Application of Machine Learning Techniques to Detecting Anomalies in Communication Networks: Datasets and Feature Selection Algorithms

Qingye Ding, Zhida Li, Soroush Haeri, and Ljiljana Trajković

Abstract Detecting, analyzing, and defending against cyber threats is an important topic in cyber security. Applying machine learning techniques to detect such threats has received considerable attention in research literature. Anomalies of Border Gateway Protocol (BGP) affect network operations and their detection is of interest to researchers and practitioners. In this Chapter, we describe main properties of the protocol and datasets that contain BGP records collected from various public and private domain repositories such as Route Views, Réseaux IP Européens (RIPE), and BCNET. We employ various feature selection algorithms to extract the most relevant features that are later used to classify BGP anomalies.

Keywords Routing anomalies · Border Gateway Protocol · Feature extraction · Feature selection · Machine learning techniques

1 Introduction

The Internet is a critical asset of information and communication technology. Cyber attacks and threats significantly impact the Internet performance. Hence, detecting such network anomalies is of great interest to researchers and practitioners. In this Chapter, we describe BGP, datasets used to detect anomalies, feature extraction process, and various feature selection algorithms. We consider BGP update messages because they contain information about the protocol status and configurations. BGP update messages are extracted from the collected data during the time periods when the Internet experienced known anomalies. BGP features are extracted and selected in order to improve the classification results. The classifiers are then used in Chapter 4 to detect anomalies and to compare classification results.

Q. Ding · Z. Li · S. Haeri · L. Trajković (✉)
Simon Fraser University, Vancouver, BC, Canada
e-mail: qingyed@sfu.ca; zhidal@sfu.ca; shaeri@sfu.ca; ljilja@cs.sfu.ca

A. Dehghantanha et al. (eds.), *Cyber Threat Intelligence*, Advances in Information Security 70, https://doi.org/10.1007/978-3-319-73951-9_3

We extract *AS-path* and *volume* features from the BGP datasets [1, 9, 11]. Subsets of features are then selected using various feature selection algorithms to reduce the dimensionality of the dataset matrix while preserving their physical meaning. The employed algorithms belong to the category of filter methods where feature selection is independent of the underlying learning algorithm [28]. For feature selection, we employ: Fisher [35], minimum redundancy maximum relevance (mRMR) [40] (including mutual information difference (MID), mutual information quotient (MIQ), and mutual information base (MIBASE)), odds ratio (OR), extended/weighted/multi-class odds ratio (EOR/WOR/MOR), class discriminating measure (CDM) [21], and decision tree [42].

In this survey, we have revised our previous research findings and results by carefully processing the considered datasets, selecting better parameters for various techniques, and reevaluating past performance results. We revised previously reported performance results [16, 24] for feature selection algorithms (Fisher, MID, MIQ, MIBASE, OR, EOR, WOR, MOR, and CDM). Other approaches include using fuzzy rough sets for feature selection [34]. Fuzzy sets and rough sets [38, 43] have greatly affected the way we compute with imperfect information. Fuzzy rough sets deal with the approximation of fuzzy sets in an approximation space [52]. Even though the classification accuracy usually improves by performing feature reduction using fuzzy rough sets, the computational complexity of the algorithm remains rather high. Hence, this approach is unsuitable in cases with large number of samples and attributes. In comparison, the decision tree algorithm is faster and may achieve acceptable classification accuracy.

This Chapter is organized as follows. We first briefly described BGP, the effect of network anomalies, and approaches for their detection. In Sect. 2, we provide details of various BGP anomalies that we have considered in our previous work and in this survey. The description of the datasets and data processing is introduced in Sect. 3. Various approaches for feature extractions are described in Sect. 4 while feature selection algorithms are described in Sect. 5. We conclude with Sect. 6. List of relevant references is also provided.

1.1 Border Gateway Protocol (BGP)

BGP [44] is a routing protocol that plays an essential role in forwarding Internet Protocol (IP) traffic between the source and the destination Autonomous Systems (ASes). An AS is a collection of BGP peers managed by a single administrative domain [51]. It consists of one or more networks that possess uniform routing policies while operating independently. Internet operations such as connectivity and data packet delivery are facilitated by various ASes.

The main function of BGP is to select the best routes between ASes based on network policies enforced by network administrators. Routing algorithms determine the route that a data packet takes while traversing the Internet. They exchange reachability information about possible destinations. BGP is an upgrade of the

Exterior Gateway Protocol (EGP) [45]. It is an interdomain routing protocol used for routing packets in networks consisting of a large number of ASs. BGP version 4 allows Classless Interdomain Routing (CIDR), aggregation of routes, incremental additions, better filtering options, and it has the ability to set routing policies. BGP employs the path vector protocol, which is a modified version of the distance vector protocol [30]. It is a standard for the exchange of information among the Internet Service Providers (ISPs).

BGP relies on the Transport Control Protocol (TCP) to establish a connection (port 179) between the routers. A BGP router establishes a TCP connection with its peers that reside in different ASes. Because of their size, BGP routing tables are exchanged once between the peering routers when they first connect. BGP allows ASes to exchange reachability information with peering ASes to transmit information about the availability of routes within an AS. Based on the exchanged information and routing policies, it determines the most appropriate path to destination. BGP allows each subnet to announce its existence to the Internet and to publish its reachability information. Hence, all sub-networks are inter-connected and are known to the Internet.

BGP is an incremental protocol that sends updates only if there are reachability or topology changes within the network. Afterwards, only updates regarding new prefixes or withdrawals of the existing prefixes are exchanged. BGP routers exchange four types of messages: *open*, *update*, *keep-alive*, and *notification* [45]. The *open* message that contains basic information such as router identifier, BGP version, and the AS number is used to open a peering session. Routers exchange all known routes using the *update* message after the BGP session is established and when there is a change of BGP routes in the routing tables. *Keep-alive* messages are exchanged between peers during inactivity periods to ensure that the connections still exist. The *notification* message closes a peering session if there is a disagreement in the configuration parameters. A sample of a BGP *update* message is shown in Table 1. It contains two Network Layer Reachability Information (NLRI) announcements, which share attributes such as the *AS-path*. The *AS-path* attribute in the BGP update message indicates the path that a BGP packet traverses among AS peers. The *AS-path* attribute enables BGP to route packets via the best path.

Propagation of the BGP routing information is susceptible to various anomalous events such as worms, malicious attacks, power outages, blackouts, and misconfigurations of BGP routers. BGP anomalies are caused by changes in network topologies, updated AS policies, or router misconfigurations. They affect the Internet servers and hosts and are manifested by anomalous traffic behavior. Anomalous events in communication networks cause traffic behavior to deviate from its usual profile. These events may spread false routing information throughout the Internet by either dropping packets or directing traffic through unauthorized ASes and, hence, risking eavesdropping. Large-scale power outages may affect ISPs due to unreliable power backup. They could also cause network equipment failures leaving affected networks isolated and their service disrupted. Configuration errors in BGP routers also induce anomalous routing behavior. Routing table leak and prefix hijack [10] events are examples of BGP configuration errors that may lead

Table 1 Sample of a BGP update message

Field	Value
Time	2003 1 24 00:39:53
Type	BGP4MP/BGP4MP_MESSAGE AFI_IP
From	192.65.184.3
To	193.0.4.28
BGP packet type	Update
Origin	IGP
AS-path	513 3320 7176 15570 7246 7246 7246 7246 7246 7246 7246 7246 7246
Next-hop	192.65.184.3
Announced NLRI prefix	198.155.189.0/24
Announced NLRI prefix	198.155.241.0/24

IGP: Interior Gateway Protocol, NLRI: Network Layer Reachability Information

to large-scale disconnections in the Internet. A routing table leak occurs when an AS such as an ISP announces a prefix from its Route Information Base (RIB) that violates previously agreed upon routing policy. A prefix hijack is the consequence of an AS originating a prefix that it does not own.

1.2 Approaches for Detecting Network Anomalies

Detailed comparison of various network intrusion techniques has been reported in the literature [18]. Demands for Internet services have been steadily increasing and anomalous events and their effects have dire economic consequences. Determining the anomalous events and their causes is an important step in assessing loss of data by anomalous routing. Hence, it is important to classify these anomalous events and prevent their effects on BGP.

