Machine Learning Models for Classification of BGP Anomalies

Nabil M. Al-Rousan and Ljiljana Trajković
Simon Fraser University
Vancouver, British Columbia
Email: {nalrousa, ljilja}@sfu.ca

***Abstract*—Worms such as Slammer, Nimda, and Code Red I are anomalies that affect performance of the global Internet Border Gateway Protocol (BGP). BGP anomalies also include Internet Protocol (IP) prefix hijacks, miss-configurations, and electrical failures. Statistical and machine learning techniques have been recently deployed to classify and detect BGP anomalies. In this paper, we introduce new classification features and apply Support Vector Machine (SVM) models and Hidden Markov Models (HMMs) to design anomaly detection mechanisms. We apply these multi classification models to correctly classify test datasets and identify the correct anomaly types. The proposed models are tested with collected BGP traffic traces and are employed to successfully classify and detect various BGP anomalies.**

I. Introduction

Border Gateway Protocol (BGP) anomalies often occur and techniques for their detection have recently gained visible attention and importance. Recent research reports describe a number of applicable detection techniques. One of the most common approaches is based on a statistical pattern recognition model implemented as an anomaly classifier [1]. Its main disadvantage is the difficulty in estimating distributions of high dimensions. Other proposed techniques are rule-based and require a priori knowledge of network conditions. An example is the Internet Routing Forensics (IRF) that was applied to classify anomaly events [2]. However, rule-based techniques are not adaptable learning mechanisms, are slow, and have high degree of computational complexity.

In this paper, we employ machine learning techniques to develop models for detecting BGP anomalies. We extract numerous BGP features in order to achieve reliable classification results. We use Support Vector Machine (SVM) models to train and test various datasets. Hidden Markov Models (HMMs) were also employed to evaluate the effectiveness of the extracted traffic features.

This paper is organized as follows. In Section II, we provide a description of the BGP data processing that consists of features extraction and selection. The design of proposed classification models and their evaluation are described in Section III and Section IV. Performance of the proposed models is discussed in Section V. We conclude with Section VI.

II. Data Processing

A. Extraction of Features

In 2001, Réseaux IP Européens (RIPE) [3] initiated the Routing Information Service (RIS) project to collect BGP update messages. Real-time BGP data were also collected by the Route Views [4] project at the University of Oregon, USA. The RIPE and Route Views BGP update messages are available to the research community in the multi-threaded routing toolkit (MRT) binary format [5], which was introduced by the Internet Engineering Task Force (IETF) to export routing protocol messages, state changes, and contents of the routing information base (RIB). We used the Zebra tool [6] to convert MRT to ASCII format and then extract traffic features. Traffic traces of three BGP anomalies along with regular RIPE traffic are shown in Fig. 1. A sample of the BGP update message format is shown in Table I. 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 Autonomous System (AS) peers. The AS-PATH attribute enables BGP to route packets via the best path.

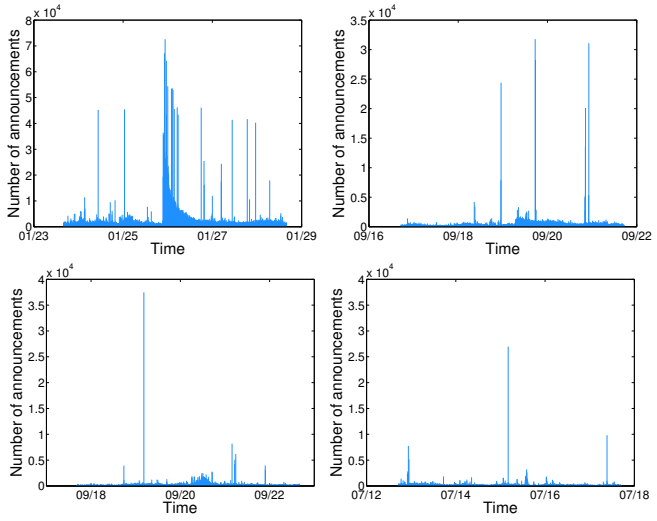

Fig. 1: Number of BGP announcements in Slammer (top left), Nimda (top right), Code Red I (bottom left), and regular RIPE (bottom right) traffic.

