DIMACS Technical Report 2008-08
August 2008

Anomaly Detection by Reasoning from Evidence in Mobile Wireless Networks

by

Nikita I. Lytkin
Dept. of Computer Science
Rutgers University
New Brunswick, New Jersey 08903

William M. Pottenger
DIMACS and Department of Computer Science
Rutgers University
New Brunswick, New Jersey 08903

Ilya B. Muchnik
DIMACS
Rutgers University
New Brunswick, New Jersey 08903

ABSTRACT

Anomaly detection is concerned with identification of abnormal patterns of behavior of a system. Traditional supervised machine learning methods of classification rely on training data in the form of labeled data instances representative of each class (e.g. normal vs anomalous data). Clustering methods, on the other hand, do not require a priori knowledge of how anomalies are represented in the data space, and are therefore particularly suitable for anomaly detection.

Partitional clustering methods such as K-means require the number K of clusters to be specified by a user. Three heuristics that rely on a joint use of two partitional clustering methods for determining an appropriate number of clusters in a dataset are proposed in this work. The heuristics were first evaluated on synthetic data and then applied on real-world data from the domain of computer network security. Experimental results demonstrated that clustering methods are adequate for detection of large-scale anomalous events in the Internet. Scalability of the heuristics across domains of application was indicated by additional experimental results obtained on several datasets from the UCI machine learning repository.

1 Introduction

Anomaly detection is concerned with identification of abnormal patterns of behavior of a system. Traditional supervised machine learning methods of classification rely on training data in the form of labeled data instances representative of each class (e.g. normal vs anomalous data). Clustering methods, on the other hand, do not require a priori knowledge of how anomalies are represented in the data space, and are therefore particularly suitable for anomaly detection.

Partitional clustering methods such as K-means require the number K of clusters to be specified by a user. Often, an appropriate value for K is estimated based on the domain knowledge. In the context of anomaly detection, however, it may not be possible to specify K in advance since that would require the knowledge of how many different anomalies are expected.

We propose three heuristics for determining an appropriate number of clusters for a dataset. The heuristics rely on a joint use of two clustering methods. One of the methods is K-means [8, 10] whose criterion can be written [13] as

$$I_1 = \sum_{\alpha=1}^{K} p_\alpha \sigma_\alpha^2, \tag{1}$$

where p_α is the probability (i.e., weight) of cluster α and σ_α^2 is its variance. The other clustering method, call it I_2, was recently proposed by [11] and is based on a criterion similar to (1), namely

$$I_2 = \sum_{\alpha=1}^{K} p_\alpha \sigma_\alpha. \tag{2}$$

Unlike (1), criterion (2) produces non-linear discriminant surfaces (i.e., cluster boundaries) when cluster standard deviations σ_α are unequal. A thorough study of criterion (2) and an algorithm of its optimization can be found in [11]. Both, K-means and I_2 clustering methods aim to minimize their respective criteria (1) and (2). The minimization corresponds to partitioning the data space so that points in the same cluster are close to each other and far apart from points in other clusters.

The paper is organized as follows. Relevant work on determining an appropriate number of clusters for a dataset is discussed in Section 2. The proposed heuristics are described in Section 3. Experimental results of application of the heuristics to synthetic as well as real-world network security data are presented in Section 4. Concluding remarks are made in Section 5.

2 Related Work

Selection of the number K of clusters in a dataset is closely related to evaluation of cluster validity. Two major types of indices [13] could be distinguished for assessing the validity of a cluster structure discovered in the data. *External* indices take into account (domain-specific) information that was not directly employed by the clustering method. Such information could, for example, be supplied in the form of true class labels of the data instances. The adequacy of the cluster structure could then be evaluated based on a number of indices that measure the similarity between partitions of objects. A comprehensive exposition of partition-partition similarity measures can be found in [8, 13].