Anomaly detection techniques have been applied in communication networks [14]. These techniques are employed to detect BGP anomalies such as intrusion attacks, worms, and distributed denial of service attacks (DDoS) [32, 39] that frequently affect the Internet and its applications. BGP data have been analyzed to identify anomalous events and design tools that have been used in anomaly predictions. Network anomalies are detected by analyzing collected traffic data and generating various classification models. A variety of techniques have been proposed to detect BGP anomalies.

Early approaches include developing traffic models using statistical signal processing techniques where a baseline profile of network regular operation is developed based on a parametric model of traffic behavior and a large collection of traffic samples to account for regular (anomaly-free) cases [27]. Anomalies may then be detected as sudden changes in the mean values of variables describing

the baseline model. However, it is infeasible to acquire datasets that include all possible cases. In a network with quasi-stationary traffic, statistical signal processing methods have been employed to detect anomalies as correlated abrupt changes in network traffic [47].

The main focus of approaches also proposed in the past is developing models for classification of anomalies. The accuracy of a classifier depends on the extracted features, combination of selected features, and underlying models. Recent research reports describe a number of applicable classification techniques. One of the most common approaches is based on a statistical pattern recognition model implemented as an anomaly classifier and detector [23]. Its main disadvantage is the difficulty in estimating distributions of higher dimensions. For example, a Bayesian detection algorithm was designed to identify unexpected route mis-configurations as statistical anomalies [17]. An instance-learning framework also employed wavelets to systematically identify anomalous BGP route advertisements [53]. Other proposed techniques are rule-based methods that have been employed for detecting anomalous BGP events. An example is the Internet Routing Forensics (IRF) that was applied to classify anomaly events [33]. However, rule-based techniques are not adaptable learning mechanisms. They are slow, have high degree of computational complexity, and require a priority knowledge of network conditions.

We view anomaly detection as a classification problem of assigning an "anomaly" or "regular" label to a data point. There are numerous machine learning methods that address these classification tasks. However, redundancies in the collected data may affect the performance of classification methods. Feature extraction and selection are used to select a subset of features from the original feature space and, thus, to reduce redundancy among features that leads to improving the classification accuracy. Feature extraction methods such as principal component analysis project the original data points onto a lower dimensional space. However, features transformed by feature extraction lose their original physical meaning. We extract BGP features based on the attributes of BGP update messages in order to achieve reliable classification results. Recent trends in designing BGP anomaly detection systems rely more frequently on machine learning techniques. Known classifiers are tested for their ability to detect network anomalies in datasets that include known BGP anomalies. In this survey, we described several machine learning techniques that we have used in the past for classification due to their superior performance compared to earlier approaches.

2 Examples of BGP Anomalies

Anomalous events considered in this Chapter are worms, power outages, and BGP router configuration errors. They are manifested by sharp and sustained increases in the number of announcement or withdrawal messages exchanged by BGP routers. *Volume* and *AS-path* features are collected over 1-min time intervals during 5-day periods for well known anomalous Internet events. While the available datasets

Table 2 Examples of known BGP Internet worms

Dataset	Class	Date		Duration (min)
		Beginning of the event	End of the event	
Slammer	Anomaly	25.01.2003 at 5:31 GMT	25.01.2003 at 19:59 GMT	869
Nimda	Anomaly	18.09.2001 at 13:19 GMT	20.09.2001 at 23:59 GMT	3,521
Code Red I	Anomaly	19.07.2001 at 13:20 GMT	19.07.2001 at 23:19 GMT	600

Table 3 Datasets of the Internet anomalous events

Event	Date	RRC	Peers
Moscow power blackout	May 2005	RIS 05	AS 1853, AS 12793, AS 13237
AS 9121 routing table leak	Dec. 2004	RIS 05	AS 1853, AS 12793, AS 13237
AS 3561 improper filtering	Apr. 2001	RIS 03	AS 3257, AS 3333, AS 286
Panix domain hijack	Jan. 2006	Route Views	AS 12956, AS 6762, AS 6939, AS 3549
AS-path error	Oct. 2001	RIS 03	AS 3257, AS 3333, AS 6762, AS 9057
AS 3356/AS 714 de-peering	Oct. 2005	RIS 01	AS 13237, AS 8342, AS 5511, AS 16034

contain data over much longer periods of time, we have selected for our analysis a 5-day period to minimize storage and computational requirements. Furthermore, selecting longer periods of regular data would make datasets ever more unbalanced. Several methods that we surveyed offer better performance when dealing with balanced datasets. Details including dates of the events, remote route collectors (RRC) that acquired data using Routing Information Service (RIS), and observed peers are given in Tables 2 and 3. For example, Slammer event occurred on January 25, 2003 and lasted almost 16 hours. Hence, BGP update messages collected between January 23, 2003 and January 27, 2003 are selected as samples for feature extraction.

The Structured Query Language (SQL) Slammer worm attacked Microsoft SQL servers on January 25, 2003 [12]. It generated random IP addresses and replicates itself by sending 376 bytes of code to those IP addresses. As a result, the update messages consumed most of the routers' bandwidth, which in turn slowed down the routers and, in some cases, caused the routers to crash. The Nimda worm [8] was released on September 18, 2001. It propagated fast through email messages, web browsers, and file systems. Viewing the email message triggered the worm payload. The worm modified the content of the web document file in the infected hosts and copied itself in local host directories. The Code Red I worm attacked web servers on July 19, 2001 [3]. The worm affected approximately half a million IP addresses a day. It took advantage of vulnerability in the Internet Information Services (IIS) indexing software. It triggered a buffer overflow in the infected hosts by writing to the buffers without checking their limits.

We consider BGP anomalous events such as Slammer [12], Nimda [8], Code Red I [3], AS 9121 routing table leak [41], Moscow power blackout, AS 3561 improper filtering, Panix domain hijack, AS path error, and AS 3356/AS 714 de-peering [54].

Slammer [12] Microsoft SQL servers were infected through a small piece of code that generated IP addresses at random. Furthermore, code replicated itself by infecting new machines through randomly generated targets. If the destination IP address was a Microsoft SQL server or a user's PC with the Microsoft SQL Server Data Engine (MSDE) installed, the server became infected and began infecting other servers. The number of infected machines doubled approximately every 9 s. Single infected machines have reported additional traffic of 50 Mb/s [13] as a consequence of increased generation of update messages.

Nimda [8] Nimda exploited vulnerabilities in the Microsoft Internet Information Services (IIS) web servers for the Internet Explorer 5. It used three methods for propagation: email, network shares, and the web. The worm propagated by sending an infected attachment that was automatically downloaded after viewing email. A user could also download it from the website or access an infected file through the network.

Code Red I [3] Although the Code Red I worm attacked Microsoft IIS web servers earlier, the peak of infected computers was observed on July 19, 2001. The worm replicated itself by exploiting weakness of the IIS servers and, unlike the Slammer worm, Code Red I searched for vulnerable servers to infect. Rate of infection was doubling every 37 min.

Records of three BGP anomalies along with regular RIPE traffic are shown in Fig. 1. The effect of Slammer worm on *volume* and *AS-path* features is shown in Fig. 2.

Moscow Power Blackout The blackout occurred on May 25, 2005 and lasted several hours. The Moscow Internet exchange was shut down during the power outage. Routing instabilities were observed due to loss of connectivity of some ISPs peering at this exchange. This effect was apparent at the RIS remote route collector in Vienna (rrc05) through a surge in announcement messages arriving from peer AS 12793, as shown in Fig. 3. Hence, volume of announcements was one of the features used to detect the anomaly.

AS 9121 Routing Table Leak It occurred on December 24, 2004 when AS 9121 announced to peers that it could be used to reach almost 70% of all prefixes (over 106,000). As a consequence, numerous networks had either misdirected or lost their traffic. The AS 9121 started announcing prefixes to peers around 9:20 GMT and the event lasted until shortly after 10:00 GMT. It continued to announce bad prefixes throughout the day. The announcement rate reached the second peak at 19:47 GMT.

