Machine Learning Models for Classification of BGP Anomalies

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

June 26, 2012

Roadmap

Contribution

- Slammer, Nimda, and Code Red I anomalies affect performance of the global Internet Border Gateway Protocol (BGP)
- BGP anomalies also include: Internet Protocol (IP) prefix hijacks, miss-configurations, and electrical failures
- We introduce new BGP features and apply Support Vector Machine (SVM) models and Hidden Markov Models (HMMs) to design anomaly detection mechanisms
- We apply 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

Datasets sources

- The RIPE and Route Views BGP update messages: multi-threaded routing toolkit (MRT) binary format
- Validity of the proposed models was checked by also using BGP traffic trace collected from the BCNET

	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

References

- RIPE RIS raw data [Online]. Available: `http://www.ripe.net/data-tools/stats/ris/ris-raw-data`.
- University of Oregon Route Views project [Online]. Available: `http://www.routeviews.org/`.
- BCNET [Online]. Available: `http://www.bc.net`.

Sampling and normalization of features

- Features are sampled every minute over five days, producing 7,200 samples for each anomaly event
- Features are normalized to have zero mean and unit variance
- This normalization reduces the effect of the Internet growth between 2003 and 2011

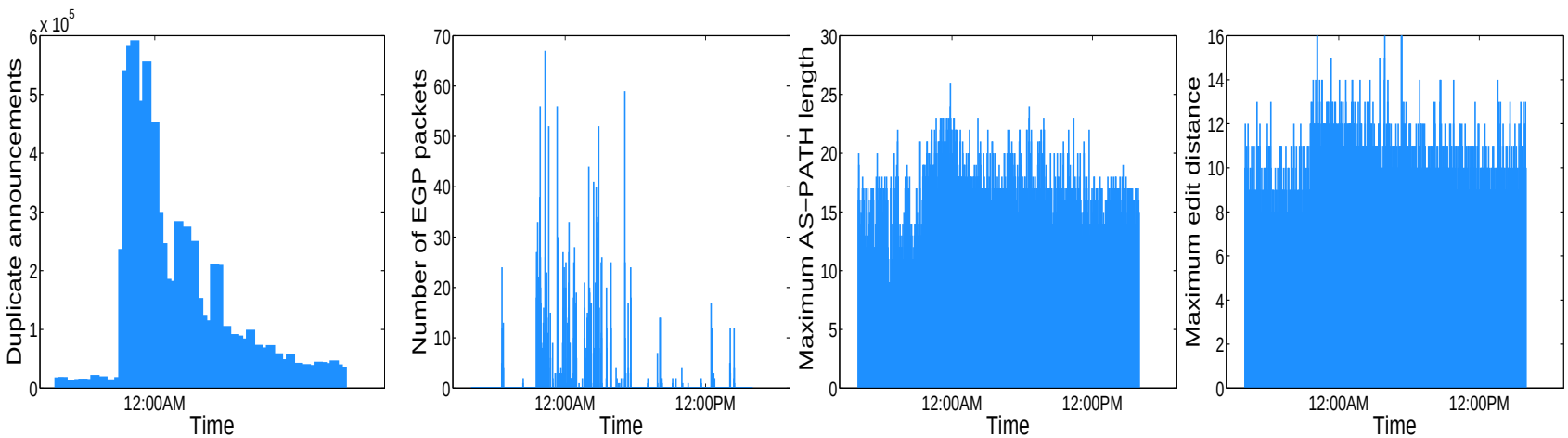
List of extracted features

- Extracted features: *volume* (number of BGP announcements) and *AS-path* (maximum edit distance) 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4	Number of withdrawn NLRI prefixes	*volume*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*