As was mentioned earlier, the distribution of data instances characterizing different types of abnormal behavior of a system is not known in advance. Moreover, the number of types of anomalies is itself unknown. Therefore, we cannot rely on external indices for determining an appropriate value of K in the context of anomaly detection.

Internal indices of cluster validity rely solely on the data used by the clustering method and measure the fit between the cluster structure and the data. Clustering criterion is a natural internal index whose usage for determining the value of K has previously been considered [4, 13]. In this case, value of the criterion is plotted as a function of K. The plot is then manually examined for an apparent "elbow" in the graph and the corresponding value of K is chosen as appropriate. Extensive reviews of other (less subjective) approaches to automatic determination of the number of clusters using internal indices can be found in [4, 12, 18]. However, the results of evaluations [4, 18] indicate that internal indices are often sensitive to the overall structure of the data and can not be applied universally.

Our work differs from the index-based approaches in that we employ an ensemble of clustering methods and consider the actual clusterings produced by each method for various values of K. The details of our approach are presented below.

3 Determining the Number of Clusters

K-means clustering as well as I_2 is sensitive to the initial settings. Given an initial partition of the data into K clusters, each of the algorithms converges to a locally optimal solution, i.e., to a clustering on which the corresponding clustering criterion achieves a local minimum. A commonly used approach to reaching a deeper value of the criterion and thus finding a better clustering involves running the clustering algorithm L times, each time starting from a different randomly generated initial partition. Here, L is a parameter specified by the user. Out of all the clusterings found during the L runs, a clustering giving the minimum value of the criterion is selected as "the best" clustering.

Heuristics presented in this section for determining an appropriate number of clusters for a given dataset are based on the following hypothesis. If the data is arranged in a (uniform) cloud of points without any prominent clustering structure, we would expect each of the L runs of a clustering algorithm to converge to a different solution regardless of the value of parameter K. On the other hand, if some clustering structure is present in the data and parameter K has been set appropriately, we would expect the algorithm to converge either to a single solution or to a small set of solutions over the L runs.

Our approach to determining the number of clusters relies on examination of histograms that show for each unique solution found during the L runs of an algorithm, the number of times the solution has been discovered, i.e., its frequency of occurrence.

We employ K-means in conjunction with I_2 for determining an appropriate number of clusters for a dataset. The two algorithms act as a committee of experts each of which has a different preference for grouping points, but both of which strongly agree with each other when the data has a clearly defined clustering structure. The description of the heuristics follows.

3.1 Heuristic One

The first heuristic lies in choosing the largest number K of clusters such that histograms for K-means and I_2 contain pronounced peaks and that the most frequent solutions give the minimum values of the corresponding clustering criteria. Figure 1 shows histograms of frequencies of discovery of unique clusterings by K-means and I_2 on a dataset comprised of three separable clusters, each generated according to a different Gaussian distribution with diagonal covariance matrix. As can be seen in Figs. 1c and 1d, seven and fifteen unique solutions were found by K-means and I_2, respectively, for $K = 3$. The most frequent solution found by K-means appeared in 87% of the runs, while the most frequent solution found by I_2 appeared in 78% of the runs. In both cases, the most frequent solution gave the minimum value of the corresponding clustering criterion.

