Cyber Attack Behaviors and a Malicious User’s Country of Origin

Samantha Roberts, M.S.

The George Washington University

**INTRODUCTION**

Phrases such as ‘hacker’ and ‘virus’ have become common, household terms; and can be translated into over 80 languages, providing shared relevance from the most rural to most populous areas on the planet (Katsev, 2018). It is represented in our entertainment, politics, economics, business affairs, and academia (Schutt & O’Neil, 2014). These concepts are so widespread, and the rate of technological growth is so rapid that it is imperative to develop new approaches for identifying and rebuffing cybersecurity attack advancements. Collectively, researchers and practitioners must share existing knowledge and innovate new discoveries that remain on pace with mal-intended tools and tactics.

Khosmood, Nico, & Woolery (2014) addressed the ability to identify an individual user based on his or her command line history in a Unix Operating System. The multi-pronged modeling approach these researchers used included the Naïve Bayes classifier, Decision Trees, and Maximum Entropy, or Multinomial Logistic Regression, models for comparison (Khosmood, Nico, & Woolery, 2014). Although they additionally used models related to language processing and analysis, the basis of their hypothesis was whether categorical variables related to a user’s behaviors could be used to predict authorship of command line executions. They ultimately were able to predict an individual based on the metadata collected 89% of the time. If researchers can accurately identify users based on syntactic patterns and behaviors, we may also be able to use a malicious user’s attack style to cultivate exploitable insight on a large scale. If successfully paired with predictive tools and algorithms, this could provide a novel approach to profiling attackers and security implementation procedures.

Another study looked at Android based technology targeted by malicious software. The researchers classified security package (permission) and metadata information from base Android phones in categorical terms, conducted several versions of feature selection, and then selected several classification algorithms to determine whether an interaction would be malicious or benign (Pehlivan, Baltaci, Acarturk, & Baykal, 2014). Among the models selected, Naïve Bayes classifier, Decision Trees, and Random Forest classifier were utilized. Their results indicated 89% and above accuracy for multiple models against various subsets of the overarching dataset to identify malicious/benign classifiers. This demonstrates the ability to use metadata of malicious attacks to successfully classify a target feature using popular classification models.

This project will review the data patterns and variable relationships of a recorded dataset from an AWS Honeypot from March 3, 2013 through September 8, 2013. The goal is to develop a model that will accurately classify which country or locale within a country an attacker is originating from based on multiple input features, including destination ports, protocol category, and host location data. The accurate prediction of geographical location may lead to other indicators of an attacker’s intent or motivations, such as time-related patterns in certain countries or even simple frequency of certain countries could be used to extrapolate meaning and further analysis about country, cultural behavior patterns, and an attacker’s habits.

**METHODOLOGY**

All coding and analysis was done using Python 3.6 Intel Distribution for Python (IDP) environment with the following packages: pandas, numpy, matplotlib, seaborn, sklearn, and scikit-learn. The dataset, Amazon Honeypot Attack Data, was collected from Kaggle.com (Casimian2000, 2018). Initial data processing included dropping several extraneous variables, removal of all NaN values, and then grouping by countries and removing all that had fewer than 10 samples or more than 20,000. This effectively removed two notable outliers, China and the United States, from the dataset.

Hyperparameter tuning was used next to find the strongest classification model with a variety of parameters specific to each model. Initial considerations for classification were Logistic Regression, Random Forest Classifier, Support Vector Machine (SVM), and Principle Component Analysis (PCA). The features used from the dataset for this project are all categorical in nature, and therefore, all the chosen models are well-known classification algorithms that inherently accept factor data types and do not require continuous data to function properly. A random sample of 100 was chosen from the greater dataset for hyperparameter tuning and parameter selection purposes. The sample size was chosen due to being large enough to change the output, but small enough to minimize run time. Upon receiving the results of the hyperparameter tuning model, the code dynamically selects the best scoring model with the associated parameter set and passed the full dataset through the predetermined pipeline. The results of this exercise demonstrate the accuracy scores of the model on train and test data that is sufficient to represent the dataset while still optimizing computational resources by utilizing the IDP environment.

**RESULTS**

The original dataset contained 451, 581 samples with 21 features. Several impractical features were removed immediately in order to simplify the dataset. The samples were reduced to 312,761 after removing all “not a number” (NaN) values from the dataset. Grouping operations reduced the dataset further to exclude countries with fewer than 10 samples or greater than 20,000 samples. This effectively reduced the dataset to a final 69,968 samples with 8 features. The features include event datetime information, host region location, protocol type, destination port, month (from datetime), count (frequency of country variable in the dataset), and geographical locale.