We collected the BGP update messages that originated from AS 513 (RIPE RIS, rcc04, CIXP, Geneva) and included a sample of the BGP traffic during time periods when the Internet experienced BGP anomalies. Details of the three anomalies and two regular traffic events considered in this paper are listed in Table II.

We developed a tool (written in C#) to parse the ASCII files

TABLE I: Sample of a BGP update packet.

Field	Value
TIME	2003 1 24 00:39:53
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 7246
NEXT-HOP	192.65.184.3
ANNOUNCED NLRI PREFIX	198.155.189.0/24
ANNOUNCED NLRI PREFIX	198.155.241.0/24

TABLE II: Details of BGP datasets.

	Class	Date	Duration (h)
Slammer	Anomaly	January 25, 2003	16
Nimda	Anomaly	September 18, 2001	59
Code Red I	Anomaly	July 19, 2001	10
RIPE	Regular	July 14, 2001	24
BCNET	Regular	December 20, 2011	24

and extract statistics of the desired features. These features are sampled every minute during a five-day period, producing 7,200 samples for each anomaly event. They are used as inputs for classification models. Samples from two days before and after each event are considered to be regular test datasets. The third day was the peak of each anomaly. The features are normalized to have zero mean and unit variance. This normalization reduces the effect of the Internet growth between 2003 and 2011. Extracted features, shown in Table III, are categorized as *volume* (number of BGP announcements) and *AS-path* (maximum edit distance) features. The effect of Slammer worm on *volume* and *AS-path* features is illustrated in Fig. 2.

TABLE III: Extracted features.

Feature	Definition	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 $n = (7, ..., 17)$	*AS-path*
25-33	Maximum AS-path length = n, where $n = (7, ..., 16)$	*AS-path*
34	Number of IGP packets	*volume*
35	Number of EGP packets	*volume*
36	Number of incomplete packets	*volume*
37	Packet size (B)	*volume*

BGP protocol generates four types of messages: open, update, keepalive, and notification. We only consider BGP update messages because they contain all features that we wish to extract. BGP update messages are either announcement or

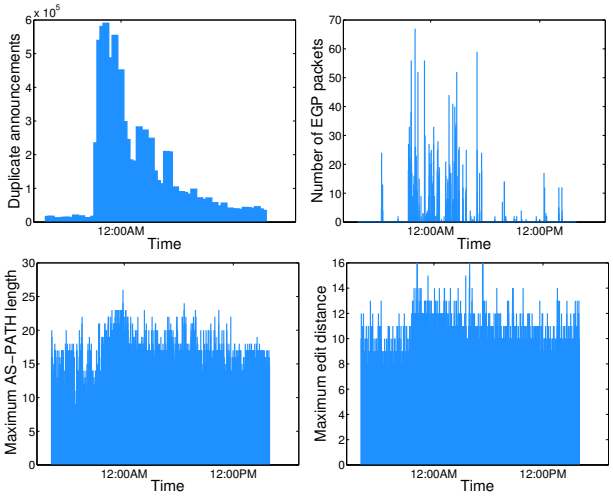

Fig. 2: Samples of extracted BGP features during the Slammer worm attack. Shown are *volume* features 8 (top left) and 35 (top right) and *AS-path* features 6 (bottom left) and 12 (bottom right).

withdrawal messages for the NLRI prefixes. Feature statistics are computed for one-minute time intervals. The NLRI prefixes that have identical BGP attributes are encapsulated and sent in one BGP packet [7]. Hence, a BGP packet may contain more than one announced or withdrawal NLRI prefix. While features 5 and 6 are the average and the maximum number of AS peers for AS-PATH attribute, respectively, feature 7 only considers the unique AS-PATH attributes. Duplicate announcements are the BGP update packets that have identical NLRI prefixes and AS-PATH attributes. Implicit withdrawals are the BGP announcements with different AS-PATHs for already announced NLRI prefixes [8]. An example is shown in Table IV. The edit distance between two AS-PATH attributes is the minimum number of insertions, deletions, and substitutions that need to be executed to match the two attributes. The value of the edit distance feature is extracted by computing the edit distance between the AS-PATH attributes in each one-minute time interval [1]. 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 most frequent values of the maximum AS-PATH length and the maximum edit distance are used to calculate features 14 to 33. Their distributions for the Slammer worm are shown in Fig. 3.