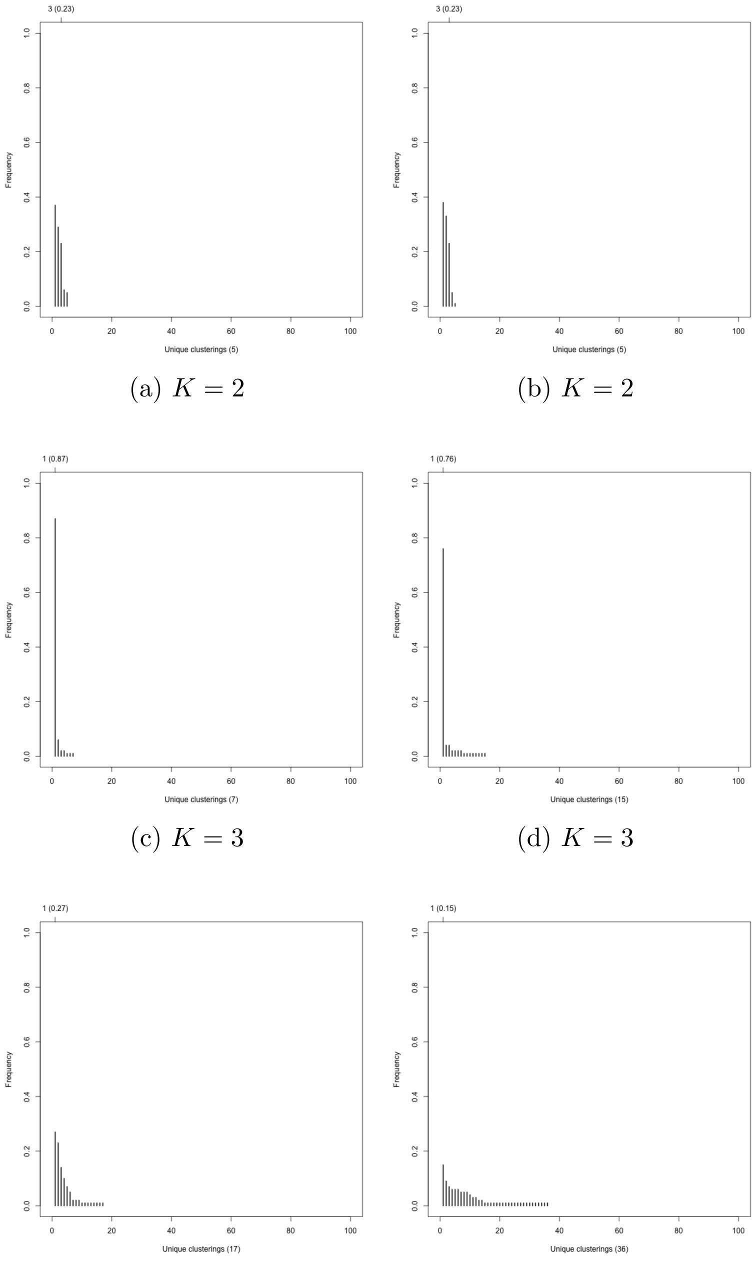
Note that histograms for $K \neq 3$ (Figs. 1a, 1b, 1e and 1f) were much smoother than for $K = 3$. A smooth histogram indicates that for a given K, the algorithm was not able to identify a clear cluster structure within the dataset, and that a different number of clusters may be more appropriate.

Figure 2 shows histograms that were generated on a one-cluster dataset. These histograms did not exhibit the peaked behavior seen in Figure 1. Lack of peaks in the histograms and a high number of unique solutions found for small values of K suggested that the dataset did not contain any separable clusters, which in fact was the case.

3.2 Heuristic Two

The second heuristic applies in situations where for a given K, a pronounced sharp peak is present in histograms for both clustering methods, but the solution minimizing one of the two clustering criteria is not the most frequent in the corresponding histogram. We demonstrate this heuristic on a dataset containing six separable clusters, each generated according to a different Gaussian distribution with a diagonal covariance matrix. The histograms are presented in Fig. 3.

As shown in Fig. 3c, the best clustering according to K-means criterion was ranked third by frequency and was found in 9% of the runs. However, Fig. 3d shows that the best clustering according to I_2 was ranked first by frequency and was found in 35% of the runs. Figures 3a and 3b show a reverse situation where the best K-means clustering was ranked first by its frequency of 36%, while the best I_2 clustering was ranked seventeenth by its frequency of 1%.