AS 3561 Improper Filtering This was a BGP mis-configuration error that occurred on April 6, 2001. AS 3561 allowed improper route announcements from its downstream customers, which created connectivity disruptions. Surge of announcement messages originating from peer AS 3257 was observed at the RIS rrc03.

Panix Domain Hijack Panix, the oldest commercial ISP in New York state, was hijacked on January 22, 2006. Its services were unreachable from the greater part

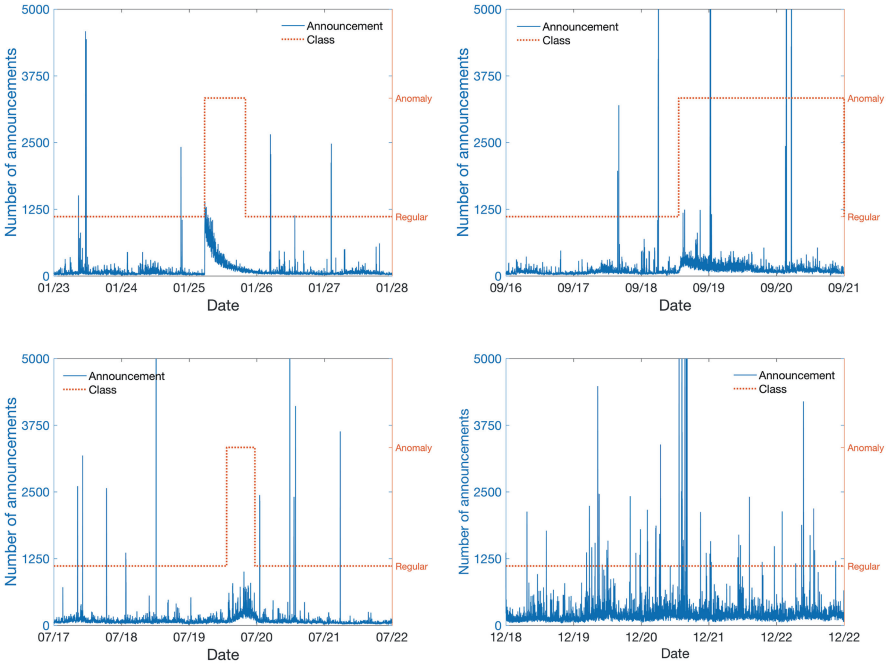

Fig. 1 Number of BGP announcements in Slammer (top left), Nimda (top right), Code Red I (bottom left), and regular RIPE (bottom right) traffic

of the Internet. Con Edison (AS 27506) advertised routes that it did not own at the time. Panix was previously a customer of Con Edison, which was once authorized to offer advertised routes. Even though AS 27506 originated improper routes, major downstream ISPs did not properly configure filters and propagated those routes, leading to excess number of update messages.

AS-path Error The AS-path error occurred on October 7, 2001. It was caused by an abnormal AS-path (AS 3300, AS 64603, AS 2008) that contained private AS 64603 that should not have been included in the path. At the time, AS 3300 and AS 2008 belonged to INFONET Europe and INFONET USA, respectively. The path was distributed to the network via mis-configured routers and caused the leak of the private AS numbers. Shown in Fig. 4, is the increase of incomplete packets around 20:00 GMT, peaking around 21:00 GMT, and slowly decreasing during the following 4 hours.

De-Peering The AS 3356/AS 714 De-Peering event occurred on October 5, 2005. Even though the Level 3 Communications (AS 3356) notified the Cogent Communications (AS 714) 2 months in advance of de-peering, the event created reachability problems for many Internet locations. Mostly affected were single-homed customers of Cogent (approximately 2,300 prefixes) and Level 3 Communications (approximately 5,000 prefixes). De-Peering resulted in partitioning of approximately 4% of prefixes in the global routing table.

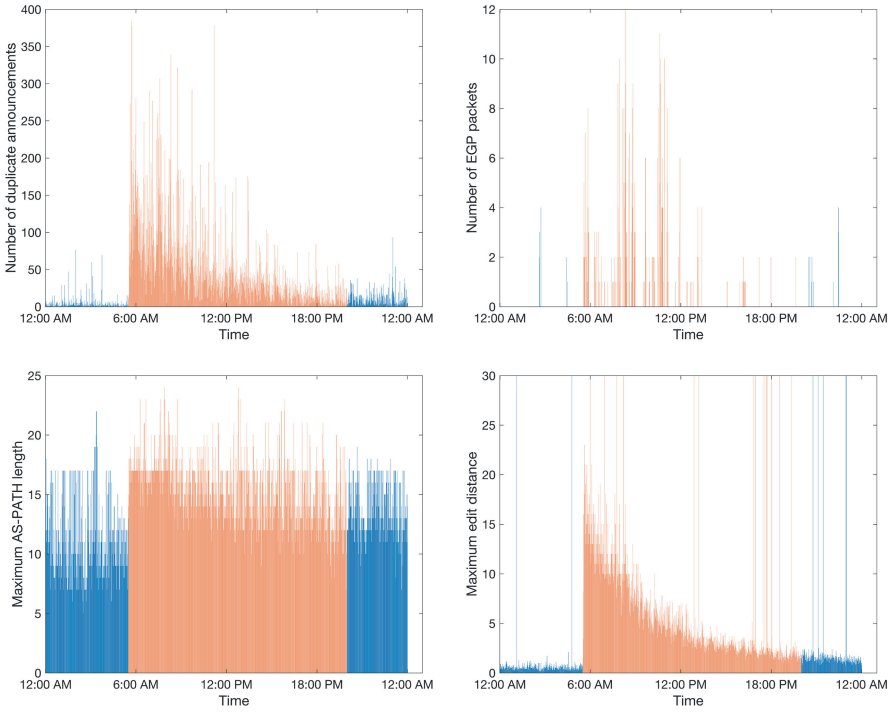

Fig. 2 BGP announcements during the Slammer worm attack: number of duplicate announcements (top left), number of EGP packets (top right), maximum AS-path length (bottom left), and maximum AS-path edit distance (bottom right)

3 Analyzed BGP Datasets

The Internet routing data used in this chapter to detect BGP anomalies are acquired from projects that provide valuable information to networking research: the Route Views project [11] at the University of Oregon, USA and the Routing Information Service (RIS) project initiated in 2001 by the Réseaux IP Européens (RIPE) Network Coordination Centre (NCC) [9]. Both projects collect and store chronological routing data that offer a unique view of the Internet topology. The Route Views and RIPE BGP update messages are publicly available to the research community. The regular BCNET dataset is collected at the BCNET location in Vancouver, British Columbia, Canada [25, 31]. We use BGP update messages that originated from AS 513 (route collector rrc04) member of the CERN Internet Exchange Point (CIXP). Only data collected during the periods of Internet anomalies are considered.

The Route Views project collects BGP routing tables from multiple geographically distributed BGP Cisco routers and Zebra servers every 2 hours. At the time of BGP anomalies considered in this study, two Cisco routers and two Zebra servers were located at the University of Oregon, USA. The remaining five Zebra servers are

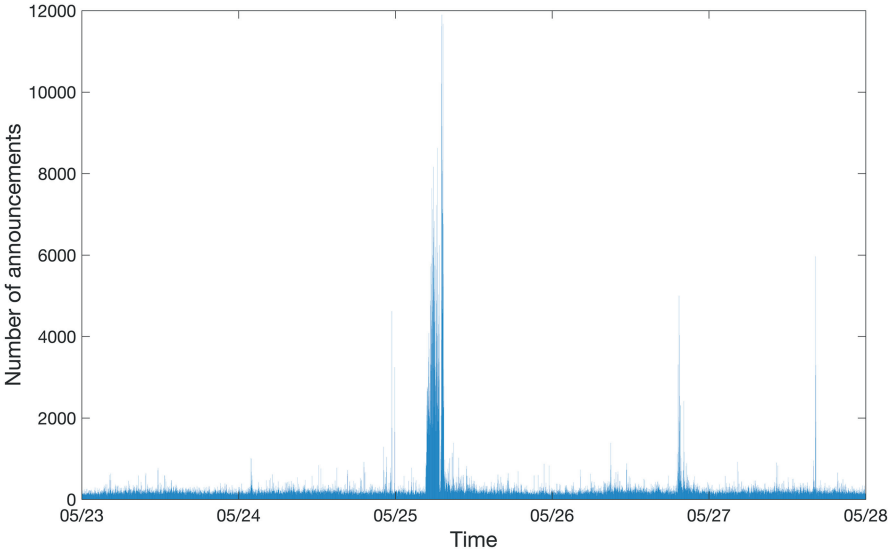

Fig. 3 Surge of announcement messages at the AS 12793 peer during the Moscow Power blackout

