BGP Dataset Generation and Feature Extraction for Anomaly Detection

（数据集生成文章，只做记录不做详细分析）

Introduction

Info

Authors: Paulo Fonseca, Edjard S. Mota, Ricardo Bennesby, Alexandre Passito. ISCC 2019.

Code: https://github.com/ufam-lia/bgp-feature-extractor

Motivation

Existing public datasets of BGP messages are not suited to feed ML models directly, because the goal of these messages is to exchange fine-grained reachability information instead of observe network behavior.

Contribution

1. Publicly available BGP feature extraction tool that allows easy integration of new network features.
2. Publicly available datasets of network features during BGP anomalies of different classes, which can be used to train customized machine learning models.
3. Proposal and evaluation of novel features that capture the relationship of complementary messages and allows easier analysis of anomalous behavior.
4. Analysis of trends in BGP behavior that can be used to distinguish regular traffic from anomalies and different types of anomalies.

Background

Four categories of BGP anomalies: Direct intended, Direct unintended, Indirect anomalies, Network failures.

(need to read Part II carefully)

Features Extraction

Volume features：

- Number of announcements/withdrawals
- Number of duplicate announcements/withdrawals
- Number of non-duplicate announcements
- Number of flaps
- Number of new announcements after withdraw
- Number of plain new announcements
- Number of implicit withdrawals with same/different path
- Number of IGP/EGP/INCOMPLETE messages
- Number of ORIGIN changes
- Number of announced prefixes
- Maximum/average announcements per prefix
- Maximum/average announcements per AS

AS path features:

- Announcements to longer/shorter paths
- Average/maximum AS path length
- Average/maximum unique AS path length
- Average/maximum edit distance
- Edit distance with k value
- Edit distance with k value (unique)
- Number of rare ASes
- Average/maximum number of rare ASes

Distribution features:

- Announcements vs withdrawals
- Origin types
- Announcements types
- Longer vs Shorter
- Implicit withdrawals vs explicit withdrawals
- Withdrawals w/ same path vs withdrawals w/ different path

Methodology

Consider anomalous events from three different types: Direct, Indirect, Outages.

1. Search for event reports that indicated approximate times of start and end of the anomaly, as well as which ASes were involved in the anomaly event.
2. Make a list of point-of-view candidates considering ASes and collectors that are close to the anomaly origin.
3. Download BGP raw data from public repositories, of the collectors identified in the last step.
4. Implement a feature extraction mechanism and generate timeseries features from all candidates considering a period that comprises the anomaly, as well as regular traffic before and after the anomaly.
5. Analyze the extracted series in order to observe trends that matched those reported (e.g. increase in number of announcements around a specific time of day).
6. Label the identified period as an anomaly event.

Anomaly and regular traffic

Features that showed significant changes in their behavior during the anomalies in most of generated datasets:

Announcements, Withdrawals, Origin changes, Rare ASes average, Distribution - New announcements

Direct, indirect and outage anomalies

Features that can be used to leverage anomaly classification and enforce mitigation techniques to each type of anomaly.

Outages

Number of rare ASes, Distribution - Announcements vsWithdrawals, Distribution - ExplicitWithdrawals and Flaps, Distribution - Longer paths and Shorter paths。

Direct

Distribution - Shorter paths

Indirect

Distribution - Implicit Withdrawals

Conclusion

- Implemented a dataset generation tool that converts raw BGP control plane messages to timeseries of features.
- Proposed novel features that allows intuitive observation of BGP behavior trends