Initial exploratory data analysis efforts (EDA) demonstrated a few interesting trends and patterns. Country entries had as few as 10 samples (Cyprus, Ivory Coast) to as many as 11,944 (Iran), with a mean of 2,283 and standard deviation of 15,913. Population size and density by country and locale were not taken into consideration for this project, however, it may very likely be another illuminating piece in how to best understand the underlying relationship. The trend by month demonstrated there was an activity spike in the month of July during this rating period. The host region analysis revealed a large skewness, with 57% contributed by ‘groucho-tokyo’. A brief search on the internet returned the AWS 101 Event presentation being released July 16, 2018. It’s likely that the additional volume and diversity of attack data in July is correlated to the unveiling of Amazon Web Services (AWS). Lastly, the protocol split appeared to portray the most realistic representation, with a ratio of roughly 1.5 times the number of TCP interactions as UDP.

Most importantly, the results of this project were somewhat successful at providing support that cyber attack behaviors can be classified and predicted semi-accurately based on a few metadata features. The hyperparameter tuning model selected PCA with Logistic Regression as the most advantageous, with parameters specified as follows:

“Pipeline(memory=None, steps=[('StandardScaler', StandardScaler(copy=True, with\_mean=True, with\_std=True)), ('PCA', PCA(copy=True, iterated\_power='auto', n\_components=11, random\_state=0, svd\_solver='auto', tol=0.0, whiten=False)), ('clf', LogisticRegression(C=1, class\_weight=None, dual=False, fit\_intercept=True, intercept\_scaling=1, max\_iter=100, multi\_class='ovr', n\_jobs=None, penalty='l2', random\_state=0, solver='newton-cg', tol=0.0001, verbose=0, warm\_start=False))])”.

The initial selection returned a 35% predictive accuracy score. During the testing, several sample sizes were attempted, and SVC would occasionally win with a lower number of samples, such as 50. They are relatively close in measurement for this particular dataset, and in a practical setting, either model may be well-suited for use in a full evaluation.

The PCA and Logistic Regression model with the full data frame returned a predicted accuracy score of 40% when predicting the country variable against host, destination port, count, protocol type, and month (of datetime). Additional evaluations were run on locale (with country removed) and locale (with country included). The results of these were a slight drop in accuracy (42%) for locale when country is not provided, and a clear increase in predictive accuracy ( 55%) of locale when country is known.

**CONCLUSIONS**

Although a significant portion of the original dataset was removed prior to analysis, a successful model was found in this project that could, with some accuracy, predict the geographical location of a malicious attacker based on other metadata information routinely collected, particularly by intrusion detection systems used in conjunction with honeypots. This information could be very useful on a large scale when attempting to rebuff specific attack styles and protect sensitive information. One example that was found in this study is the potential increase in attacks due to the AWS 101 release in July 2013. A company interested in identifying it's weaknesses may want to identify company events, releases, or socioeconomic variables among others that could catch the intrigue of attackers.

Future research recommendations include focus on socioeconomic and political events that may cause motivated attackers and groups to surge attacks against a target. It would also be of interest to determine if there are regional cultural differences among attackers, such as styles that may differ between individualistic and collectivist cultures. This type of information could pave the way for more complex models to monitor and predict attacks based on known network and host metadata gathered coupled with web scraping of other patterns that might correlate in attack patterns, such as political events, weather patterns and natural disasters, terrorist activities, or economic stability. All of these factors could be influenced by the actual location of an individual attacker or attack group, and thus this project can be a stepping stone.

**REFERENCES**

Casimian2000. (2018). AWS Honeypot Attack Data. Retrieved from <https://www.kaggle.com/casimian2000/aws-honeypot-attack-data/home>

Katsev, I. (2018). How to Say Hacker in Different Languages. Retrieved October 3, 2018, from <https://www.indifferentlanguages.com/words/hacker>

Khosmood, F., Nico, P. L., & Woolery, J. (2014). User identification through command history analysis. 2014 IEEE Symposium on Computational Intelligence in Cyber Security (CICS). doi:10.1109/cicybs.2014.7013363

Pehlivan, U., Baltaci, N., Acarturk, C., & Baykal, N. (2014). The analysis of feature selection methods and classification algorithms in permission based Android malware detection. 2014 IEEE Symposium on Computational Intelligence in Cyber Security (CICS). doi:10.1109/cicybs.2014.7013371

Schutt, R., & O'Neil, C. (2014). Introduction: What Is Data Science? In Doing Data Science. Sebastopol, CA: O'Reilly Media.