TABLE IV: Definitions 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

We also introduce three new 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

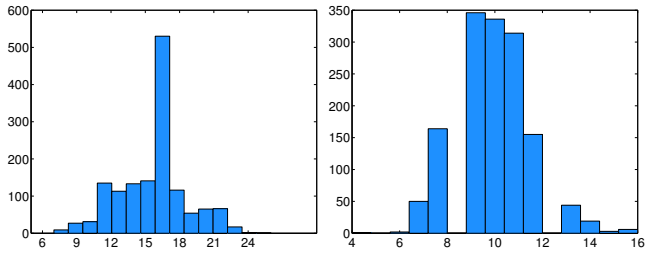

Fig. 3: Distributions of (left) the maximum AS-PATH length and (right) the maximum edit distance.

values: IGP (generated by an Interior Gateway Protocol), EGP (generated by the Exterior Gateway Protocol), and Incomplete. The EGP protocol is the BGP predecessor, which is not currently used by Internet Service Providers (ISPs). However, EGP packets still appear in traffic traces containing BGP updates messages. Under a worm attack, BGP traces contain large number of EGP packets. Incomplete update messages imply that the announced NLRI prefixes are generated from unknown sources. They usually originate from BGP redistribution configurations [7].

B. Selection of Features

We use the Fisher [9] and minimum Redundancy Maximum Relevance (mRMR) [10] feature scoring algorithms to select the most relevant features. These algorithms measure the correlation and relevancy among features and, hence, help improve the classification accuracy. We select the top ten features for the Fisher feature selection and thus neglect the weak and distorted features in the classification models.

Each training datasets is represented as a real matrix $\mathbf{X}_{7200\times 37}$. Each column vector $\mathbf{X}_k, k = 1, ..., 37$ corresponds to one feature. The Fisher score for $\mathbf{X}_k$ is computed as:

$$\text{F-Score} = \frac{m_a^2 - m_r^2}{s_a^2 + s_r^2} = \frac{\frac{1}{N_a}\sum_{i\in anomlay} x_{i,k}^{\ 2} - \frac{1}{N_r}\sum_{i\in regular} x_{i,k}^{\ 2}}{\frac{1}{N_a}\sum_{i\in anomlay}\left(x_{i,k} - m_a\right)^2 + \frac{1}{N_r}\sum_{i\in regular}\left(x_{i,k} - m_r\right)^2}, \quad (1)$$

where N_a and N_r are the numbers of anomaly and regular data points, respectively and m_a and s_a^2 (m_r and s_r^2) are the mean and the variance for anomaly (regular) class, respectively. The Fisher algorithm maximizes the inter-class separation $m_a^2 - m_r^2$ and minimizes the intra-class variances s_a^2 and s_r^2.

The mRMR algorithm maximizes the relevance of features with respect to the target class while minimizing the redundancy among features. We use three variants of the mRMR algorithm: Mutual Information Difference (MID), Mutual Information Quotient (MIQ), and Mutual Information Base (MIBASE). The mRMR relevance of a feature set $S = \{\mathbf{X}_1, ..., \mathbf{X}_k, \mathbf{X}_l, ..., \mathbf{X}_{37}\}$ for a class vector $\mathbf{Y}$ is based on the mutual information function $\mathcal{I}$:

$$\mathcal{I}(\mathbf{X}_k, \mathbf{X}_l) = \sum_{k,l} p(\mathbf{X}_k, \mathbf{X}_l) log \frac{p(\mathbf{X}_k, \mathbf{X}_l)}{p(\mathbf{X}_k)p(\mathbf{X}_l)}. \quad (2)$$

The mRMR variants are defined by the criteria:

$$\begin{aligned} \text{MID: } &\max\left[V(\mathcal{I}) - W(\mathcal{I})\right] \\ \text{MIQ: } &\max\left[V(\mathcal{I})/W(\mathcal{I})\right], \end{aligned} \quad (3)$$

where:

$$V(I) = \frac{1}{|S|}\sum_{\mathbf{X}_k\in S} \mathcal{I}(\mathbf{X}_k, \mathbf{Y})$$

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

Constant $|S|$ is the length of the set S. The MIBASE feature scores are ordered based on their value (2). The Fisher and mRMR scores are obtained for the set of features captured on January 25, 2003. The test set contains 1,440 samples where 869 samples are labeled as anomaly. The top ten features using the Fisher and mRMR algorithms are listed in Table V. They are evaluated later by using the SVM classification.