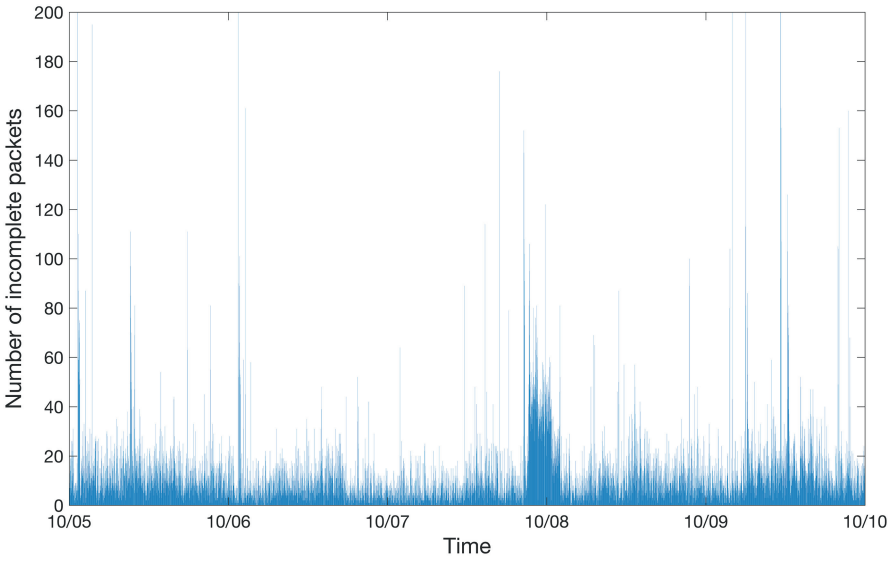

Fig. 4 Surge of incomplete packets at the AS 6762 peer during the AS-path error

located at Equinix-USA, ISC-USA, KIXP-Kenya, LINX-Great Britain, and DIXIE-Japan [11]. Most participating ASes in the Route Views project are located in North America.

The RIPE NCC began collecting and storing Internet routing data in 2001 through the RIS project [9]. The data were exported every 15 min until July 2003. The interval between consecutive exports was later decreased to 5 min. BGP update messages are collected by the RRCs and stored in the multi-threaded routing toolkit (MRT) binary format [7]. The Internet Engineering Task Force (IETF) [4] introduced MRT to export routing protocol messages, state changes, and content of the routing information base (RIB). We transformed BGP update messages from MRT into ASCII format by using libBGPdump library [5] on a Linux platform. LibBGPdump is a C library maintained by the RIPE NCC and it is used to analyze dump files, which are in MRT format.

We use data from the Route Views and RIPE projects data collectors. Only data collected during the periods of Internet anomalies are considered. BGP update messages originated from RIS route collectors: rrc01 (LINX, London), rcc03 (AMS-IX, Amsterdam), rrc04 (CIXP, Geneva), and rrc05 (VIX, Vienna).

3.1 Processing of Collected Data

BGP update messages are collected during the time period when the Internet experienced anomalies. Datasets are concatenated to increase the size of training datasets and thus improve the classification results. Anomaly datasets and their concatenations used for training and testing are shown in Table 4.

We consider a 5-day period for each anomaly: the days of the attack (anomalous data points) and 2 days prior and 2 days after the attack (regular data points). The exception is Nimda dataset where the anomaly lasted longer than 2 days and, hence, we only use 2 days prior to the event as regular data points. Datasets consist of 14,400 ($2\times$ 7,200) data points represented by $14,400 \times 37$ and $14,400 \times 10$ matrices that correspond to 37 and 10 features, respectively. In some cases choosing 15 features was suitable for detecting anomalous events [22] leading to the feature matrices of dimension $7{,}200 \times 15$. In addition to anomalous test datasets, we also use regular datasets collected from RIPE [9] and BCNET [1]. Details of the three anomalies are listed in Table 5.

Table 4 Training and test datasets

Training dataset	Anomalies	Test dataset
1	Slammer and Nimda	Code Red I
2	Slammer and Code Red I	Nimda
3	Nimda and Code Red I	Slammer
4	Slammer	Nimda and Code Red I
5	Nimda	Slammer and Code Red I
6	Code Red I	Slammer and Nimda
7	Slammer, Nimda, and Code Red I	RIPE or BCNET

Table 5 Duration of analyzed BGP events

	Anomaly (min)	Regular (min)
Slammer	869	6,331
Nimda	3,521	3,679
Code Red I	600	6,600

Table 6 List of features extracted from BGP *update* messages

Feature	Name	Category
1	Number of announcements	*volume*
2	Number of withdrawals	*volume*
3	Number of announced NLRI prefixes	*volume*
4	Number of withdrawn NLRI prefixes	*volume*
5	Average *AS-path* length	*AS-path*
6	Maximum *AS-path* length	*AS-path*
7	Average unique *AS-path* length	*AS-path*
8	Number of duplicate announcements	*volume*
9	Number of duplicate withdrawals	*volume*
10	Number of implicit withdrawals	*volume*
11	Average edit distance	*AS-path*
12	Maximum edit distance	*AS-path*
13	Inter-arrival time	*volume*
14–24	Maximum edit distance = n, where $n = (7, \ldots, 17)$	*AS-path*
25–33	Maximum *AS-path* length = n, where $n = (7, \ldots, 15)$	*AS-path*
34	Number of Interior Gateway Protocol (IGP) packets	*volume*
35	Number of Exterior Gateway Protocol (EGP) packets	*volume*
36	Number of incomplete packets	*volume*
37	Packet size (B)	*volume*

4 Extraction of Features from BGP Update Messages

Feature extraction and selection are the first steps in the classification process. We developed a tool (written in C#) [15] to parse the ASCII files and extract statistics of the desired features. The *AS-path* is a BGP update message attribute that enables the protocol to select the best path for routing packets. It indicates a path that a packet may traverse to reach its destination. If a feature is derived from the *AS-path* attribute, it is categorized as an *AS-path* feature. Otherwise, it is categorized as a *volume* feature. There are three types of features: continuous, categorical, and binary. We extracted *AS-path* and *volume* features shown in Table 6 [15].

Definitions of the extracted features are listed in Table 7. BGP update messages are either announcement or withdrawal messages for the NLRI prefixes. The NLRI prefixes that have identical BGP attributes are encapsulated and sent in one BGP packet [37]. Hence, a BGP packet may contain more than one announced or withdrawn NLRI prefix. The average and the maximum number of AS peers are used for calculating *AS-path* lengths. Duplicate announcements are the BGP update packets

Table 7 Definition of *volume* and *AS-path* features extracted from BGP *update* messages

Feature	Name	Definition
1	Number of announcements	Routes available for delivery of data
2	Number of withdrawals	Routes no longer reachable
3/4	Number of announced/withdrawn NLRI prefixes	BGP update messages that have type field set to announcement/withdrawal
5/6/7	Average/maximum/average unique *AS-path* length	Various *AS-path* lengths
8/9	Number of duplicate announcements/withdrawals	Duplicate BGP update messages with type field set to announcement/withdrawal
10	Number of implicit withdrawals	BGP update messages with type field set to announcement and different *AS-path* attribute for already announced NLRI prefixes
11/12	Average/maximum edit distance	Average/maximum of edit distances of messages
34/35/36	Number of IGP, EGP or, incomplete packets	BGP update messages generated by IGP, EGP, or unknown sources

Table 8 Example of BGP features

Time	Definition	BGP update type	NLRI	*AS-path*
t_0	Announcement	Announcement	199.60.12.130	13455 614
t_1	Withdrawal	Withdrawal	199.60.12.130	13455 614
t_2	Duplicate announcement	Announcement	199.60.12.130	13455 614
t_3	Implicit withdrawal	Announcement	199.60.12.130	16180 614
t_4	Duplicate withdrawal	Withdrawal	199.60.12.130	13455 614

that have identical NLRI prefixes and the *AS-path* attributes. Implicit withdrawals are the BGP announcements with different *AS-paths* for already announced NLRI prefixes [48]. The edit distance between two *AS-path* attributes is the minimum number of deletions, insertions, or substitutions that need to be executed to match the two attributes [23]. For example, the edit distance between *AS-path* 513 940 and *AS-path* 513 4567 1318 is two because one insertion and one substitution are sufficient to match the two *AS-paths*. The maximum *AS-path* length and the maximum edit distance are used to count Features 14–33. We also consider Features 34, 35, and 36 based on distinct values of the origin attribute that specifies the origin of a BGP update packet and may assume three values: IGP, EGP, and incomplete. Even though the EGP protocol is the predecessor of BGP, EGP packets still appear in traffic traces containing BGP updates messages. Under a worm attack, BGP traces contain large volume of EGP packets. Furthermore, incomplete update messages imply that the announced NLRI prefixes are generated from unknown sources. They usually originate from BGP redistribution configurations [37]. Examples are shown in Table 8 while various distributions are shown in Figs. 5 and 6.