Samples of extracted BGP features during the Slammer worm attack

- *Volume* features 8 and 35
- *AS-path* features 6 and 12



BGP messages

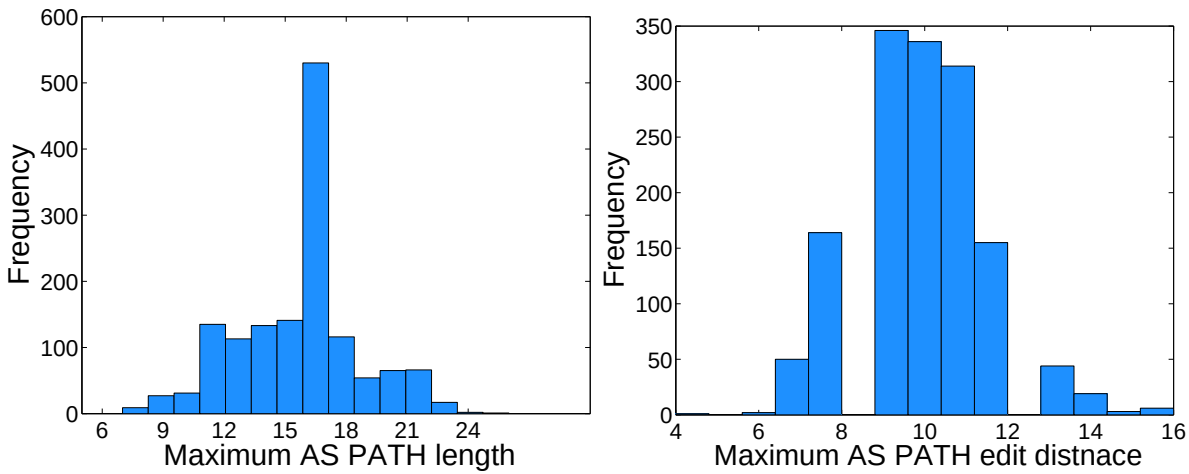
- BGP protocol generates four types of messages:
  - open,update, keepalive, and notification
- Only BGP update messages were considered
- BGP update messages: announcements or withdrawals

References

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

Definition of BGP features: sample

- Features 14 to 33: the most frequent values of the maximum AS-PATH length and the maximum edit distance
- Distributions of (left) the maximum AS-PATH length and (right) the maximum edit distance during the Slammer worm attack:



Feature selection algorithms

- Features scoring algorithms:
  - Fisher
  - minimum Redundancy Maximum Relevance (mRMR)
- These algorithms measure the correlation and relevancy among features
- The top ten features were selected for the Fisher feature selection

References

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

Fisher algorithm

- Training datasets: a real matrix $\mathbf{X}_{7200\times 37}$.
- Column vector $\mathbf{X}_k, k = 1, ..., 37$ corresponds to one feature
- The Fisher score for $\mathbf{X}_k$:

$$\text{F-score} = \frac{m_a^2 - m_r^2}{s_a^2 + s_r^2}$$

$$= \frac{\frac{1}{N_a}\sum_{i\in anomaly} {x_{ik}}^2 - \frac{1}{N_r}\sum_{i\in regular} {x_{ik}}^2}{\frac{1}{N_a}\sum_{i\in anomaly} {(x_{ik} - m_a)}^2 + \frac{1}{N_r}\sum_{i\in regular} {(x_{ik} - m_r)}^2},$$

- N_a and N_r: number of anomaly and regular data points
- m_a and s_a^2 (m_r and s_r^2): the mean and the variance of anomaly (regular) class

Fisher algorithm

- 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
- mRMR algorithm: maximizes the relevance of features with respect to the target class while minimizing the redundancy among features
- Variants of the mRMR algorithm:
  - Mutual Information Difference (MID)
  - Mutual Information Quotient (MIQ)
  - Mutual Information Base (MIBASE)

mRMR algorithm

- mRMR relevance between a feature set $S = \{\mathbf{X}_1, ..., \mathbf{X}_k, \mathbf{X}_l, ..., \mathbf{X}_{37}\}$ and a class vector $\mathbf{Y}$ is based on the mutual information function $\mathcal{I}$:

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

- Criteria for mRMR variants:
  - MID: $\max [V(\mathcal{I}) - W(\mathcal{I})]$
  - MIQ: $\max [V(\mathcal{I})/W(\mathcal{I})]$

