UNITECH 2014 GABROVO
INTERNATIONAL SCIENTIFIC CONFERENCE
21 – 22 November 2014, GABROVO
TU GABROVO 50 YEARS

USING DATABASES FOR BGP DATA ANALYSIS

Marijana Ćosović, Slobodan Obradović
Faculty of Electrical Engineering, East Sarajevo, Bosnia and Herzegovina

Ljiljana Trajković
Simon Fraser University, Vancouver, Canada

Abstract

Boarder Gateway Protocol is an Exterior Gateway Protocol used between Autonomous Systems (ASes) to send update information upon changes in the network topology. Network reachability information is contained within BGP update messages. Recent trends in BGP anomaly detection systems employ machine learning techniques to mine network data. In the proposed approach, we consider diversity of anomalous events and the volume of BGP data to be processed. Creating efficient environment to access data is useful in collaborative research projects. We collect raw data, import BGP update messages into a database, issue appropriate SQL queries to extract features, and analyze query results. Obtained data may be used for machine learning modeling and development of BGP anomaly detection systems.

Keywords: Border gateway protocol, feature extraction, database, machine learning

INTRODUCTION

Data travels across Internet in packets and routers determine the route a data packet takes through the Internet. Routers need to decide how to forward a received data packet and, hence, they need to exchange information about reachability of possible destinations. A physical interconnection of various independently administrated networking regions called Autonomous Systems facilitates the Internet operations such as connectivity and data packet delivery. The Border Gateway Protocol (BGP) is a routing protocol that manages forwarding of IP traffic between the source and the destination Autonomous Systems [1]. The Internet is a critical asset [2] of information and communications technology. Hence, recognition and detection of BGP instabilities is of significant interest.

Machine learning techniques have been recently employed in designing BGP anomaly detection systems [3]-[5]. A first step in the process is obtaining the input data for modeling using feature extraction process. In this paper, we use a database for feature extraction. As an alternative to analyzing raw data directly, we import raw data into database, issue SQL queries, and extract information. Creating an efficient data accessing environment allows users to access and analyze data from remote access points. It is also beneficial to extract statistics from a large volume of data already stored in the database.

Feature selection is a process of selecting a subset of original features according to a certain criteria. It is a frequently used technique for data preprocessing to reduce problem dimension in machine learning. It reduces the number of features and removes irrelevant, redundant, and noisy features [6]. Feature selection follows a feature extraction process.

BGP UPDATE MESSAGES

BGP is an incremental protocol that sends updates only if there are changes of reachability or topology within the network. BGP routers exchange four types of messages: 'Open', 'Update', 'Notification', and 'Keep Alive'. 'Open' message that contains basic information such as router identifier, BGP version, and the AS number is used to open up peering session. Notification message closes down peering session if there is a disagreement in the configuration parameters. Once the BGP session is established, routers exchange all known routes using the 'Update' message

shown in Fig.1, and after that only when there is a change of BGP routes in the routing tables.

```
TIME: 01/25/03 15:45:53
TYPE: BGP4MP/MESSAGE/Update
FROM: 192.65.184.3 AS513
TO: 192.65.185.40 AS12654
ORIGIN: IGP
ASPATH: 513 3320 209 16738
NEXT_HOP: 192.65.185.4
ANNOUNCE
  198.3.128.0/24
  204.255.70.0/24
```

Fig. 1. *BGP update message*

'Keep Alive' messages are exchanged between peers during inactivity periods to make sure that the connection still exists.

Processing of BGP update messages

The Réseaux IP Européens Network Coordination Centre (RIPE NCC) collects and stores Internet routing data through the Routing Information Service (RIS) project [7]. BGP update messages are collected by the Remote Route Collectors (RRCs) and stored in multi-threaded routing toolkit (MRT) format [8], [9]. These messages were collected every fifteen minutes until July of 2003 after which a five minute interval between the consecutive BGP messages has been adopted.

In this paper, we use BGP update messages that originated from AS 513 (route collector rcc04). Only data collected during the periods of Internet anomalies are considered. BGP update messages are transformed from MRT into ASCII format by using *bgpdump* library on Linux platform. A *bash* script was used to process data files in batches. Bgpdump is a C library, maintained by RIPE Network Coordination Centre, used to analyze dump files produced by MRT. Furthermore, concatenation of all the messages is performed to optimize loading the database.

SQL loader utility is used to load BGP update messages into tables of an Oracle database. All BGP update messages are imported sequentially into the database. Feature statistics are computed every minute during five-day periods for three well known attacks on the Internet: Slammer, Nimda, and Code-Red I worms. By querying the database and accessing the BGP update messages, we generated the following volume features.