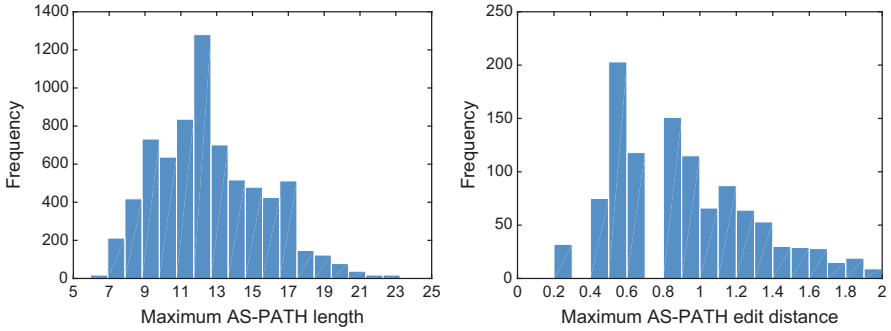

Fig. 5 Distributions of the maximum *AS-path* length (left) and the maximum edit distance (right) collected during the Slammer worm

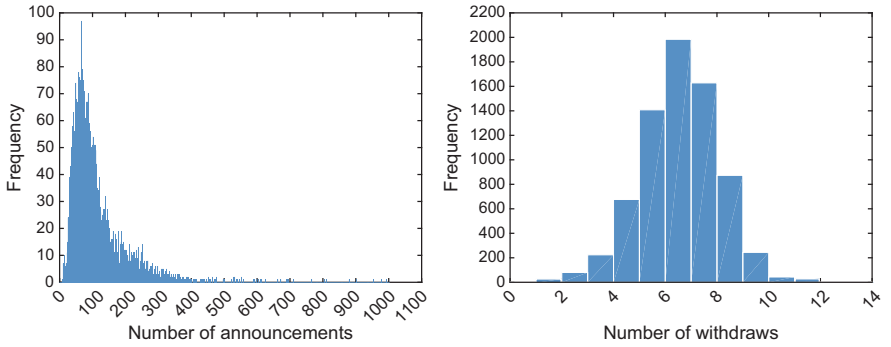

Fig. 6 Distribution of the number of BGP announcements (left) and withdrawals (right) for the Code Red I worm

Performance of the BGP protocol is based on trust among BGP peers because they assume that the interchanged announcements are accurate and reliable. This trust relationship is vulnerable during BGP anomalies. For example, during BGP hijacks, a BGP peer may announce unauthorized prefixes that indicate to other peers that it is the originating peer. These false announcements propagate across the Internet to other BGP peers and, hence, affect the number of BGP announcements (updates and withdrawals) worldwide. This storm of BGP announcements affects the quantity of *volume* features. For example, we have observed that 65% of the influential features are *volume* features. They proved to be more relevant to the anomaly class than the *AS-path* features, which confirms the known consequence of BGP anomalies on the volume of announcements. Hence, using BGP *volume* features is a feasible approach for detecting BGP anomalies and possible worm attacks in communication networks.

The top selected *AS-path* features appear on the boundaries of the distributions. This indicates that during BGP anomalies, the edit distance and *AS-path* length of the BGP announcements tend to have a very high or a very low value and, hence, large variance. This implies that during an anomaly attack, *AS-path* features are the

distribution outliers. For example, approximately 58% of the *AS-path* features are larger than the distribution mean. Large length of the *AS-path* BGP attribute implies that the packet is routed via a longer path to its destination, which causes large routing delays during BGP anomalies. In a similar case, very short lengths of *AS-path* attributes occur during BGP hijacks when the new (false) originator usually gains a preferred or shorter path to the destination [10].

5 Review of Feature Selection Algorithms

Machine learning models classify data points using a feature matrix. The rows correspond to the data points while the columns correspond to the features. Even though machine learning provides general models to classify anomalies, it may easily misclassify test data points due to the redundancy or noise contained in datasets. By providing a sufficient number of relevant features, machine learning models overcome this deficiency and may help design a generalized model to classify data with small error rates [28, 29]. Performance of anomaly classifiers is closely related to feature selection algorithms [20].

Feature selection is used to pre-process data prior to applying machine learning algorithms for classification. Selecting appropriate combination of features is essential for an accurate classification. For example, the scatterings of anomalous and regular classes for Feature 9 (*volume*) vs. Feature 1 (*volume*) and vs. Feature 6 (*AS-path*) are shown in Fig. 7 (left) and (right), respectively. The graphs indicate spatial separation of features. While selecting Feature 9 and Feature 1 may lead to a feasible classification based on visible clusters, using only Feature 9 and Feature 6 would lead to poor classification.

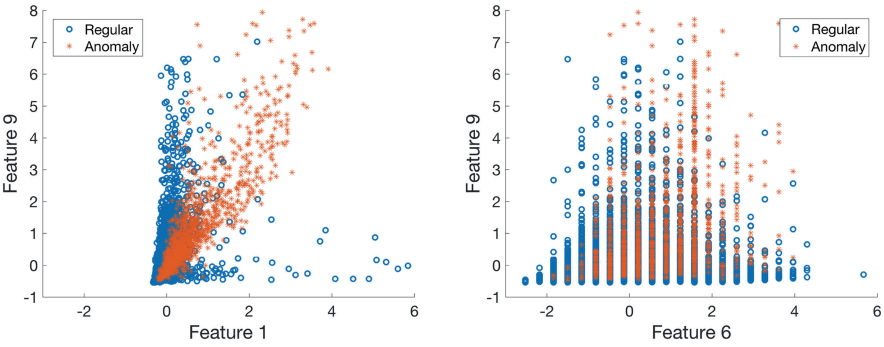

Fig. 7 Scattered graph of Feature 9 vs. Feature 1 (left) and vs. Feature 6 (right) extracted from the BCNET traffic. Feature values are normalized to have zero mean and unit variance. Shown are two traffic classes: regular (open circle) and anomaly (asterisk)

Feature selection follows a feature extraction process and it is used to decrease dimension of the dataset matrix by selecting a subset of original features to create a new matrix according to certain criteria. The number of features is reduced by removing irrelevant, redundant, and noisy features [36]. Feature selection reduces overfitting by minimizing the redundancies in data, improves modeling accuracy, and decreases training time. It also reduces computational complexity and memory usage. Performance of classification algorithms may also be improved by using preselection of features that are most relevant to the classification task. We select the top ten features while dismissing weak and distorted features. We employ Fisher [26, 35, 49], mRMR [40] (including MID, MIQ, and MIBASE), OR (including EOR, WOR, MOR, and CDM) [21], and Decision Tree [42] feature selection algorithms to select relevant features from BGP datasets.

Each sample is a point in n-dimensional space, where kth dimension is a column vector $\mathbf{X}_k$ representing one feature. For example, $\mathbf{X}_1$ is a column vector representing 7,200 announcements in each sampling window of 1 min.

