Project Title:

Classification of Border Gateway Protocol Anomalies using Machine Learning

Shaji R.Nathan**

Executive summary

Border Gateway Protocol (BGP) is a standardized mechanism that powers the exchange of routing info between autonomous systems (AS) on the Internet. As separate networks need to peer with each other to form a global web, they promote its presence by communicating the routing information. This data is further stored in a routing information base (RIB).

RIB acts as a huge, constantly updating map that exists to path the way across a variety of destinations. BGP can access the RIB database, listing every possible route to deliver data and choosing the most efficient one. In case BGP fails, one network (Facebook in this case) can't advertise its presence, therefore, other networks can't reach it anymore. As a result, the affected network seems to be cut out of the Internet.Border Gateway Protocol (BGP) is an interdomain routing protocol designed to provide loop-free routing between separate routing domains that contain independent routing policies (autonomous systems). IP Infusion's software implementation of BGP version 4 includes multiprotocol extensions to allow BGP to carry routing information for IP multicast routes and multiple Layer 3 protocol address families including IP Version 4 (IPv4), IP Version 6 (IPv6), and Virtual Private Networks version 4 (VPNv4).

BGP is mainly used to connect a local network to an external network to gain access to the Internet or to connect to other organizations. When connecting to an external organization, external BGP (eBGP) peering sessions are created. BGP is referred to as an exterior gateway protocol (EGP). However, many existing networks within an organization are becoming complex in nature. So there has been widespread adoption of BGP to simplify the internal network used within the organization. BGP peers within the same organization exchange routing information through internal BGP (iBGP) peering sessions. When a TCP connection is established between peers, each BGP peer initially exchanges all its routes—the complete BGP routing table with the other peer. After this initial exchange, only incremental updates are sent when there has been a topology change in the network, or when a routing policy has been implemented or modified. In the periods of inactivity between these updates, peers exchange special messages called keepalives.

Current version of BGP is susceptible to Internet worms such as Slammer, Nimda, and Code Red I. These anomalies affect performance of the global Internet routing systems. BGP anomalies often also involve Internet Protocol (IP) prefix hijacks, operator mis-configurations, and routing system (switch hardware/network operating system) failures. Current advances in statistical and machine learning techniques can be used to classify and detect BGP routing anomalies.

Rationale
Why should anyone care about this question?

In short BGP allows the millions of networks (also known as autonomous systems or AS) worldwide that make up the global internet to reach each other. Just like the electric transmission grid, it almost always "just works," so internet users do not notice it — until it stops working or is abused. When one of those things happens, however, the effects can be severe and widely felt. Internet traffic between large swaths of the internet might break or bog down, taking large service providers or entire countries offline and affecting millions of companies and users.

Some recent examples of BGP in the news:
1. Facebook outage: https://engineering.fb.com/2021/10/05/networking-traffic/outage-details/
2. CodeRed Worm Infections in 24.0.0.0/8 on July 19, 2001, https://www.caida.org/catalog/software/walrus/examples/codered/
3. Nanog talk on Global Routing Instabilities During Code Red II and Nimda Worm Propagation, https://www.youtube.com/watch?v=ob_vxgEmC0s
4. https://www.forbes.com/sites/zakdoffman/2020/04/18/russia-and-china-behind-internet-hijack-risk-heres-how-to-check-youre-now-secure/?sh=79c04e865b16

Research Question
What are you trying to answer?

In this project I am looking to classify BGP updates on the provider edge routers in a typical 4G/5G backhaul network. A typical backhaul network will look like the one shown below. The end goal is to use a route reflector in conjunction with the classifier to analyze the BGP updates from the gateways and push a packet filter rule to block anomalous BGP traffic whenever the classifier detects an anomalous BGP update.

In this practical application, my goal is to compare the performance of the machine learning based classifiers namely K Nearest Neighbor, Logistic Regression, Decision Trees, and Support Vector Machines in detecting protocol anomalies in Border Gateway Routing Protocol.

Data Sources
What data will you use to answer you question?

I will utilize Border Gateway Protocol (BGP) datasets with routing records collected from Reseaux IP Europeens (RIPE) and BCNET to implement an anomaly detection model. For the supervised learning datasets from five well-known anomalies are used. They are namely WannaCrypt, Moscow blackout, Slammer, Nimda, Code Red I, that occurred in May 2017, May 2005, January 2003, September 2001, and July 2001. The datasets were made publicly available by Reseaux IP Europeens (RIPE) Network Coordination

Centre(NCC):[RIPE]http://www.ensc.sfu.ca/~ljilja/cnl/projects/BGP_datasets/BGP_RIPE_datasets_for_anomaly_detection_csv_revised_19022021.zip)__.