TABLE V: Top ten features used for selection algorithms.

Fisher		mRMR					
		MID		MIQ		MIBASE	
Feature	Score	Feature	Score	Feature	Score	Feature	Score
11	0.39	34	0.94	34	0.94	34	0.94
6	0.35	32	0.02	2	0.33	36	0.63
25	0.29	33	0.02	8	0.34	2	0.47
9	0.27	2	0.01	24	0.31	8	0.34
2	0.18	31	0.02	9	0.33	9	0.27
36	0.12	24	0.01	14	0.30	3	0.13
37	0.12	8	0.01	1	0.35	1	0.13
24	0.12	14	0.02	36	0.36	6	0.10
8	0.11	30	0.02	3	0.30	12	0.08
14	0.08	22	0.02	25	0.27	11	0.06

III. Classification using Support Vector Machines

We use the SVM classification as supervised deterministic model to classify BGP anomalies. MATLAB libsvm-3.1 toolbox [11] is used to train and test the SVM classifiers. The dimension of feature matrix is $7,200 \times 10$ and corresponds to a five-day period. Each matrix row corresponds to the top ten selected features within the one-minute interval. For each training dataset $\mathbf{X}_{7200\times 37}$, we target two classes: anomaly (true) and regular (false). The SVM solves a loss function as an optimization problem [12] with the constraints:

$$\begin{aligned} &\min C\sum_{m=1}^{M}\xi_m + \frac{1}{2}\|w\|^2 \\ &t_m y(\mathbf{X}_m) \geq 1 - \xi_m. \end{aligned} \quad (4)$$

Constant $C > 0$ controls the importance of the margin while slack variable ξ_m solves the non-separable data points classification problem. A regularization parameter $\frac{1}{2}\|w\|^2$ is used to avoid over-fitting problem. SVM classifies each data point

$\mathbf{X}_m$ with a training target class t_m either as anomaly $y = 1$ or regular $y = -1$. $\mathbf{X}_m$ corresponds to a row vector where $m = 1, ..., 7200$. The SVM solution maximizes the margin between the data points and the decision boundary. Data points that have the minimum distance to the decision boundary are called support vectors. The Radial Basis Function (RBF) kernel is used to avoid the high dimension of the feature matrix:

$$\mathcal{K}(\mathbf{X}_k, \mathbf{X}_l) = exp(-\gamma * \|\mathbf{X}_k - \mathbf{X}_l\|^2). \quad (5)$$

The RBF kernel $\mathcal{K}$ depends on the Euclidean distance between $\mathbf{X}_k$ and $\mathbf{X}_l$ features. Constant γ influences the number of support vectors. The datasets are trained using 10-fold cross validation to select parameters (C, γ) that give the best accuracy. We apply SVM on sets listed in Table VI to classify BGP anomalies.

TABLE VI: The SVM dataset.

SVM	Training dataset	Test dataset		
		Code Red I	Nimda	Slammer
SVM_1	Slammer and Nimda	√	x	x
SVM_2	Slammer and Code Red I	x	√	x
SVM_3	Code Red I and Nimda	x	x	√

Three measures are used for performance indices: *sensitivity*, *specificity*, and *precision*. *Sensitivity* indicates the ability of the model to identify anomalies (true positive) among all labeled anomalies (true). *Specificity* reflects the ability of the model to identify the regular traffic (true negative) among all regular traffic (false). *Precision* is the ability of the model to identify anomalies (true positive) among all data points that are identified as anomaly (positive). Accuracy is calculated by dividing the summation of true positives and true negatives over all possible outcomes. An alternative performance index is the balanced accuracy, which is equal to the average of *sensitivity* and *specificity*. While accuracy and balanced accuracy give equal importance to the regular and the anomaly traffic, F-Score captures the harmonic mean of both *sensitivity* and *precision*:

$$\text{F-Score} = 2 \times \frac{precision \times sensitivity}{precision + sensitivity}.$$

In a two-way classification, all anomalies are treated as one class. Its performance is shown in Table VII. SVM_3 achieves the best F-Score (86.1%) using MIQ selected features. We check validity of the proposed models by also applying two-way SVM classification on BGP traffic trace collected from the BCNET [13] on December 20, 2011. All data points in the BCNET traffic trace are labeled as regular traffic. Hence, $y = -1$. The classification accuracy of 79.2% indicates the number of data points that are classified as regular traffic. Since all data points in BCNET and RIPE test datasets contain no anomalies, they have low sensitivities and low F-Scores. Hence, we calculated instead accuracy as the performance measure. Data points that are classified as anomalies (false positive) are shown in Fig. 4. The best two-way classification result is achieved by using SVM_2.