5.1 Fisher Algorithm

The Fisher feature selection algorithm [26, 35, 49] computes the score Φ_k for the kth feature as a ratio of inter-class separation and intra-class variance. Features with higher inter-class separation and lower intra-class variance have higher Fisher scores. If there are N_a^k anomalous samples and N_r^k regular samples of the kth feature, the mean values m_a^k of anomalous samples and m_r^k of regular samples are calculated as:

$$m_a^k = \frac{1}{N_a^k} \sum_{i \in \mathbf{a}_k} x_{ik}$$

$$m_r^k = \frac{1}{N_r^k} \sum_{i \in \mathbf{r}_k} x_{ik}, \tag{1}$$

where $\mathbf{a}_k$ and $\mathbf{r}_k$ are the sets of anomalous and regular samples for the kth feature, respectively. The Fisher score for the kth feature is calculated as:

$$\Phi_k = \frac{|\left(m_a^k\right)^2 - \left(m_r^k\right)^2|}{\frac{1}{N_a^k} \sum_{i \in \mathbf{a}_k} \left(x_{ik} - m_a^k\right)^2 + \frac{1}{N_r^k} \sum_{i \in \mathbf{r}_k} \left(x_{ik} - m_r^k\right)^2}. \tag{2}$$

Table 9 The top ten features selected using the Fisher feature selection algorithm and Dataset 2 (two-way classification)

Feature	9	10	8	3	6	11	1	34	36	2
Fisher score	0.2280	0.1665	0.0794	0.0656	0.0614	0.0610	0.0528	0.0526	0.0499	0.0336

Table 10 The top ten features selected using the Fisher feature selection algorithm and Dataset 7 (four-way classification)

Feature	9	10	6	11	8	36	3	37	1	34
Fisher score	0.1259	0.0502	0.0414	0.0409	0.0281	0.0271	0.0240	0.0239	0.0210	0.0203

Examples of features selected using the Fisher algorithm applied to various training datasets are shown in Tables 9 and 10.

5.2 *Minimum Redundancy Maximum Relevance (mRMR) Algorithms*

The mRMR algorithm [6, 40] relies on information theory for feature selection. It selects a subset of features that contains more information about the target class while having less pairwise mutual information. A subset of features $S = \{\mathbf{X_1}, \ldots, \mathbf{X_k}, \ldots\}$ with $|S|$ elements has the minimum redundancy if it minimizes:

$$\mathscr{W} = \frac{1}{|S|^2} \sum_{\mathbf{X}_k, \mathbf{X}_l \in S} \mathscr{I}(\mathbf{X}_k, \mathbf{X}_l). \tag{3}$$

It has maximum relevance to the classification task if it maximizes:

$$\mathscr{V} = \frac{1}{|S|} \sum_{\mathbf{X}_k \in S} \mathscr{I}(\mathbf{X}_k, \mathbf{C}), \tag{4}$$

where $\mathbf{C}$ is a class vector and $\mathscr{I}$ denotes the mutual information function calculated as:

$$\mathscr{I}(\mathbf{X}_k, \mathbf{X}_l) = \sum_{k,l} p(\mathbf{X}_k, \mathbf{X}_l) log \frac{p(\mathbf{X}_k, \mathbf{X}_l)}{p(\mathbf{X}_k)\, p(\mathbf{X}_l)}. \tag{5}$$

The mRMR algorithm offers three variants for feature selection: Mutual Information Difference (MID), Mutual Information Quotient (MIQ), and Mutual Information Base (MIBASE). MID and MIQ select the best features based on $\max_{S \subset \Omega}[\mathscr{V} - \mathscr{W}]$ and $\max_{S \subset \Omega}[\mathscr{V}/\mathscr{W}]$, respectively, where Ω is the set of all features.

Table 11 The top ten features selected using the mRMR feature selection algorithms and Dataset 2 (two-way classification)

mRMR					
MID		MIQ		MIBASE	
Feature	Score	Feature	Score	Feature	Score
34	0.0554	34	0.0554	34	0.0554
10	0.0117	10	0.7527	1	0.0545
20	0.0047	8	0.6583	8	0.0469
25	0.0014	20	0.5014	10	0.0469
24	0.0012	4	0.4937	3	0.0421
23	0.0008	36	0.4095	9	0.0411
4	0.0007	1	0.3720	36	0.0377
8	0.0007	9	0.3260	4	0.0367
22	0.0006	3	0.2824	6	0.0205
21	0.0005	6	0.2809	11	0.0201

Table 12 The top ten features selected using the mRMR feature selection algorithms and Dataset 7 (four-way classification)

mRMR					
MID		MIQ		MIBASE	
Feature	Score	Feature	Score	Feature	Score
9	0.0407	9	0.0407	9	0.0407
20	0.0030	34	0.4797	1	0.0308
36	0.0024	36	0.3790	34	0.0305
34	0.0024	10	0.3730	36	0.0305
22	0.0017	5	0.3333	3	0.0234
21	0.0017	8	0.3322	8	0.0225
5	0.0003	1	0.3156	10	0.0200
10	0.0003	6	0.2920	6	0.0179
29	0.0002	37	0.2387	11	0.0177
23	0.0000	3	0.2299	37	0.0175

The top ten features selected by mRMR from various datasets are shown in Tables 11 and 12. They are used for two-way and four-way classifications. Selected features shown in Table 11 are generated by using Dataset 2 (Table 4) and are intended for two-way classification. Features shown in Table 12 are generated by using Dataset 7 (Table 4) and are used for four-way classification.

5.3 Odds Ratio Algorithms

The odds ratio (OR) algorithm and its variants perform well when selecting features to be used in binary classification using naive Bayes models. In case of a binary classification with two target classes c and $\bar{c}$, the odds ratio for a feature $\mathbf{X}_k$ is calculated as:

$$OR(\mathbf{X}_k) = \log \frac{\Pr(\mathbf{X}_k|c)\big(1 - \Pr(\mathbf{X}_k|\bar{c})\big)}{\Pr(\mathbf{X}_k|\bar{c})\big(1 - \Pr(\mathbf{X}_k|c)\big)}, \tag{6}$$

where $\Pr(\mathbf{X}_k|c)$ and $\Pr(\mathbf{X}_k|\bar{c})$ are the probabilities of feature $\mathbf{X}_k$ being in classes c and $\bar{c}$, respectively.

The extended odds ratio (EOR), weighted odds ratio (WOR), multi-class odds ratio (MOR), and class discriminating measure (CDM) are variants that enable multi-class feature selections in case of $\gamma = \{c_1, c_2, \ldots, c_J\}$ classes:

$$
\begin{aligned}
EOR(\mathbf{X}_k) &= \sum_{j=1}^{J} \log \frac{\Pr(\mathbf{X}_k|c_j)\big(1 - \Pr(\mathbf{X}_k|\bar{c_j})\big)}{\Pr(\mathbf{X}_k|\bar{c_j})\big(1 - \Pr(\mathbf{X}_k|c_j)\big)} \\
WOR(\mathbf{X}_k) &= \sum_{j=1}^{J} \Pr(c_j) \times \log \frac{\Pr(\mathbf{X}_k|c_j)\big(1 - \Pr(\mathbf{X}_k|\bar{c_j})\big)}{\Pr(\mathbf{X}_k|\bar{c_j})\big(1 - \Pr(\mathbf{X}_k|c_j)\big)} \\
MOR(\mathbf{X}_k) &= \sum_{j=1}^{J} \left| \log \frac{\Pr(\mathbf{X}_k|c_j)\big(1 - \Pr(\mathbf{X}_k|\bar{c_j})\big)}{\Pr(\mathbf{X}_k|\bar{c_j})\big(1 - \Pr(\mathbf{X}_k|c_j)\big)} \right| \\
CDM(\mathbf{X}_k) &= \sum_{j=1}^{J} \left| \log \frac{\Pr(\mathbf{X}_k|c_j)}{\Pr(\mathbf{X}_k|\bar{c_j})} \right|,
\end{aligned} \tag{7}
$$

where $\Pr(\mathbf{X}_k|c_j)$ is the conditional probability of $\mathbf{X}_k$ given the class c_j and $\Pr(c_j)$ is the probability of occurrence of the class j. The OR algorithm is extended by calculating $\Pr(\mathbf{X}_k|c_j)$ for continuous features. If the sample points are independent and identically distributed, (6) is written as:

$$
OR(\mathbf{X}_k) = \sum_{i=1}^{|\mathbf{X}_k|} \log \frac{\Pr(X_{ik} = x_{ik}|c)\big(1 - \Pr(X_{ik} = x_{ik}|\bar{c})\big)}{\Pr(X_{ik} = x_{ik}|\bar{c})\big(1 - \Pr(X_{ik} = x_{ik}|c)\big)}, \tag{8}
$$

where $|\mathbf{X}_k|$ denote the size of the kth feature vector, X_{ik} is the ith element of the kth feature vector, and x_{ik} is realization of the random variable X_{ik}. Other variants of the OR feature selection algorithm are extended to continuous cases in a similar manner. The top ten selected features used for two-way and four-way classifications are shown in Tables 13 and 14, respectively.