The regular data is collected from IP Infusion's BGP gateways. Live BGP updates are also available from the Center for Applied Internet Data Analysis at UC San Diego. [CAIDA]https://bgpstream.caida.org/data

Methodology
What methods are you using to answer the question?

In this project we first used basic exploratory data analysis to get a preliminary understanding of the RIPE BGP announcement dataset. Based on the analysis, data was scaled, and cleaned for the classification algorithms. In the initial base model building phase using Logistic Regression the data was left unbalanced. Datasets also contained known anomalies from actual Internet outage events while building this supervised model. Then this multiple Models were then built using KNN, Logistic Regression, SVM and Random Forest. The performance characteristics of each model was analyzed. In the next phase GridSearch and HalvingGridSearch was used to tune the hyperparameters. Based on this the best model using SVM was developed. The final model used balanced data, and this was achieved using the SMOTE library. The feature importance for the final model was derived. The model predictions were tested against the new BGP announcements from RIPE live feeds and IP Infusion’s internal BGP feeds to check the viability of this model for classifying

Results
What did your research find?

Findings and Actionable Insights:
The latest model based on GridSearchCV is pretty accurate. It got one prediction wrong which is in line with the accuracy rate on test data and ROC characterestics**

1. The objective of this capstone project was to come up with an optimum classification model to predict whether a bgp update message can be classified as anomalous or normal based on the update message attributes.
2. I analyzed 37 numerical variables which capture the various attributed of a typical BGP protocol update to build the model.
3. Exploratory data analysis showed absence of null values in the dataset, and the data is imbalanced, where “1” anomalous message is the majority class.
4. Univariate analysis revealed that Average_AS_Path length and Number of Implicit withdrawals does not help very much when it comes to predicting the target variable. Some numerical features tend to predict the target variable much better (for example: ['Maximum AS-path length', 'Number of duplicate announcements', 'Maximum edit distance', 'Number of Exterior Gateway Protocol (EGP) packets'] etc.)
5. Dataset preprocessing of numerical data was done using standscaler and MinMax scaler
6. Basic models were built using K Nearest Neighbor, Logistic Regression, Decision Trees, and Support Vector Machines.

7. The most important features in predicting whether a BGP update is anomalous based on the Support Vector model was
['Hour and Minutes', 'Hour', 'Minutes', 'Seconds',
'Number of announcements', 'Number of withdrawals',
'Number of announced NLRI prefixes',
'Number of withdrawn NLRI prefixes', 'Average AS-path length',
'Maximum AS-path length', 'Average unique AS-path length',
'Number of duplicate announcements',
'Number of duplicate withdrawals',
'Number of implicit withdrawals']

8. GridSearch and Halving GridSearch was used to find the best parameters. The best parameters derived from Gridsearch were as follows:

{'C': 1, 'gamma': 0.0001, 'kernel': 'rbf'}
Classification results show that the best achieved F-Scores of the SVM models are 86.1%. Support Vector model gave the best performance with Halving GridSearchCV and the best test AUC was 0.86 which was similar to the results reported by researchers at CAIDA, RIPE, and Simon Fraser University in previous research reports using a similar dataset.

Next Steps in the Classifier Model Enhancement:

1. Fine Tuning the model based on domain knowledge and feature importance results
2. We could reduce the dimensions/features further to tune the model.
3. Support Vector Model training even with HalvingGridSearch was very slow.
4. There are some new libraries like T-POT https://epistasislab.github.io/tpot/using/ that use genetic algorithms for hyperparameter tuning to derive the best pipeline for this classification problem. For best feature selection we could use a library like YellowBrick https://www.scikit-yb.org/en/latest/api/model_selection/importances.html
5. Ensemble models, XGBoost could be used as next step in improving the performance of the current SVM model
6. Neural Networks/Autoencoders and LSTM based model seems to be well suited for this class of problems.

Outline of project

- [Link to notebook 1]()
- [Link to notebook 2]()
- [Link to notebook 3]()

Contact and Further Information

e-mail: srnathan@pacbell.net, shaji.nathan@ipinfusion.com
Phone: (408) 400-1503
IP Infusion Inc.
3965 Freedom Cir #200, Santa Clara, CA 95054