Android Malware Detection Powered by Formal Analysis and Machine Learning Approaches

The goal of this project was to develop machine learning techniques to detect malicious applications (apps) found in Android markets (such as Google Play). We leverage on static program analysis techniques to construct program dependency graph (PDG) representations of apps. Subsequently, semantic features are extracted from PDGs using which machine learning classifiers are trained to perform malware detection. To this end, we have made two contributions: (1) developed a graph kernel that facilitates capturing structural and contextual information from PDGs (2) developed an unsupervised learning technique to learn semantic representations of PDG subgraphs. These two contributions are explained in brief detail in the two following paragraphs.

Application-specific Graph Kernel. Recent research on malware analysis has revealed that besides capturing neighborhood (i.e., structural) information from PDGs it is important to capture the context under which the neighborhoods are reachable to accurately detect malicious neighborhoods. While many existing graph kernels such as Weisfeiler-Lehman Kernel (WLK), excel in capturing structural information from PDGs, they fail to capture the contextual information, as it is a strong domain- and application-specific requirement. To bridge this research gap, we have developed the Contextual WLK (CWLK), which precisely capture both the aforementioned types of information from PDGs. With CWLK, we have designed a framework as presented in Figure 1 to perform context-aware malware detection. Through our large-scale experiments with more than 50,000 real-world Android apps, we demonstrate that CWLK outperforms two state-of-the-art graph kernels (including WLK) and three malware detection techniques by more than 5.27% and 4.87% F-measure, respectively, while maintaining high efficiency. This high accuracy and efficiency make CWLK suitable for large-scale real-world malware detection.

Subgraph Representation Learning. Many existing graph kernels (incl. WLK and CWLK) use rooted subgraph patterns from a given pair of graphs to evaluate their similarity. However, they do not account for the similarity that exists among these subgraph patterns. Since these kernels regard subgraphs as separate features, the dimensionality of the feature space often grows exponentially with the number of rooted subgraphs. Consequently, only a few subgraphs will be common across graphs. This leads to kernel diagonal dominance, that is, a given graph is similar to itself but not to any other graph in the dataset. This leads to poor classification accuracy. To address this research gap, we developed subgraph2vec, a novel approach for learning latent representations of rooted subgraphs from large graphs inspired by recent advancements in Deep Learning and Natural Language Processing. These latent representations encode PDG substructure dependencies in a continuous vector space, which is easily exploited by statistical models for tasks such as malware detection. Also, we show that the subgraph vectors could be used for building a deep learning variant of WLK. Our experiments reveal that subgraph2vec achieves significant improvements in malware detection accuracies over existing graph kernels.

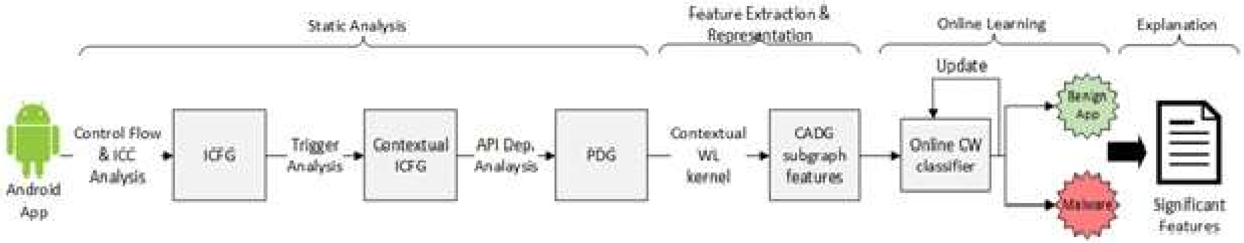

Figure 1. Context-aware Android Malware Detection Framework. The framework uses CWLK and an online classifier to perform accurate and efficient malware detection.

Principal Investigators	*Assoc. Prof. Chen Lihui (EEE)*
Student	Narayanan Annamalai
School / Dept	School of Electrical & Electronic Engineering / Infinitus
Collaborator	Asst. Prof Lui Yang
Grant Agency	NTU Research Scholarship