Familial Clustering For Weakly-Labeled Android Malware Using Hybrid Representation Learning

Yanxin Zhang; Baodi Ning; Shirui Pan[3]; Ivor Tsang[2]; Yulei Sui[1]
University of Technology Sydney[4], Faculty of Engineering and IT[5]

Abstract

This paper presents ANDRE, a new hybrid representation learning approach to clustering weakly-labeled malware by preserving heterogeneous information from multiple sources (including the results of static code analysis, the meta-information of an app, and the raw-labels of the AV vendors) to jointly learn a hybrid representation for accurate clustering.

The learned representation is then fed into our outlier-aware clustering to partition the weakly-labeled malware into known and unknown families. The malware whose malicious behaviours are close to those of the existing families on the network, are further classified using a three layer Deep Neural Network (DNN). The unknown malware are clustered using a standard density-based clustering algorithm.

Introduction

A substantial number of weakly-labeled malware are unable to be clustered because their raw labels are often incomplete, inconsistent and overly generic, as reported by incompatible commercial vendors with different capabilities in the presence of rapidly evolving malware. A strongly-labeled malware with a clear and unambiguous family name is easy to be clustered. However, relying on vendors' raw labels as the only source of information for labeling weakly-labeled malware is inherently partial and shallow.

Deep representation learning (DRL) is a new and promising branch of machine learning. DRL learns the representation of the target data via deep architectures in a layer-wise manner, through which the higher abstraction level of the features is embedded in a lower unified representation, making it easy for later classification and clustering tasks.

Inspired by the recent advances in DRL, ANDRE jointly learns a hybrid representation that allows heterogeneous information to be integrated into one neural network pipeline that distills the discriminative features for accurate clustering. ANDRE is based on a new Android malware network, in which every node represents an app whose label contains its meta-info and the raw reports from AV vendors, and every edge denotes the similarity between two apps inferred by our static analysis which exploits a pairwise analysis of code similarity.

Methods and Materials

Our framework consists of the following three major components.

1. Feature Extraction. There are two steps for extracting features, including analysing code similarity and the Android manifest files.

2. Hybrid Representation Learning. Our hybrid method couples two neural networks to learn the representation from node structure and meta-info.

3. Outlier-aware weakly-labeled malware clustering. ANDRE performs an outlier-aware clustering that partition all the weakly-labeled malware apps into inlier apps, and anomalous apps.

Results

Table 1 gives the inliers and outliers of the 3324 weakly-labeled malware. The number of outlier samples is relatively small, comprising only 2.5%. A major portion of the weakly-labeled malware are detected as inliers. For the 1534 empty-labeled malware, 35 and 1449 are outliers and inliers respectively. Of the 1790 malware who have controversial family names (i.e., the top two most frequent family names reported by an equal number of vendors), 47 and 1743 are outliers and inliers respectively.

Figure 1 shows the outlier results with 58 outlier (unknown) families clustered by anomaly detection. 47 out of 58 outliers contain only 1 malware sample, and the remaining 11 include more than 1 candidate. The largest outlier cluster contains 9 samples.

For inliers, there are 60 families out of the 176 known families. The distribution is uneven. The malware apps in the top 10 families occupy the majority (88.26%) of all the weakly-labeled malware.

RESULTS OF OUTLIERS AND INLIERS OF WEAKLY-LABELED MALWARE INCLUDING MALWARE WITH EMPTY LABELS AND CONTROVERSIAL LABELS (I.E., THE TOP TWO MOST FREQUENT FAMILY NAMES REPORTED BY AN EQUAL NUMBER OF VENDORS USING EUPHONY).

#Weakly-labeled malware	#Malware in outlier		#Malware in inlier	
	Empty-labeled	Dispute-labeled	Empty-labeled	Dispute-labeled
# 3324	35	47	1499	1743