(a) $K = 2$

(b) $K = 2$

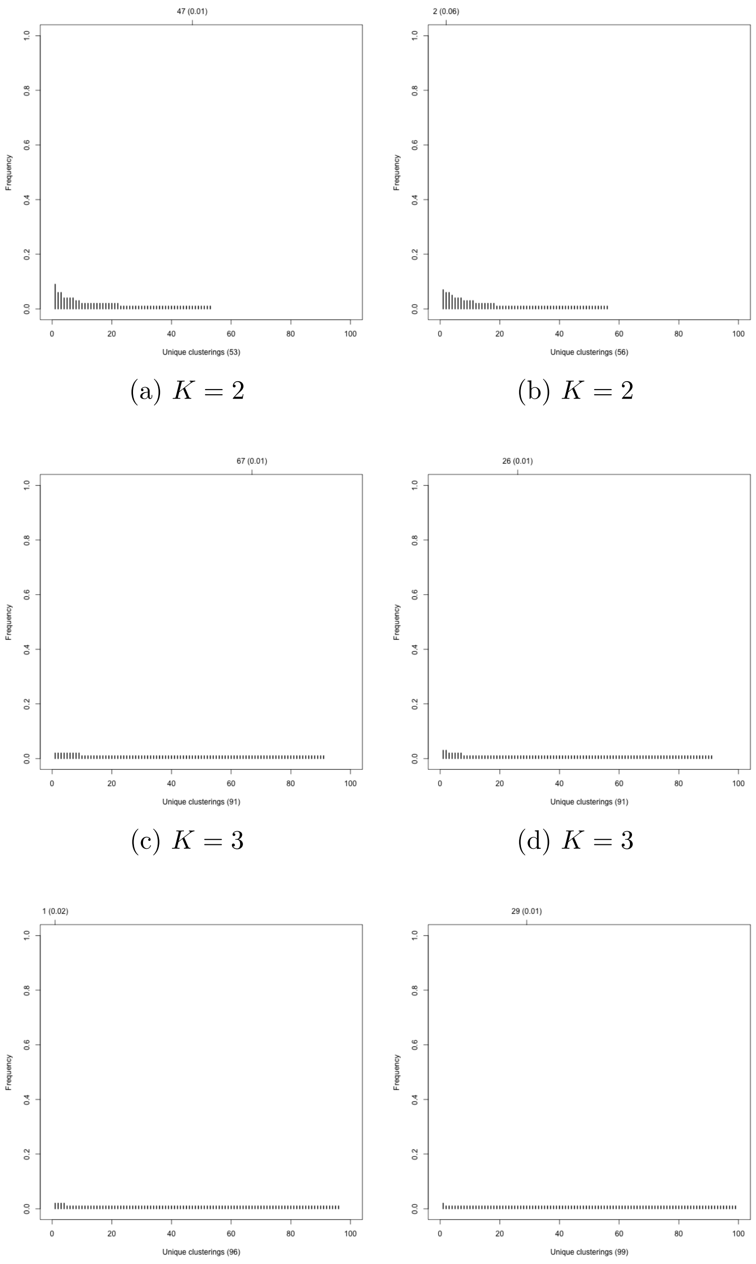
(c) $K = 3$

(d) $K = 3$

(e) $K = 4$

(f) $K = 4$

Figure 1: Histograms of frequencies of occurrence of unique clusterings discovered by K-means (left column) and I_2 (right column) over 100 runs on a three-cluster dataset. Clusterings are ranked by decreasing frequency. Rank and frequency of the clustering giving the minimal value of the corresponding criterion for a particular value of K are marked at the top of each histogram.