5.4 Decision Tree Algorithm

The decision tree approach is commonly used in data mining to predict the class label based on several input variables. A classification tree is a directed tree where the root is the source sample set and each internal (non-leaf) node is labeled with an input feature. The tree branches are prediction outcomes that are labeled with possible feature values while each leaf node is labeled with a class or a class probability distribution [50]. A top-down approach is commonly used for

Table 13 The top ten features selected using the OR feature selection algorithms and Dataset 2 (two-way classification)

Odds Ratio variants							
OR		WOR		MOR		CDM	
Feature	Score $\times 10^4$	Feature	Score $\times 10^4$	Feature	Score $\times 10^5$	Feature	Score $\times 10^5$
13	−2.7046	12	3.9676	12	1.0789	12	1.0713
7	−2.8051	1	3.4121	34	0.9214	34	0.9199
5	−2.8064	34	3.4095	1	0.9213	1	0.9198
29	−2.8774	3	3.3482	3	0.8908	3	0.8885
15	−2.8777	4	3.3468	4	0.8775	4	0.8702
28	−2.9136	23	2.9348	9	0.7406	9	0.7224
14	−2.9137	24	2.7628	36	0.7264	36	0.7201
6	−2.9190	22	2.7051	37	0.7229	37	0.7192
11	−2.9248	21	2.6662	23	0.7208	2	0.7145
30	−2.9288	20	2.5821	2	0.6782	8	0.6624

Table 14 The top ten features selected using the OR feature selection algorithms and Dataset 7 (four-way classification)

Odds Ratio variants							
EOR		WOR		MOR		CDM	
Feature	Score $\times 10^5$	Feature	Score $\times 10^4$	Feature	Score $\times 10^5$	Feature	Score $\times 10^5$
3	−1.5496	12	1.7791	12	3.0894	12	3.0700
13	−1.5681	1	1.3293	34	2.5964	34	2.5924
9	−1.6063	34	1.3273	1	2.5723	1	2.5688
11	−1.6184	4	1.1140	4	2.5252	4	2.5190
6	−1.6184	36	1.0763	36	2.4617	36	2.4422
37	−1.6191	2	1.0140	2	2.3024	2	2.2300
5	−1.6499	23	0.8669	23	2.2832	8	2.2068
7	−1.6522	24	0.8529	24	2.2733	9	2.1357
29	−1.6783	21	0.8508	21	2.2696	10	2.1168
15	−1.6784	20	0.8504	22	2.2696	3	2.0848

constructing decision trees. At each step, an appropriate variable is chosen to best split the set of items. A quality measure is the homogeneity of the target variable within subsets and it is applied to each candidate subset. The combined results measure the split quality [19, 46].

The C5 [2] software tool is used to generate decision tree for both feature selections and anomaly classifications. The C5 decision tree algorithm relies on the information gain measure. The continuous attribute values are discretized and the most important features are iteratively used to split the sample space until a certain portion of samples associated with the leaf node has the same value as the target attribute. For each training dataset, a set of rules used for classification is extracted from the constructed decision tree.

Table 15 Selected features using the decision tree algorithm

Training dataset	Selected features
Dataset 1	1–21, 23–29, 34–37
Dataset 2	1–22, 24–29, 34–37
Dataset 3	1–29, 34–37

We apply the decision tree algorithm for feature selection to form the training datasets shown in Table 4. These datasets are also used in the classification stage. The selected features are shown in Table 15. Based on the outcome of the decision tree algorithm, some features are removed in the constructed trees. Fewer features are selected either based on the number of leaf nodes with the largest correct classified samples or based on the number of rules with maximum sample coverage. The features that appear in the selected rules are considered to be important and, therefore, are preserved.

6 Conclusion

Detecting network anomalies and intrusions are crucial in fighting cyber attacks and insuring cyber security to service providers and network customers. Machine learning techniques are one of the most promising approaches for detecting network anomalies and have been employed in analyzing BGP behavior. In this chapter, we introduce BGP datasets, investigate BGP anomalies, and describe various feature selection techniques. Datasets used in these experiments are examples of known anomalies that proved useful for developing anomaly detection models. We have processed and extracted features from known BGP anomalies such as Slammer, Nimda, and Code Red I worms as well as the Moscow power blackout, AS 9121 routing table leak, Panix hijack, and AS-path error datasets. Various feature selection and attribute reduction algorithms are used to select a subset of features important for classification. After the feature selection process, extracted features are used as input to machine learning classification algorithms described in the follow-up Chapter.

Acknowledgements We thank Yan Li, Hong-Jie Xing, Qiang Hua, and Xi-Zhao Wang from Hebei University, Marijana Ćosović from University of East Sarajevo, and Prerna Batta from Simon Fraser University for their helpful contributions in earlier publications related to this project.

References

1. (Mar. 2018) BCNET. [Online]. Available: http://www.bc.net.
2. (Mar. 2018) Data Mining Tools See5 and C5.0. [Online]. Available: http://www.rulequest.com/see5-info.html.

3. (Mar. 2018) Sans Institute. The mechanisms and effects of the Code Red worm. [Online]. Available: https://www.sans.org/reading-room/whitepapers/dlp/mechanisms-effects-code-red-worm-87.
4. (Mar. 2018) The Internet Engineering Task Force (IETF) [Online]. Available: https://www.ietf.org/.
5. (Mar. 2018) bgpdump [Online]. Available: https://bitbucket.org/ripencc/bgpdump/wiki/Home.
6. (Mar. 2018) mRMR feature selection (using mutual information computation). [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/14608-mrmr-feature-selection--using-mutual-information-computation-.
7. (Mar. 2018) MRT rooting information export format. [Online]. Available: http://tools.ietf.org/html/draft-ietf-grow-mrt-13.
8. (Mar. 2018) Sans Institute. Nimda worm—why is it different? [Online]. Available: http://www.sans.org/reading-room/whitepapers/malicious/nimda-worm-different-98.
9. (Mar. 2018) RIPE NCC: RIPE Network Coordination Center. [Online]. Available: http://www.ripe.net/data-tools/stats/ris/ris-raw-data.
10. (Mar. 2018) YouTube Hijacking: A RIPE NCC RIS case study [Online]. Available: http://www.ripe.net/internet-coordination/news/industry-developments/youtube-hijacking-a-ripe-ncc-ris-case-study.
11. (Mar. 2018) University of Oregon Route Views project [Online]. Available: http://www.routeviews.org/.
12. (Mar. 2018) Center for Applied Internet Data Analysis. The Spread of the Sapphire/Slammer Worm [Online]. Available: http://www.caida.org/publications/papers/2003/sapphire/.
13. (Mar. 2018) Sans Institute. Malware FAQ: MS-SQL Slammer. [Online]. Available: https://www.sans.org/security-resources/malwarefaq/ms-sql-exploit.
14. T. Ahmed, B. Oreshkin, and M. Coates, "Machine learning approaches to network anomaly detection," in *Proc. USENIX Workshop on Tackling Computer Systems Problems with Machine Learning Techniques*, Cambridge, MA, Apr. 2007, pp. 1–6.
15. N. Al-Rousan and Lj. Trajković, "Machine learning models for classification of BGP anomalies," in *Proc. IEEE Conf. on High Performance Switching and Routing (HPSR)*, Belgrade, Serbia, June 2012, pp. 103–108.
16. N. Al-Rousan, S. Haeri, and Lj. Trajković, "Feature selection for classification of BGP anomalies using Bayesian models," in *Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC)*, Xi'an, China, July 2012, pp. 140–147.
17. K. El-Arini and K. Killourhy, "Bayesian detection of router configuration anomalies," in *Proc. Workshop Mining Network Data*, Philadelphia, PA, USA, Aug. 2005, pp. 221–222.
18. M. Bhuyan, D. Bhattacharyya, and J. Kalita, "Network anomaly detection: methods, systems and tools," *IEEE Commun. Surveys Tut.*, vol. 16, no. 1, pp. 303–336, Mar. 2014.
19. L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, Aug. 1996.
20. Y.-W. Chen and C.-J. Lin, "Combining SVMs with various feature selection strategies," *Strategies*, vol. 324, no. 1, pp. 1–10, Nov. 2006.
21. J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with naive Bayes," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5432–5435, Apr. 2009.
22. M. Ćosović, S. Obradović, and Lj. Trajković, "Classifying anomalous events in BGP datasets," in *Proc. The 29th Annu. IEEE Can. Conf. on Elect. and Comput. Eng. (CCECE)*, Vancouver, Canada, May 2016, pp. 697–700.
23. S. Deshpande, M. Thottan, T. K. Ho, and B. Sikdar, "An online mechanism for BGP instability detection and analysis," *IEEE Trans. Comput.*, vol. 58, no. 11, pp. 1470–1484, Nov. 2009.
24. Q. Ding, Z. Li, P. Batta, and Lj. Trajković, "Detecting BGP anomalies using machine learning techniques," in *Proc. IEEE Int. Conf. Syst., Man, and Cybern.*, Budapest, Hungary, Oct. 2016, pp. 3352–3355.
25. T. Farah, S. Lally, R. Gill, N. Al-Rousan, R. Paul, D. Xu, and Lj. Trajković, "Collection of BCNET BGP traffic," in *Proc. 23rd ITC*, San Francisco, CA, USA, Sept. 2011, pp. 322–323.