TABLE VII: Performance of the two-way SVM classification.

SVM	Feature	Performance index			
		Accuracy (%)			F-Score (%)
		Test dataset	RIPE normal	BCNET	Test dataset
SVM_1	All features	64.1	55.0	62.0	63.2
SVM_1	Fisher	72.6	63.2	58.5	73.4
SVM_1	MID	63.1	52.2	59.4	61.2
SVM_1	MIQ	60.7	47.9	61.7	57.8
SVM_1	MIBASE	79.1	74.3	60.9	80.1
SVM_2	All features	68.6	97.7	79.2	22.2
SVM_2	Fisher	67.4	96.6	74.8	16.3
SVM_2	MID	67.9	97.4	72.5	19.3
SVM_2	MIQ	67.7	97.5	76.2	15.3
SVM_2	MIBASE	67.5	96.8	78.8	17.8
SVM_3	All features	81.5	92.0	69.2	84.6
SVM_3	Fisher	89.3	93.8	68.4	75.2
SVM_3	MID	75.4	92.8	71.7	79.2
SVM_3	MIQ	85.1	92.2	73.2	86.1
SVM_3	MIBASE	89.3	89.7	69.7	80.1

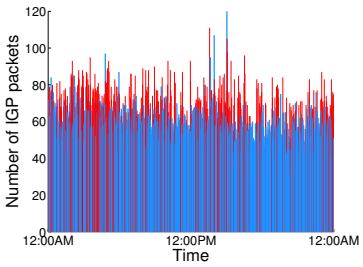

Fig. 4: Shown in red is incorrectly classified (anomaly) traffic collected on December 20, 2011.

We extended the proposed classifier to implement multi-class SVMs and used one-versus-one multi-class classification [14] on four training datasets: Slammer, Nimda, Code Red I, and RIPE. Data points are classified by $n(n-1)/2$ classifiers, where n is the number of classes. The four-way classification detects and classifies the specific type of traffic: Slammer, Nimda, Code Red I, or regular. Classification performance is shown in Table VIII. BCNET dataset is also tested using the multi-class SVM and achieved 91.4% accuracy.

TABLE VIII: Accuracy of the four-way SVM classification.

Feature	Average accuracy (%)	
	RIPE	BCNET
All features	77.1	91.4
Fisher	82.8	85.7
MID	67.8	78.7
MIQ	71.3	89.1
MIBASE	72.8	90.2

Test data points from the Slammer worm that are incorrectly classified in the two-way classification (false positive and false negative) are shown in Fig. 5 (left). Correctly classified as anomaly (true positive) with calculated *sensitivity* of 88.3% during the 16 hours time interval are shown in Fig. 5 (right).

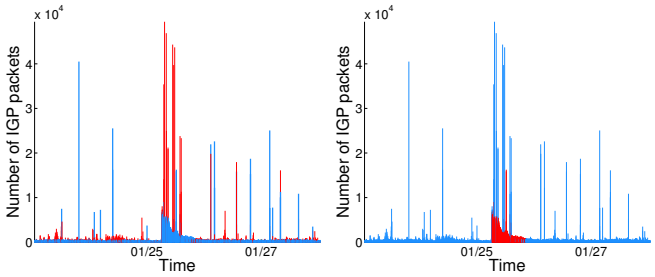

Fig. 5: Shown in red are incorrectly classified regular and anomaly traffic (left) and correctly classified anomaly traffic (right) on January 25, 2003.