(a) $K = 2$

(b) $K = 2$

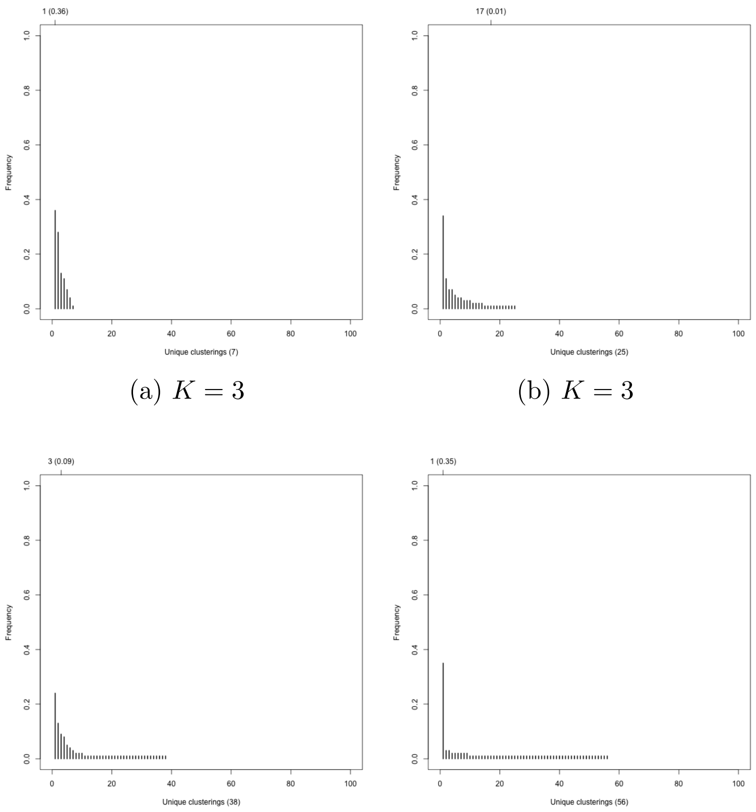
(c) $K = 3$

(d) $K = 3$

(e) $K = 4$

(f) $K = 4$

Figure 2: Histograms of frequencies of occurrence of unique clusterings discovered by K-means (left column) and I_2 (right column) over 100 runs on a one-cluster dataset



(a) $K = 3$

(b) $K = 3$

(c) $K = 6$

(d) $K = 6$

Figure 3: Histograms of frequencies of occurrence of unique clusterings discovered by K-means (left column) and I_2 (right column) over 100 runs on a six-cluster dataset. Clusterings are ranked by decreasing frequency. Rank and frequency of the clustering giving the minimal value of the corresponding criterion for a particular value of K are marked at the top of each histogram.

Further examination of the similarity measured by Rand index [7, 16] between the best clusterings found by the two methods for a given K helped to determine the appropriate number of clusters for this dataset. Similarity between the best clusterings was about 0.7 for $K = 3$ and was close to the maximum possible value of one for $K = 6$. Even though the best clustering found by K-means for $K = 6$ was not ranked first, it was much more similar to the best clustering found by I_2 than was the case for $K = 3$. Together, Fig. 3 and pairwise similarity scores between clusterings found by K-means and I_2 suggest correctly that the appropriate number of clusters for this dataset is six. The histograms for $K > 6$ were smooth, did not satisfy the conditions of the heuristics proposed in this work and are omitted due to space constraints.

3.3 Heuristic Three

The third heuristic for determining the number of clusters is based on the following observation. K-means method takes into account only the cluster means when constructing discriminant surfaces. I_2, on the other hand, takes into account not only the cluster means, but also standard deviations of the clusters when constructing discriminant surfaces (see [11] for details). Sensitivity to cluster variance makes I_2 more reluctant towards generating clusters with small numbers of points and high variance in presence of neighboring points nearby.

Contingency tables presented in Table 1 demonstrate a situation where the two methods produced similar clusterings for $K = 2$ and very different clusterings for $K = 3$. In the latter case, K-means split off six points into a separate cluster, while maintaining a large 316-point cluster. I_2, however, produced two clusters of nearly equal sizes of 186 and 135 points. The third cluster under both methods was of moderate size of about 30 points. Such drastic difference in clusterings resulted in a sharp decrease in the Rand index from about one for $K = 2$ to approximately 0.6 for $K = 3$.

