UNIVERSITY of GUELPH

COLLEGE of ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

MSc Seminar

Friday December 13, 2019 at 1PM in Reynolds, Room 2224

An Optimized Positive-Unlabeled Learning Method for Detecting a Large Scale of Malware Variants

Mohammad Khan

Advisor: Dr. Xiaodong Lin

Advisory Committee: Dr. Charlie Obimbo

ABSTRACT:

A study at the University of Maryland shows that there is a cybersecurity attack every 39 seconds. The severity of these attacks varies, but just for data breaches, since 2013, there have been close to a million records stolen, DDOS attacks have increased by 500%. Just in terms of cost, according to Accenture, the average cost of a malware attack on a company is 2.4 million, and the costs of cybersecurity is increased by 22% just from 2016 to 2017. This goes to show that it is imperative for us to fight against these threats.

There are two traits many new malware incorporate (polymorphic and metamorphic), where the idea is to change the unimportant parts of a malware such that it's difficult to compare the old and new version. Current malware detectors rely on the fact that every new malware that is studied, is a copy of a previously known malware. Thus most detectors will be unable to detect polymorphic or metamorphic malware. This brings the notion that machine learning will be the next step towards combating these malicious traits.

A big industrial problem that currently exists is that companies receive too many executable softwares, and they don't have the resources to label these software as malware or benign. Currently, these executables are split into 3 sections, malware, benign, and unlabeled. When training machine learning models to detect new malware, companies treat the unlabeled executables as benign. Which can be detrimental when it comes to creating a trustworthy model, because unlabeled section may contain malware inside of them. Our solution is to create a cost-efficient method that can be implemented into various malware detection models such that when working with unlabeled data, they will attach a weight to each iteration of training. This will make it so that even if we treat the unlabeled data as benign, our model will be trustworthy.