An Epidemiological Study of Malware Encounters in a Large Enterprise

Jeff Malavasi
Dept. of Computing Security
Rochester Institute of Technology
Email: jm3378@rit.edu

I. Summary

A. Problem

In a large enterprise, it is difficult to detect and contain malware due to geographic diversity as well as other factors. This is especially true as more and more organizations are shifting to a remote workforce, which dramatically increases the number of networks corporate assets interact with. When creating a security program for an enterprise, it is important to consider how risk may change across countries, organizational units, and networks.

B. Solution

The authors propose a novel risk model that uses regression in order to determine the likelihood of a particular employee encountering malware. This is achieved by first creating a dataset comprised of 85,000 hosts and then implementing a two-stage feature selection process comprised of three categories: demographics, VPN activity, and web activity [1].

C. Methodology

The researchers started by collecting anti-virus reports from a large enterprise. They collected results over a four-month period using MacAfee anti-virus agents which was responsible for reporting the host name, malware detected, relevant file path and the detection time. After deduping the results and removing outliers, the researchers enriched the dataset with an employee demographic database, windows authentication logs, VPN logs, and web proxy logs. Specifically, each employee was assigned a level based on hierarchy and additionally categorized based on job role. VPN logs were used in order to determine whether or not the encounter occurred on or off the corporate network.

In order to select the most significant features to use for their risk detection system, the researchers first create a logistic regression model for each category. This helps to determine which features are most effective, and are then loaded into the final model.

D. Results

After analyzing the dataset, the authors found that 15% of the machines investigated encountered malware. The results were especially concentrated in the United States, India, Ireland and China. It was determined that although slow, the most common delivery method was via an external drive [1]. Additionally, when connected to the corporate network, hosts appear to be less vulnerable to web attacks, likely attributed to content filters and zero-trust network appliances. However, this pattern was not true for all countries. For example, EMEA employees were more likely to encounter malware on the corporate network, as they often do not work outside the office as much as their American counterparts. When comparing employee levels and roles, the authors found that higher level employees were less likely to encounter malware. Additionally, job functions that require the use of computers, increased the risk significantly. Finally, among the 30% of malware that was detected on network, the majority of it was categorized as either business, communications, or search.

The researchers found that user demographics is the most effective category at predicting user risk, with VPN activity closely behind. Shockingly, they found web activity to be an extremely weak feature with less than 5% of encounters originating from the internet.

II. Evaluation

A. Strengths

The researchers successfully create a risk model based on user demographics, VPN and web activity that is able to accurately predict the likelihood of a specific host or user encountering malware. Additionally, the researchers gathered data from a variety of users and hosts which helps to support their reproducibility. Finally, they implement a feature reduction process in order to select relevant features without adding extraneous overhead to the model.

B. Weaknesses

While the authors are able to accurately determine risk in the evaluated enterprise, their study also has many limitations. For example, they only evaluated one organization. Each industry carries a different risk model and would likely have unique attack vectors. Additionally, they only utilized one anti-virus collection agent. This means that the dataset is skewed towards the signatures provided by that application. Finally, their features were based on indirect signaling which required the researchers to correlate events together in order to infer information about the host or user. This creates the potential for error in the underlying dataset.

C. Significance

The problem addressed in the paper is significant for a couple of reasons. First, the researchers evaluate whether risk changes when the employee moves on and off the corporate network. This is an extremely relevant problem as organizations allow hybrid and remote work. Additionally, organizations often lack the resources to create granular risk policies. The researchers address this by creating a risk scoring system that can be used to shape policy.

III. Proposal

There are a few ways that the researchers could improve their work. First, through the addition of a log aggregation tool such as ElasticSearch or Splunk, they could better correlate the events between various systems together. This would help to reduce the errors caused by measuring indirect features. Additionally, they could test organizations in other industries, in order to determine if the risk profile changes.

References

[1] T.-F. Yen, V. Heorhiadi, A. Oprea, M. K. Reiter, and A. Juels, "An epidemiological study of malware encounters in a large enterprise," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 1117–1130. [Online]. Available: https://doi.org/10.1145/2660267.2660330