Table 1: Contingency tables for clusterings found by K-means and I_2 on Blackout dataset for $K \in \{2, 3\}$

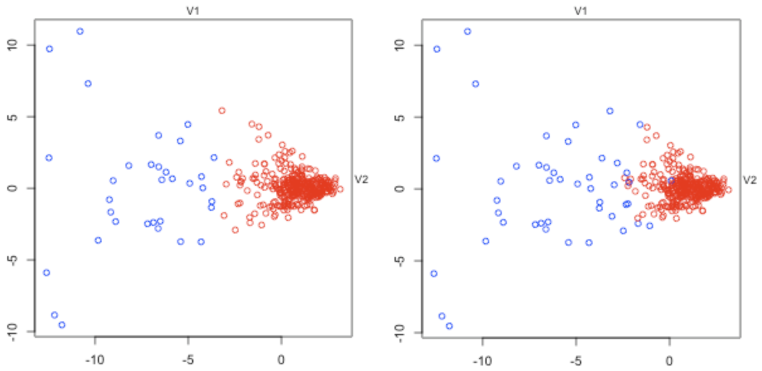
		I_2		
	Cluster	2	1	Size
K-means	1	308	13	321
	2	0	33	33
	Size	308	46	

		I_2			
	Cluster	1	3	2	Size
K-means	1	186	130	0	316
	2	0	0	6	6
	3	0	5	27	32
	Size	186	135	33	

When considered together, behavior of the two methods indicates that the data contains a dense core of points, which should not be split into separate clusters. The fact that the two methods produced highly different clusterings as a result of increasing the number of clusters from K to $K + 1$ suggests that the appropriate number of clusters for this dataset is no greater than K. Visualizations of clusterings of the dataset in the space of the first two principal components are shown in Fig. 4, and support the hypothesis that $K = 2$ is an appropriate number of clusters in this case.

4 Experimental Results

The heuristics described in Sect. 3 were first evaluated on a collection of nine synthetic datasets containing $2, 3, \ldots, 9$ and 10 clusters, respectively. The clusters were generated according to Gaussian distributions with different means and with one of the clusters having



(a) K-means

(b) I_2

Figure 4: Plot of twenty-dimensional Blackout event dataset in the space of the first two principal components. Clusterings for $K = 2$ are indicated by colors.

larger variance than the others. Using the developed heuristics, we were able to identify the correct number of clusters in datasets with moderate number ($K \leq 7$) of clusters.

We then applied the heuristics in the domain of cyber security on data collected [15] from Border Gateway Protocol (BGP) for Internet routing. The dataset contained information on four anomalous events: outbreak of the Slammer [14] worm, east coast power Blackout [1] and outbreak of the WItty [17] worm. The dataset consisted of four-dimensional time series, with each of the event data preceded by a pre-event time series reflecting the normal behavior of BGP. The features for characterizing BGP were selected based on prior works [6, 9] in this domain. Following [6], we used 3-second aggregate counts of the number of BGP announcements, BGP withdrawals, announced prefixes and withdrawn prefixes.

The four-dimensional vectors within each time series were standardized, feature-wise, by subtracting the mean value of a feature and dividing by its standard deviation. Standardization brought each of the four features of a time series to the same scale, thus making the data more suitable for clustering.

In order to capture time-local dynamics of the data, each time series was further transformed into a twenty-dimensional time series by combining windows of five consecutive four-dimensional vectors into the corresponding twenty-dimensional vectors. The resulting Slammer, Blackout and Witty datasets contained 922, 741 and 447 vectors, respectively. These twenty-dimensional vectors were used as input to K-means and I_2 clustering algorithms.