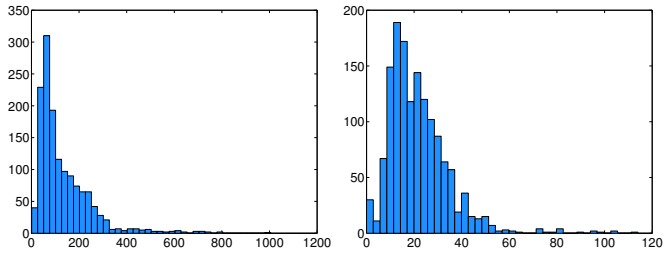

Fig. 6: Distribution of the number of BGP announcements (left) and withdrawals (right) for the Code Red I worm.

IV. Classification using Hidden Markov Models

The second model for classification is based on the first order HMMs. HMMs are statistical tools used to model stochastic processes that consist of two embedded processes: the observable process that maps BGP features and the unobserved hidden process that has the Markov property. We assume that the observations are independent and identically distributed. Even though the HMMs considered in this paper belong to non-parametric supervised classification methods, we used 10-fold cross validation to select number of hidden states as a parameter in order to improve the accuracy of the model. We implemented the HMMs using the MATLAB statistical toolbox.

Each HMM model is specified by a tuple $\lambda = (N, M, \alpha, \beta, \pi)$, where:

N = number of hidden states (cross-validated)
M = number of observations (11)
α = transition probability distribution $N \times N$ matrix
β = emission probability distribution $N \times M$ matrix
π = initial state probability distribution matrix.

The proposed detection model consists of three stages:

- *Sequence extractor and mapping*: All features are mapped to 1-D observation vector.
- *Training*: Two HMMs for two-way classification and four HMMs for four-way classification are trained to identify the best α and β for each class. HMMs are trained and validated for various number of hidden states N.
- *Classification*: Maximum likelihood probability $p(x|\lambda)$ is used to classify the test observation sequences.

In the sequence extraction stage, the BGP feature matrix is mapped to a sequence of observations by adding the BGP announcements (feature 1) to the BGP withdrawals (feature 2). We also add the maximum AS-PATH length (feature 6) to the maximum edit distance (feature 12). In both cases, we divide the result to eleven observations using a logarithmic scale, which solves the high skew of heavy tailed probability distribution of the BGP *volume* features in the training datasets. The distribution for BGP announcements during the Code Red I worm attack is shown Fig. 6.

HMMs are trained and validated for various number of hidden states. A 10-fold cross-validation with the Balm-Welch algorithm [12] is used for training to find the best α and β for each HMM. The best transition and emission matrices are validated by obtaining the largest maximum likelihood probability $p(x|\lambda_{\text{HMM}_x})$. We construct six and twelve HMM models for two-way and four-way classifications, respectively. Various HMMs are listed in Table IX and Table X. We evaluate the test observation sequences for each HMM and calculate its maximum likelihood probability.

In the classification stage, each test observation sequence is classified based on the largest maximum likelihood probability for HMMs with the same number of hidden states. For example, HMM_1, HMM_4, HMM_7, and HMM_{10} shown in Table X correspond to HMMs with two hidden states for various training datasets.

TABLE IX: HMMs: two-way classification.

	Number of hidden states		
Training dataset	2	4	6
Slammer, Nimda, and Code Red I	HMM_1	HMM_2	HMM_3
RIPE/BCNET	HMM_4	HMM_5	HMM_6

TABLE X: HMMs: four-way classification.

	Number of hidden states		
Training dataset	2	4	6
Slammer	HMM_1	HMM_2	HMM_3
Nimda	HMM_4	HMM_5	HMM_6
Code Red I	HMM_7	HMM_8	HMM_9
RIPE/BCNET	HMM_{10}	HMM_{11}	HMM_{12}

The accuracy of each HMM is defined as:

$$\frac{Number\ of\ correctly\ classified\ observation\ sequences}{Total\ number\ of\ observation\ sequences}. \quad (6)$$

The numerator is calculated using the highest maximum likelihood probability $p(x|\lambda_{\text{HMM}_x})$. Sequences in the denominator share the same number of hidden states. The correctly classified observation sequence is generated by a model that has the highest probability when tested with itself.

We use RIPE and BCNET datasets to test the three anomalies. Two sets of features (*volume*) and (*AS-path*) are mapped to create one observation sequence for each HMM. We mapped *volume* feature set (1, 2) and *AS-path* feature set (6, 12) to two observation sequences. HMMs have better F-Score using set (1, 2) than set (6, 12), as shown in Table XI. The RIPE and BCNET test datasets have the highest F-Score when tested using HMMs with two hidden states.

