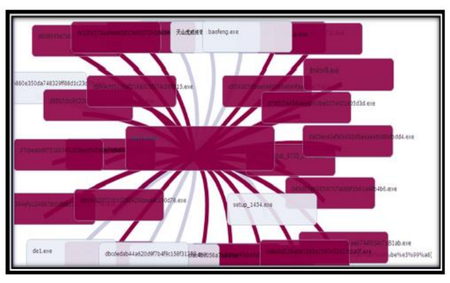
1. Besides file contents, what kinds of newly novel features can be used for malware detection?
2. How to construct an effective model to detect the unknown malware utilizing both content-based and relation-based features?
3. How to develop resilient techniques that are robust and secure against adversarial attacks?

Approach

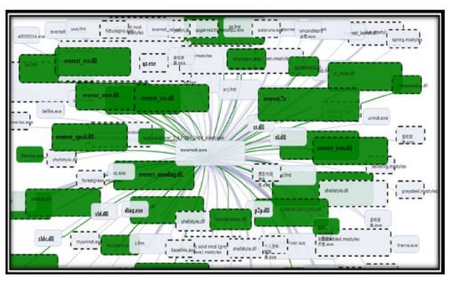
This project is to create a resilient platform against adversarial malware attacks.

- Design newly novel relation-based features for malware representations.

  E.g., as the moral says "*man is known by the company he keeps*", a file's legitimacy can be judged by the other files that often co-occur/co-exist with it on the users' machines.



File co-occurrences between a Trojan-downloader and its related Trojans



File co-occurrences between a benign app and its related dynamic link files

- Design and develop a semi-supervised learning framework utilizing both content-based (e.g., CFGs, system calls) and relation-based features (e.g., file co-occurrences, file co-operations) for malware detection.

- Design and develop resilient techniques against adversarial attacks on machine learning/data mining based models.

  Machine learning and data mining techniques offer unparalleled flexibility in intelligent malware detection. However, machine learning or data mining algorithms themselves can be a target of attack by a malicious adversary. Specifically, use of machine learning or data mining techniques may open the possibility of an adversary who maliciously "mis-trains" a learning system (e.g., by changing the data distribution or feature importance) in a malware detection system. In this project, we will investigate:

  ✓ What techniques can an adversary use in their attacks to confuse a learning system for malware detection?

  ✓ How to develop resilient techniques that are robust and secure in adversarial scenarios?

The developed techniques are designed to be arms race capable so that they can also be used in other security domains, such as anti-spam, fraud detection, and counter-terrorism.

Publications supported by this funding

- **Yanfang Ye,** Tao Li, Donald Adjeroh, **Katerina Goseva-Popstojanova**, S. Sitharama Iyengar. "A Survey on Malware Detection Using Data Mining Techniques", ***ACM Computing Surveys***, 2016. (Accepted)
- Gongde Guo, Lifei Chen, **Yanfang Ye,** Qingshan Jiang. "Cluster Validation Method for Determining the Number of Clusters in Categorical Sequences", ***IEEE Transactions on Neural Networks and Learning Systems***, Vol. PP (99): 1-13, 2016.
- Shifu Hou, Aaron Saas, **Yanfang Ye,** Lifei Chen. "Deep4MalDroid: A Deep Learning Framework for Android Malware Detection Based on Linux Kernel System Call Graphs", ***Proceedings of the IEEE/WIC/ACM WI Workshop on Advanced Methods in Optimization and Machine Learning (WI-BIH),*** 2016.
- Shifu Hou, Aaron Saas, **Yanfang Ye,** and Lifei Chen. "DroidDelver: An Android Malware Detection System Using Deep Belief Network Based on API Call Blocks", ***Proceedings of International Conference on Web-Age Information Management Workshop on Mobile Web Data Analytics (WAIM-MWDA),*** 2016.
- William Hardy, Lingwei Chen, Shifu Hou, **Yanfang Ye**, and Xin Li. "DL4MD: A Deep Learning Framework for Intelligent Malware Detection", ***The 12th International Conference on Data Mining (DMIN),*** 2016.

Integrating Research with Education

- **Curriculum Development Activities.** PI Ye is teaching a graduate level course *CS591L Cyber Security and Big Data Analytics* and Co-PI Goseva-Popstojanova is teaching an undergraduate level course *CS465 Introduction to Computer Security* at WVU.
- **Robust Outreach Efforts** to K-12, general public, undergraduate, graduate, minority, and women in cyber security.

PI Ye: Growing Roots in STEM - Engineering Challenge Camp (All Female)
"Cyber Security in Your Everyday Life", June 26-July 1, 2016.

Interested in meeting the PIs? Attach post-it note below!

The 3rd Biennial NSF Secure and Trustworthy Cyberspace Principal Investigator Meeting

Jan. 9 – 11th 2017

Arlington, Virginia

National Science Foundation
WHERE DISCOVERIES BEGIN