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

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

mRMR algorithm

- MIBASE: ordered based on the $\mathcal{I}(X_k, X_l)$ function

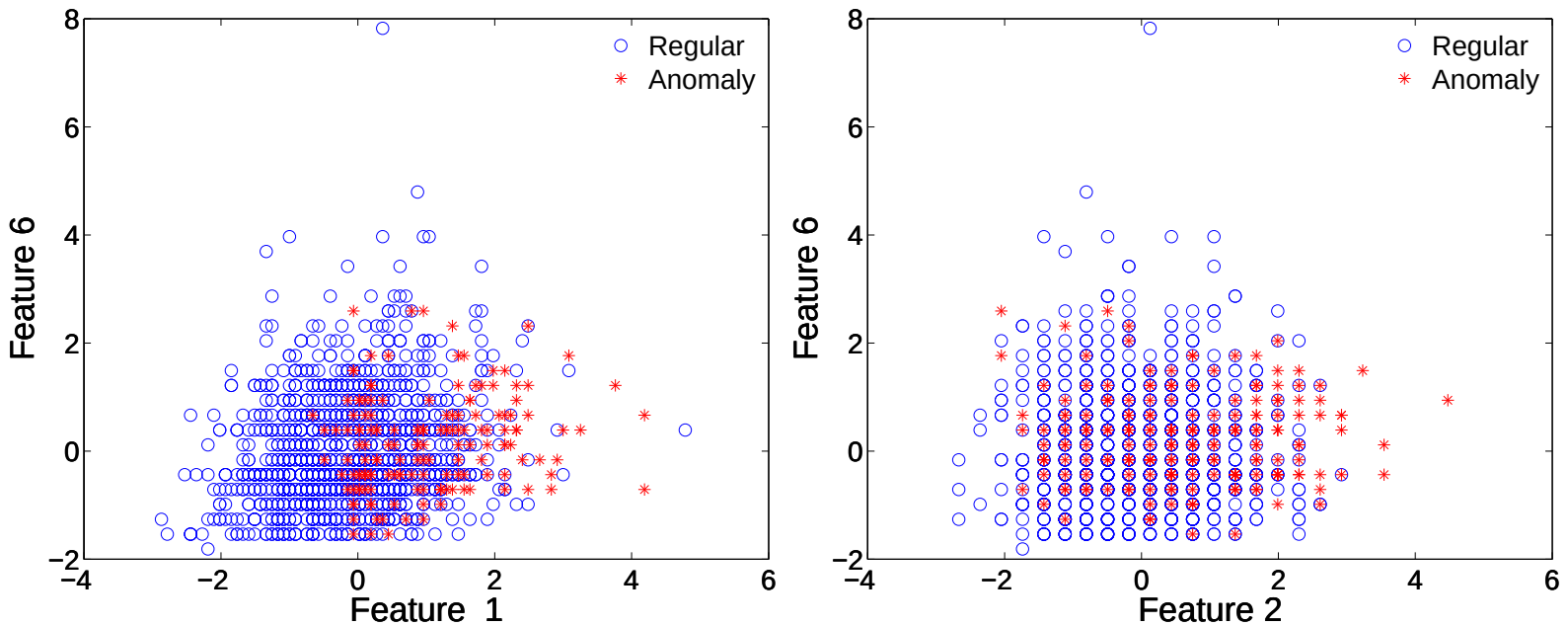
- The set of features was captured on January 25, 2003
- Test set contains 1,440 samples where 869 samples are labeled as anomaly

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

Normalized scattering graphs

- Feature 1, Feature 2, and Feature 6:



- Selecting appropriate combination of features is essential for a an accurate classification

Definitions

- We considered: accuracy, balanced accuracy, and F-score
- Definitions:
  - True positive (TP): is number of anomalous training data points that are classified as anomaly
  - True negative (TN): is number of regular training data points that are classified as regular
  - False positive (FP): is number of regular training data points that are classified as anomaly
  - False negative (FN): is number of anomalous training data points that are classified as regular

Performance measures and indices

- Performance measures:

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

- Performance indices:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{balanced accuracy} = \frac{\text{sensitivity} + \text{precision}}{2}$$

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

Performance indices

- Sensitivity: ability of the model to identify the anomalies (TP) among all labeled anomalies (true).
- Precision: ability of the model to identify the anomalies (TP) among all data points that were identified as anomalous (positive)
- Accuracy: treats the regular data points as important as anomalous training data (not a good performance measure)
- F-score:
  - often used as a performance index to compare classification models
  - the harmonic mean of the sensitivity and the precision
  - reflects the success of detecting anomalies rather than detecting both anomalies and regular data points

Feature matrix

- Support vector machines were introduced by V. Vapnik in 1970s
- SVMs perform more accurately for datasets with high dimensional complexity
- For each training dataset $\mathbf{X}_{7200\times37}$, we target two classes: anomaly (true) and regular (false)
- Dimension of feature matrix: $7,200 \times 10$
- Each row contains the top ten selected features within the one-minute interval

References

- Support Vector Machine - The Book [Online]. Available: `http://www.support-vector.net/chapter_6.html`.
- Libsvm–a library for support vector machines [Online]. Available: `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