26. Q. Gu, Z. Li, and J. Han, "Generalized Fisher score for feature selection," in *Proc. Conf. Uncertainty in Artificial Intelligence*, Barcelona, Spain, July 2011, pp. 266–273.
27. H. Hajji, "Statistical analysis of network traffic for adaptive faults detection," *IEEE Trans. Neural Netw.*, vol. 16, no. 5, pp. 1053–1063, Sept. 2005.
28. G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Proc. Int. Conf. Machine Learning*, New Brunswick, NJ, USA, July 1994, pp. 121–129.
29. M. N. A. Kumar and H. S. Sheshadri, "On the classification of imbalanced datasets," *Int. J. Comput. Appl.*, vol. 44, no. 8, pp. 1–7, Apr. 2012.
30. J. Kurose and K. W. Ross, *"Computer Networking: A Top-Down Approach (6th edition)."* Addison-Wesley, 2012, pp. 305–431.
31. S. Lally, T. Farah, R. Gill, R. Paul, N. Al-Rousan, and Lj. Trajković, "Collection and characterization of BCNET BGP traffic," in *Proc. 2011 IEEE Pacific Rim Conf. Commun., Comput. and Signal Process.*, Victoria, BC, Canada, Aug. 2011, pp. 830–835.
32. F. Lau, S. H. Rubin, M. H. Smith, and Lj. Trajković, "Distributed denial of service attacks," in *Proc. IEEE Int. Conf. Syst., Man, and Cybern., SMC 2000*, Nashville, TN, USA, Oct. 2000, pp. 2275–2280.
33. J. Li, D. Dou, Z. Wu, S. Kim, and V. Agarwal, "An Internet routing forensics framework for discovering rules of abnormal BGP events," *SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 5, pp. 55–66, Oct. 2005.
34. Y. Li, H. J. Xing, Q. Hua, X.-Z. Wang, P. Batta, S. Haeri, and Lj. Trajković, "Classification of BGP anomalies using decision trees and fuzzy rough sets," in *Proc. IEEE Trans. Syst., Man, Cybern.*, San Diego, CA, USA, Oct. 2014, pp. 1331–1336.
35. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification.* Hoboken, NJ, USA: Wiley-Interscience Publication, 2001.
36. H. Liu, H. Motoda, Eds., *Computational Methods of Feature Selection*. Boca Raton, FL, USA: Chapman and Hall/CRC Press, 2007.
37. (Mar. 2018) D. Meyer, "BGP communities for data collection," RFC 4384, *IETF*, Feb. 2006. [Online]. Available: http://www.ietf.org/rfc/rfc4384.txt.
38. Z. Pawlak, "Rough sets," *Int. J. Inform. and Comput. Sci.*, vol. 11, no. 5, pp. 341–356, Oct. 1982.
39. C. Patrikakis, M. Masikos, and O. Zouraraki, "Distributed denial of service attacks," *The Internet Protocol*, vol. 7, no. 4, pp. 13–31, Dec. 2004.
40. H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
41. (Mar. 2018) A. C. Popescu, B. J. Premore, and T. Underwood, The anatomy of a leak: AS9121. Renesys Corporation, Manchester, NH, USA. May 2005. [Online]. Available: http://50.31.151.73/meetings/nanog34/presentations/underwood.pdf.
42. J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986.
43. A. M. Radzikowska and E. E. Kerre, "A comparative study of fuzzy rough sets," *Fuzzy Sets and Syst.*, vol. 126, no. 2, pp. 137–155, Mar. 2002.
44. (Mar. 2018) Y. Rekhter and T. Li, "A Border Gateway Protocol 4 (BGP-4)," RFC 1771, *IETF*, Mar. 1995. [Online]. Available: http://tools.ietf.org/rfc/rfc1771.txt.
45. (Mar. 2018) Y. Rekhter, T. Li, and S. Hares, "A Border Gateway Protocol 4 (BGP-4)," RFC 4271, *IETF*, Jan. 2016. [Online]. Available: http://tools.ietf.org/rfc/rfc4271.txt.
46. L. Rokach and O. Maimon, "Top-down induction of decision trees classifiers—a survey," *IEEE Trans. Syst., Man, Cybern., Appl. and Rev.*, vol. 35, no. 4, pp. 476–487, Nov. 2005.
47. M. Thottan and C. Ji, "Anomaly detection in IP networks," *IEEE Trans. Signal Process.*, vol. 51, no. 8, pp. 2191–2204, Aug. 2003.
48. L. Wang, X. Zhao, D. Pei, R. Bush, D. Massey, A. Mankin, S. F. Wu, and L. Zhang, "Observation and analysis of BGP behavior under stress," in *Proc. 2nd ACM SIGCOMM Workshop on Internet Meas.*, New York, NY, USA, 2002, pp. 183–195.

49. J. Wang, X. Chen, and W. Gao, "Online selecting discriminative tracking features using particle filter," in *Proc. Comput. Vision and Pattern Recognition*, San Diego, CA, USA, June 2005, vol. 2, pp. 1037–1042.
50. X.-Z. Wang, L. C. Dong, and J. H. Yan, "Maximum ambiguity based sample selection in fuzzy decision tree induction," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 8, pp. 1491–1505, Aug. 2012.
51. D. P. Watson and D. H. Scheidt, "Autonomous systems," *Johns Hopkins APL Technical Digest*, vol. 26, no. 4, pp. 368–376, Oct.–Dec. 2005.
52. D. S. Yeung, D. G. Chen, E. C. C. Tsang, J. W. T. Lee, and X.-Z. Wang, "On the generalization of fuzzy rough sets," *IEEE Trans. Fuzz. Syst.*, vol. 13, no. 3, pp. 343–361, June 2005.
53. J. Zhang, J. Rexford, and J. Feigenbaum, "Learning-based anomaly detection in BGP updates," in *Proc. Workshop Mining Netw. Data*, Philadelphia, PA, USA, Aug. 2005, pp. 219–220.
54. Y. Zhang, Z. M. Mao, and J. Wang, "A firewall for routers: protecting against routing misbehavior," in *Proc. 37th Annu. IEEE/IFIP Int. Conf. on Dependable Syst. and Netw.*, Edinburgh, UK, June 2007, pp. 20–29.