1. *Number of announcements* is a number of routes that are available for delivery of the data while *number of withdrawals* is the number of the routes that are no longer reachable. Number of announcements SQL query is shown in Fig. 2. Number of announcements during the Slammer worm attack is shown in Fig. 3.

```
select substr(d.txt_line,1,20)time,
count(a.id)no_ann
from maja_day1 a ,maja_v_day1_dan d
,maja_v_day1_from f,maja_v_day1_tip t
where a.id=d.id and a.id=f.id and
a.id=t.id
and upper(a.txt_line) like '%ANNOUNCE%'
group by substr(d.txt_line,1,20)
```

Fig. 2. *Number of announcements SQL query*

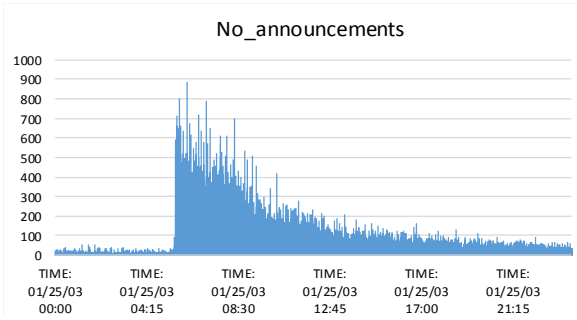

Fig. 3. *Number of announcements during Slammer worm attack*

2. *Number of announced and withdrawn NLRI prefixes* is the number of announced/withdrawn Network Layer Reachability Information *(NLRI) pr*efixes within BGP update messages that have type field set to announcement /withdrawal during one minute interval.
   Number of duplicate announcements and withdrawals is a number of duplicate BGP update messages that have type field set to announcement /withdrawal during one minute interval.
3. *Number of implicit withdrawals* is the number of BGP update messages that have type field set to announcement and different AS-PATH attribute for already announced NLRI prefixes during one minute interval.

4. *Number of IGP, EGP and incomplete packets* is the number of BGP update messages that are generated by Interior Gateway Protocol (IGP), Exterior Gateway Protocol (EGP), and of unknown sources during one minute interval. Number of EGP packets SQL query is shown in Fig. 4. Number of EGP packets during Slammer worm attack is shown in Fig. 5.

```
select time,sum(egp) from (
select substr(d.txt_line,1,20)
time,count(a.id) egp
from maja_day1 a ,maja_v_day1_dan d
,maja_v_day1_from f,maja_v_day1_tip t
where a.id=d.id
and a.id=f.id
and a.id=t.id
and upper(a.txt_line) like '%EGP%'
group by substr(d.txt_line,1,20)
union
select substr(v.txt_line,1,20),sum(0)
from maja_day1 a,maja_v_day1_dan v
where a.id=v.id
and substr(v.txt_line,1,20)!='TIME:
01/23/03 17:|&'
group by substr(v.txt_line,1,20))
group by time
order by 1
```

Fig. 4. *Number of EGP packets*

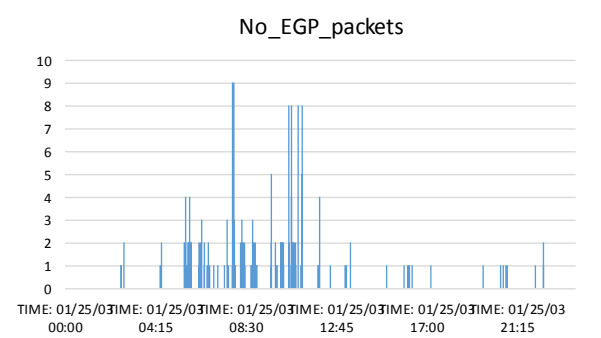

Fig. 5. *Number of EGP packets during Slammer worm attac.*

AS-PATH features are derived from AS-PATH attribute of a BGP update message. Parsing the ASCII files and feature statistics generation is performed using data filtering in the database. Additional views are created in database in order to obtain various requests. More complex tasks have required writing PL/SQL code. AS-PATH features and their respective SQL queries are:

1. *Average AS-PATH length* is the average length of AS-PATHs of all messages during one minute interval.
2. *Maximum AS-PATH length* is the maximum length of AS-PATHs of all messages during one minute interval. The maximum AS-PATH length SQL query is shown in Fig. 6. The maximum AS-PATH length during Slammer worm attack is shown in Fig. 7.

```
select substr(d.txt_line,1,20)
time,max(duzina_as) from
maja_day1_aspath asp,
maja_v_day1_dan d
where asp.id=d.id
group by substr(d.txt_line,1,20)
order by 1
```

Fig. 6. *Maximum AS-PATH length SQL query*

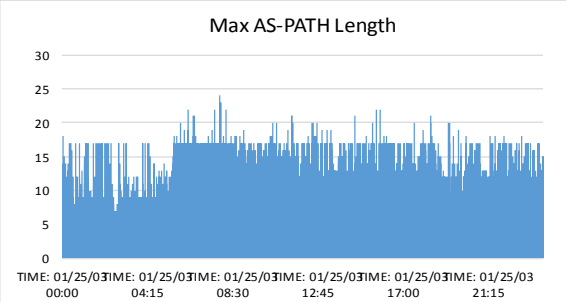

Fig. 7. *Maximum AS-PATH length during Slammer worm attack*

3. *Average unique AS-PATH length* is the average of unique length of AS-PATHs of all messages during one minute interval.
4. *Average edit distance* is the average of edit distances among all the messages during one minute interval. The average edit distance SQL query is shown in Fig. 8 Average edit distance during Slammer worm attack is shown in Fig. 9.

```
select substr(d.txt_line,1,20)
time,round(avg(ed),0) avg_ed
from maja_day3_ed_ld a,
maja_v_day3_dan d
where a.id=d.id
group by substr(d.txt_line,1,20)
```