SVM solution

- SVM solves a loss function as an optimization problemwith the constraints:

$$\min C \sum_{m=1}^{M} \xi_m + \frac{1}{2}\|w\|^2$$

$$t_m y(\mathbf{X}_m) \geq 1 - \xi_m$$

  - constant $C > 0$ controls the importance of the margin
  - slack variable ξ_m solves the non-separable data points classification problem
  - regularization parameter $\frac{1}{2}\|w\|^2$: used to avoid over-fitting
  - $\mathbf{X}_m$ corresponds to a row vector where $m = 1, ..., 7200$
  - training target class t_m: 1 (anomaly), -1 (regular)
  - tested target class y: 1 (anomaly), -1 (regular)

References

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

SVM two-way datasets

		Test dataset		
SVM	Training dataset	Code Red I	Nimda	Slammer
SVM_1	Slammer and Nimda	$\surd$	x	x
SVM_2	Slammer and Code Red I	x	$\surd$	x
SVM_3	Code Red I and Nimda	x	x	$\surd$

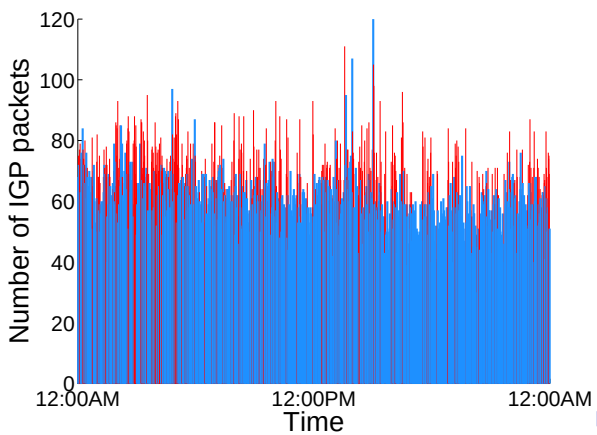
Two-way classification: performance

- All anomalies are treated as one class

		Performance index			
		Accuracy (%)			F-score (%)
SVM	Feature	Test dataset	RIPE	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

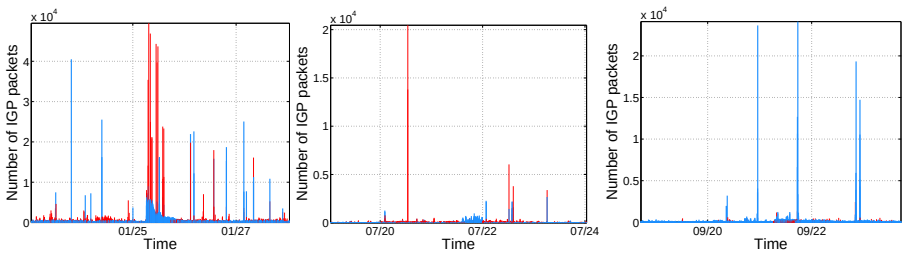
Classification results

- SVM_3 achieves the best F-score (86.1%) using features selected by MIQ
- SVM_2: the best overall two-way classifier
- BCNET and RIPE test datasets contain no anomalies and have low sensitivities and low F-scores
- Performance measure: accuracy
- Incorrectly classified (anomaly) BCNET traffic collected on December 20, 2011 (red):

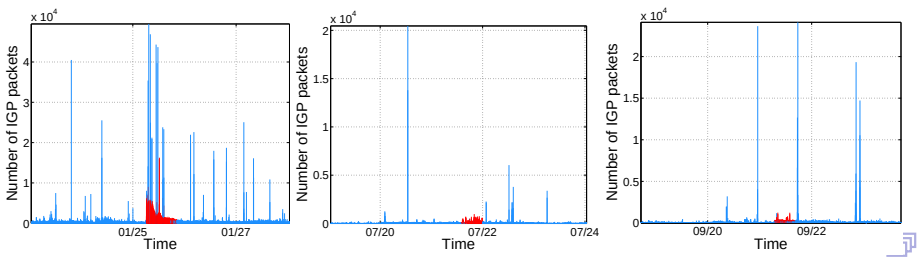


Classification results

- Incorrectly classified regular and anomaly traffic:
  - Slammer (left)
  - Code Red I (middle)
  - Nimda (right)



- Correctly classified anomaly traffic:
  - Slammer (left)
  - Code Red I (middle)
  - Nimda (right)



Four-way classification: performance

- Multi-class SVMs are used on training datasets: Slammer, Nimda, Code Red I, and RIPE/BCNET

	Average accuracy (%)	
Feature	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

References

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.