In all the experiments discussed in this section, we ran K-means and I_2 clustering $L = 100$ times, each time starting from a different randomly generated initial setting. Two sets of experiments were conducted on BGP data. Experiments presented in Sect. 4.1 addressed the question of differentiability of the events from normal BGP behavior by application of clustering methods to BGP data. Section 4.2 presents the results of further identification of internal structure within the event portions of BGP datasets using the developed heuristics.

Preliminary experimental results on scalability across domains of the proposed heuristics are reported in Sect. 4.3.

4.1 Detection of Anomalous Events in the Internet

In this subsection we present the results of application of the developed heuristics in order to determine if large scale anomalous events in the Internet could be detected using clustering methods. The experiments were performed on each of Slammer, Blackout and Witty datasets.

Based on the heuristic described in Sect. 3.1, two clusters were detected in each of the three datasets. The corresponding histograms for Slammer and Blackout datasets are shown in Fig. 5. Clustering results obtained by K-means on Slammer and Blackout data were superimposed for illustrative purposes on one of the four original time series features and are shown in Fig. 6. Clusterings obtained by I_2 were extremely similar to those obtained by K-means and are not shown due to space limitations. As can be seen in Fig. 6a, the two clusters discovered in Slammer dataset corresponded to non-event and event data.

The two surges comprising Cluster 2 and marked by cyan color in Fig. 6b happened between 4:10pm and 4:18pm EDT on August 14, 2003 and coincided with the beginning of the Northeast Blackout that was fully on from 4:10pm EDT[1] that day.

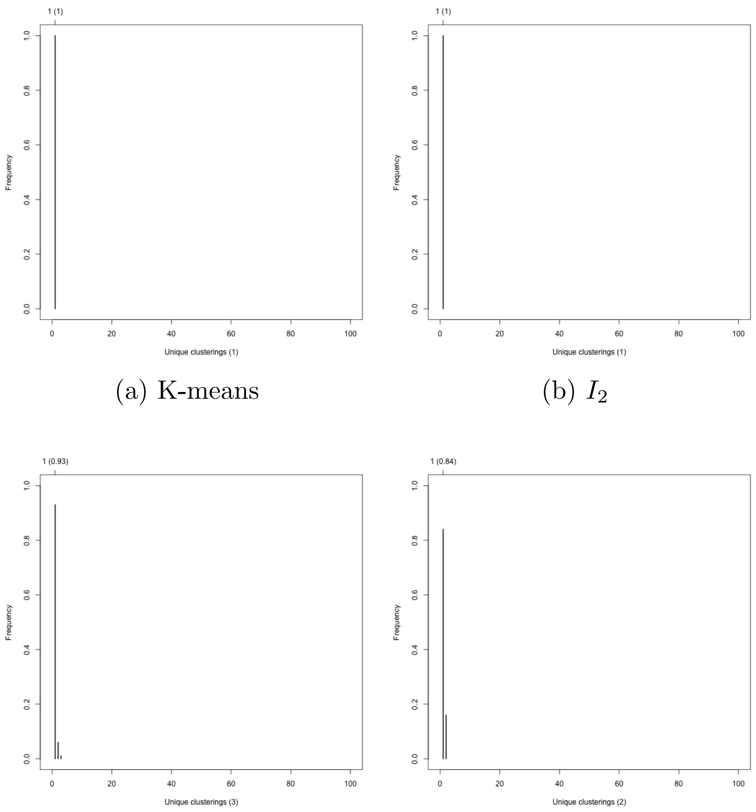
One of the two clusters discovered in Witty dataset comprised of very few (outlier) points: two in case of K-means and four in case of I_2 clustering. The rest of the points resided inside a dense core with no apparent clustering structure. We therefore concluded that Witty event was not distinguishable from the normal BGP behavior based on the first-order information used in the experiments. This result is consistent with the analysis performed by [17], who estimated that the total size of the population vulnerable to Witty worm was on the order of 12,000 computers, which is about one order of magnitude less than the number of machines infected by Slammer. Additionally, Witty carried a destructive payload that would eventually render an infected machine non-operational and thus unable to attempt to spread the worm. Hence, the disturbance caused by Witty worm may not have been large enough to be noticeable on the level of first-order BGP data.