Table 1. Inliers and Outliers of the 3324 weakly-labeled malware

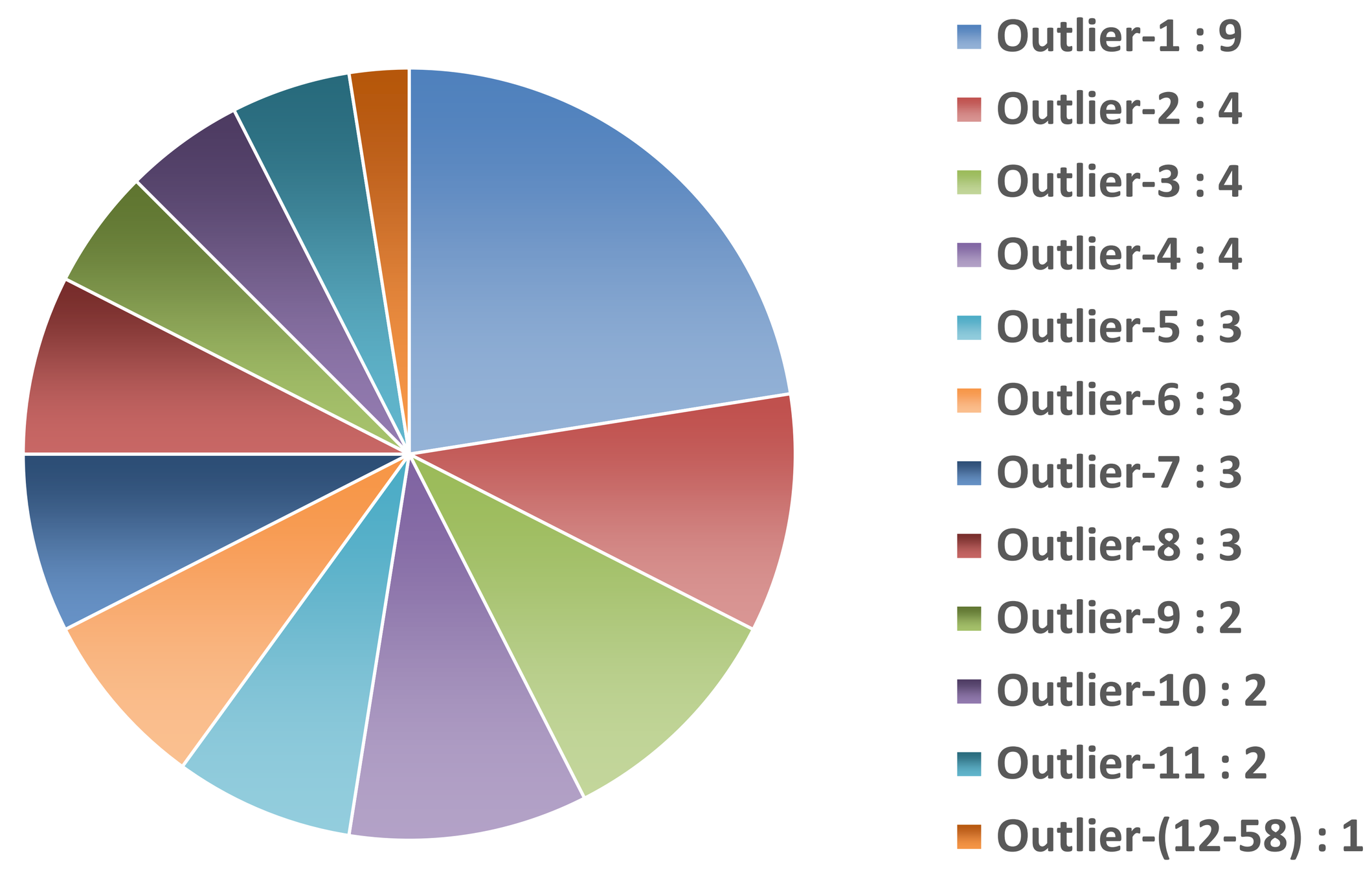

Figure 1. Weakly-labeled malware detected as outliers

Conclusions

This paper proposes ANDRE, a new approach to Android malware clustering that utilizes heterogeneous information including code similarity, the raw labels of AV vendors and meta-data information to jointly learn an effective representation that embeds all malware in the network into a low dimensional and compact hybrid feature space for effectively clustering weakly-labeled malware. The experimental results show that ANDRE achieves comparable accuracy to the state-of-the-art approaches for clustering ground-truth samples and that ANDRE can effectively cluster weakly-labeled malware which cannot be clustered by those approaches.

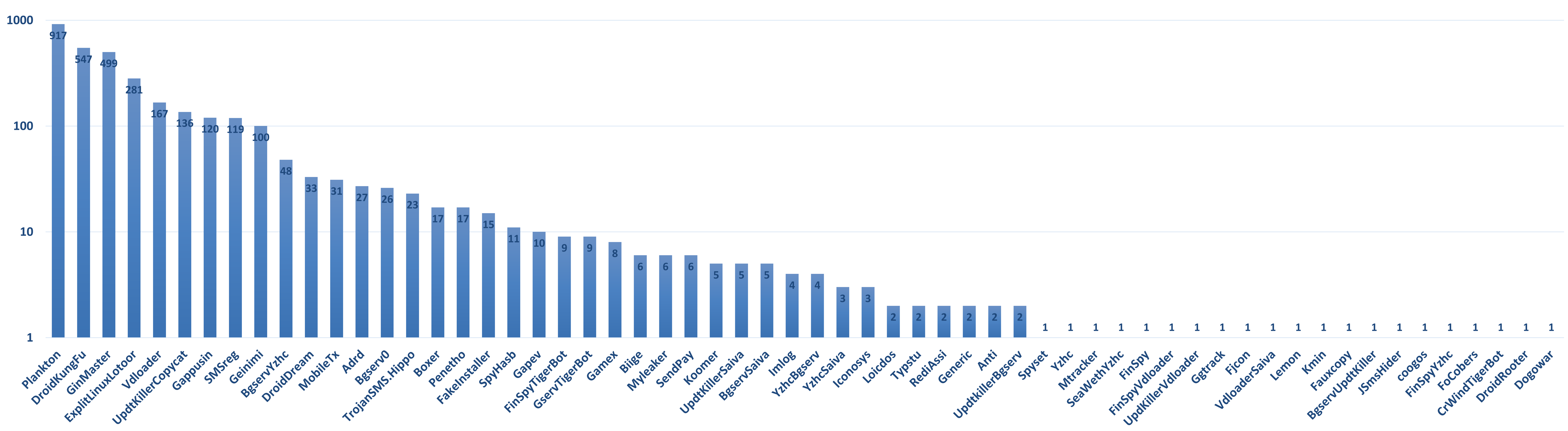

Figure 2. Weakly-labeled malware detected as Inliers

Contact

Yanxin Zhang
School of Software
Faculty of Engineering and IT
University of Technology Sydney
Email: yanxin.zhang@student.uts.edu.au

References

1. https://www.uts.edu.au/staff/yulei.sui
2. https://www.uts.edu.au/staff/ivor.tsang
3. https://research.monash.edu/en/persons/shirui-pan
4. https://www.uts.edu.au/
5. https://www.uts.edu.au/about/faculty-engineering-and-information-technology