Hidden Markov Models

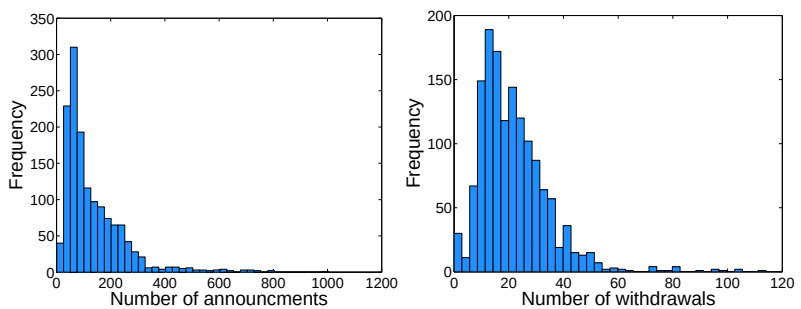
- First order HMMs are used to model stochastic processes that consist of two embedded processes:
  - observable process that maps BGP features
  - unobserved hidden process that has the Markov property
- Assumption: observations are independent and identically distributed

HMM classification stages

- HMM model is specified by a tuple $\lambda = (N, M, \alpha, \beta, \pi)$:
  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
  - *Classification*: maximum likelihood probability $p(x|\lambda)$ is used to classify the test observation sequences.

Sequence extraction and mapping

- BGP feature matrix is mapped to a sequence of observations by adding:
  - BGP announcements (Feature 1) to BGP withdrawals (Feature 2)
  - maximum AS-PATH length (Feature 6) to the maximum edit distance (Feature 12)
- In both cases, results are divided to eleven observations using a logarithmic scale (solves the high skew of heavy tailed probability distribution of the BGP *volume* features)



Training

- HMMs are trained and validated for various number of hidden states
- The best α and β for each HMM was found by using 10-fold cross-validation with the Balm-Welch algorithm
- Validated by obtaining the largest maximum likelihood probability $p(x|\lambda_{\mathrm{HMM}_x})$

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

Classification

- Six (two-way) and twelve (four-way) HMM models were constructed
- Test observation sequences were evaluated using maximum likelihood probability
- 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

- 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}

Classification

- HMMs with the same number of hidden states are compared
- Example: HMM_1, HMM_4, HMM_7, and HMM_{10} correspond to HMMs with two hidden states for various training datasets
- HMM accuracy:

$$\frac{\textit{Number of correctly classified observation sequences}}{\textit{Total number of observation sequences}}$$

Classification

- The correctly classified observation sequence is generated by a model that has the highest probability when tested with itself
- RIPE and BCNET were datasets to test the three anomalies.
- Two sets of features (*volume*) and (*AS-path*) are mapped to create one observation sequence for each HMM
- *Volume* feature set (1, 2) and *AS-path* feature set (6, 12) are mapped to two observation sequences.
- RIPE and BCNET test datasets have the highest F-score when tested using HMMs with two hidden states

Two-way classification: performance

		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

- HMMs have better F-score using set (1, 2) than set (6, 12)