TABLE XI: Accuracy of the two-way HMM classification.

		Performance index			
		Accuracy (%)		F-Score (%)	
N	Feature set	RIPE	BCNET	RIPE	BCNET
2	(1,2)	86.0	94.0	84.4	93.8
2	(6,12)	79.0	71.0	76.2	60.7
4	(1,2)	78.0	87.0	72.2	85.0
4	(6,12)	64.0	60.0	48.0	35.9
6	(1,2)	85.0	91.0	84.3	90.1
6	(6,12)	81.0	65.0	80.1	50.2

Similar tests are applied using RIPE and BCNET datasets with four-way HMM classification. The classification accuracies are averaged over four HMMs for each dataset and are listed in Table. XII.

TABLE XII: Accuracy of the four-way HMM classification.

		Average accuracy (%)	
N	Feature set	RIPE	BCNET
2	(1,2)	72.50	77.50
2	(6,12)	38.75	41.25
4	(1,2)	66.25	76.25
4	(6,12)	26.25	33.75
6	(1,2)	70.00	76.25
6	(6,12)	43.75	42.50

V. Discussion

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. As shown in Table V, 65% of the selected features are *volume* features. Hence, they are more relevant to the anomaly class than the *AS-path* features, which confirms the known effect of BGP anomalies on the volume of the BGP announcements.

The top selected *AS-path* features appear on the boundaries of the distributions shown in Fig. 3. For example, *AS-path* features 25, 32, and 24 have the highest Fisher, MID, and MIQ scores, respectively. 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. Approximately 58% of the *AS-path* features shown in Table V are larger than the distribution mean. For example, 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 [8]. 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 [15]. The SVM models exhibited better performance than the HMMs in two-way and four-way classifications. The SVM models based on Code Red I and Nimda datasets and the HMMs with two hidden states have the highest accuracies. HMMs based on the number of announcements and number of withdrawals (feature 1 and feature 2) offer better accuracy in two-way and four-way classifications than models with the maximum number of AS-PATH length (feature 6) and the maximum edit distance (feature 12). Both SVM and HMM two-way classifications produced better results than four-way classifications because of the common semantics among BGP anomalies. For example, BGP Slammer is more correlated to Nimda than to regular RIPE mapped sequence.

VI. Conclusions

We have investigated BGP anomalies and proposed detection models based on the SVM and HMM classifiers. Classification results show that the best achieved F-Scores of the SVM and HMM models are 86.1% and 84.4%, respectively. Furthermore, *volume* mapped sequences generate models with better accuracy than *AS-path* mapped sequences. Hence, using the BGP *volume* features is a viable approach for detecting possible worm attacks. Since BGP anomalies have similar properties and effect on BGP features, the proposed models may be used as online mechanisms to predict new BGP anomalies and detect the onset of worm attacks.

References

[1] S. Deshpande, M. Thottan, T. K. Ho, and B. Sikdar, "An online mechanism for BGP instability detection and analysis," *IEEE Trans. Computers*, vol. 58, no. 11, pp. 1470–1484, Nov. 2009.

[2] 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, pp. 55–66, Oct. 2005.

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

[4] University of Oregon Route Views project [Online]. Available: http://www.routeviews.org/.

[5] T. Manderson, "Multi-threaded routing toolkit (MRT) border gateway protocol (BGP) routing information export format with geo-location extensions," RFC 6397, *IETF*, Oct. 2011 [Online]. Available: http://www.ietf.org/rfc/rfc6397.txt.

[6] Zebra BGP parser [Online]. Available: http://www.linux.it/~md/software/zebra-dump-parser.tgz.

[7] D. Meyer, "BGP communities for data collection," RFC 4384, *IETF*, 2006 [Online]. Available: http://www.ietf.org/rfc/rfc4384.txt.

[8] 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 Measurment*, New York, NY, USA, 2002, pp. 183–195.

[9] Y.-W. Chen and C.-J. Lin, "Combining SVMs with various feature selection strategies," *Strategies*, vol. 324, no. 1, pp. 1–10, Nov. 2006.

[10] 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 Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[11] Libsvm–a library for support vector machines [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer-Verlag, 2006.

[13] BCNET [Online]. Available: http://www.bc.net.

[14] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[15] 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.