Fig. 8. *Average edit distance SQL query*

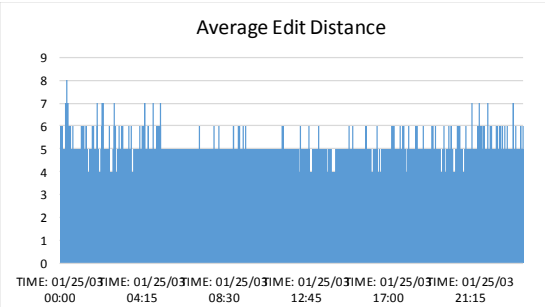

Fig. 9. *Average edit distance during Slammer worm attack*

5. *Maximum edit distance* is the average of edit distances among all the messages during one minute interval. The maximum edit distance SQL query is shown in Fig. 10. The maximum edit distance during Slammer worm attack is shown in Fig. 11.

```
select substr(d.txt_line,1,20)
time,max(ed) max_ed
from maja_day3_ed_ld a,maja_v_day3_dan d
where a.id=d.id
group by substr(d.txt_line,1,20)
```

Fig. 10. *Maximum edit distance SQL query*

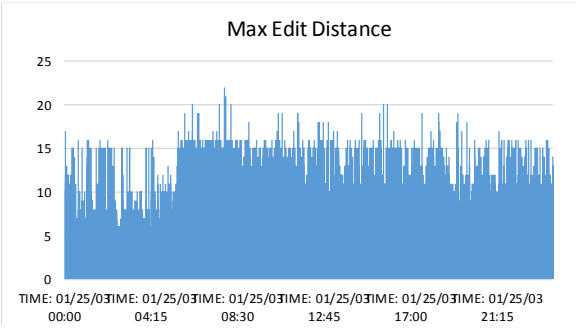

Fig. 11. *Maximum edit distance during Slammer worm attack*

Function created for calculating Levenshtein distance [10] has been imported into the database. Levenshtein distance is a measure of the similarity between two strings. The distance is the minimal number of deletions, insertions, or substitutions required to transform one string into another. PL/SQL code was written to use the function and have AS-PATH attributes in a form that could be used as input to Levenshtein function. SQL queries calculated the average and maximum value of edit distance per one minute.

CONCLUSION

We have created a database to collect, process, and extract features from known sources of BGP anomalies such as Slammer, Nimda, and Code Red I worms. Database may be used for creating friendly user environment for extracting information by issuing queries. This study may be extended to include other BGP anomalies. After feature selection process, extracted features may be used as input to machine learning algorithms because using BGP volume features is a feasible approach for detecting possible worm attacks.

REFERENCE

[1] Y. Rekhter and T. Li, "A border gateway protocol 4 (bgp-4)," IETF RFC 1771, Mar. 1995.

[2] European Union Agency for Network and Information Security [Online]. Available: http://www.enisa.europa.eu/activities/Resilience-and-CIIP.

[3] N. Al-Rousan and Lj. Trajkovic, "Machine learning models for classification of BGP anomalies," Proc. IEEE Conf. High Performance Switching and Routing, HPSR 2012, Belgrade, Serbia, June 2012, pp. 103-108.

[4] N. Al-Rousan, S. Haeri, and Lj. Trajkovic, "Feature Selection for Classification of BGP Anomalies using Bayes Models," ICMLC 2012, July 2012, Xi'an, China.

[5] Y. Li, H. J. Xing, Q. Hua, X.-Z. Wang, P. Batta, S. Haeri, and Lj. Trajkovic, "Classification of BGP anomalies using decision trees and fuzzy rough sets," in Proc. IEEE International Conference on Systems, Man, and Cybernetics (SMC 2014), San Diego, CA, October 2014, pp. 1331-1336.

[6] H. Liu, H. Motoda, Eds. Computational Methods of Feature Selection, Chapman and Hall/CRC Press, 2007.

[7] Ripe RIS raw data [Online]. Available: http://www.ripe.net/data-tools/stats/ris/ris-raw-data.

[8] T. Manderson, "Multi-threaded routing toolkit (MRT) Border Gateway Protocol (BGP) routing information export format with geo-location extensions," RFC 6397, IETF, Oct. 2011.

[9] MRT rooting information export format [Online]. Available: http://tools.ietf.org/html/draft-ietf-grow-mrt-13.

[10] Levenshtein algorithm PL/SQL [Online]. Available: http://richmurnane.blogspot.com/2006/02/levenshtein-distance-algorithm-oracle.html.