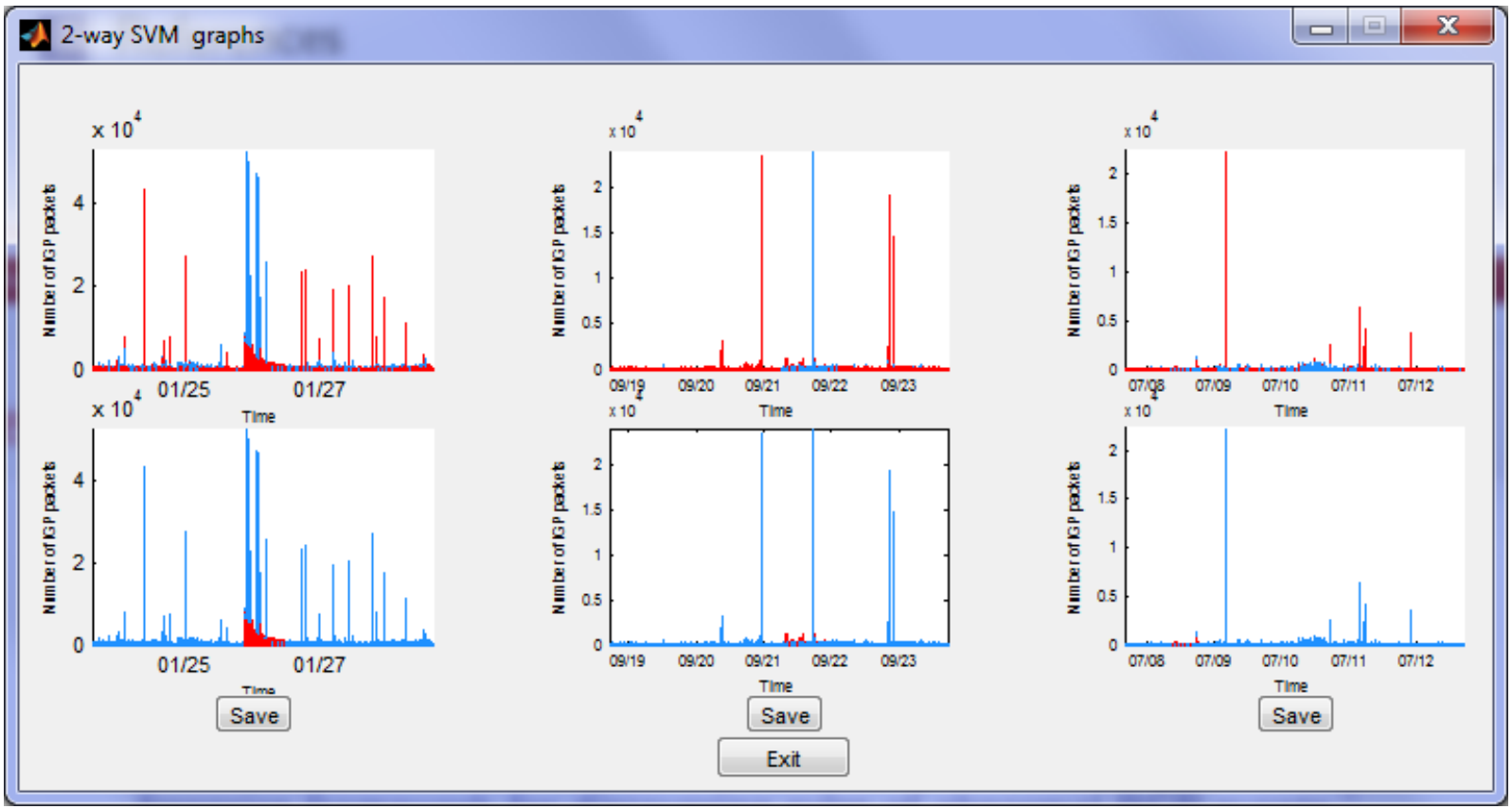
Four-way classification: performance

- 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

		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

BGPAD

- Displays anomalous traffic



Discussion: feature extraction and selection

- The trust relationship among BGP peers is vulnerable during anomaly attacks
- Example: during BGP hijacks, a BGP peer may announce unauthorized prefixes that indicate to other peers that it is the originating peer
- Effect of anomalies on *volume* features:
  - False announcements propagate across the Internet and affect the number of BGP announcements (updates and withdrawals)

Discussion: feature extraction and selection

- Effect of anomalies on *AS-path* features:
  - large length of the AS-PATH BGP attribute implies that the packet is routed via a longer path to its destination
  - 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
  - edit distance and AS-PATH length of the BGP announcements tend to have a very high or a very low value (large variance)
- The top selected *AS-path* features appear on the boundaries of the distributions: *AS-path* features 25, 32, and 24 have the highest Fisher, MID, and MIQ scores

Discussion: classification

- SVM models exhibited better performance than the HMMs in two-way and four-way classifications
- 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 than models with the maximum number of AS-PATH length (Feature 6) and the maximum edit distance (Feature 12)
- SVM and HMM two-way classifications produced better results than four-way classifications because of the common semantics among BGP anomalies

Conclusions

- We have investigated BGP anomalies and proposed detection models based on the SVM and HMM classifiers
- The best achieved F-scores: SVM (86.1%), HMM(84.4%)
- *Volume* features are more relevant to the anomaly class than the *AS-path* features (known effect)
- Using the BGP *volume* features is a viable approach for detecting possible worm attacks
- The proposed models may be used as online mechanisms to predict new BGP anomalies and detect the onset of worm attacks