Using higher-order relations within BGP data, [6] demonstrated promising results on detection of BGP events, including Witty, in a *supervised* machine learning setting. The incorporation of higher-order information into clustering (i.e., *unsupervised*) methods constitutes one of the directions for future research.

4.2 Identification of Internal Structure of BGP Events

Results of clustering Slammer event data satisfied the conditions of the heuristic described in Sect. 3.2. Two-cluster structure was thus discovered within Slammer event, and would have to be further examined by a domain expert.

[1] http://www.tipmagazine.com/tip/INPHFA/vol-9/iss-5/p8.html



(a) K-means

(b) I_2

(c) K-means

(d) I_2

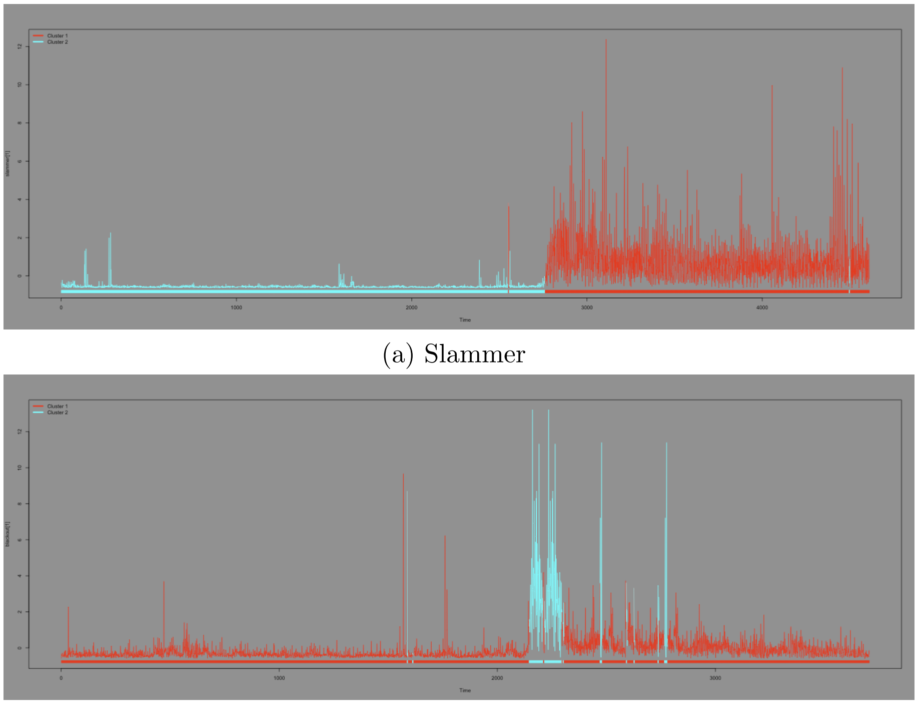
Figure 5: Histograms of frequencies of occurrence of unique clusterings discovered by K-means and I_2 over 100 runs on Slammer (top row) and Blackout (bottom row) datasets for $K = 2$

By applying the heuristic given in Sect. 3.3, a two-cluster structure was discovered within the Blackout event data as well. Points forming the two prominent surges coinciding with the beginning of the Northeast Blackout event in Blackout data were again placed into a separate cluster.

The clusterings discovered by K-means and I_2 within Slammer and Blackout events were very similar. For brevity, only the clusterings by K-means are shown in Fig. 7. The results of clustering Witty event data showed no presence of any apparent cluster structure within that dataset.

4.3 Scalability of the Heuristics

Preliminary experiments on four additional datasets (Wine, Ecoli, Iris and Wifi) were conducted in order to assess the scalability of the heuristics described in Sect. 3. Wine, Ecoli



(a) Slammer

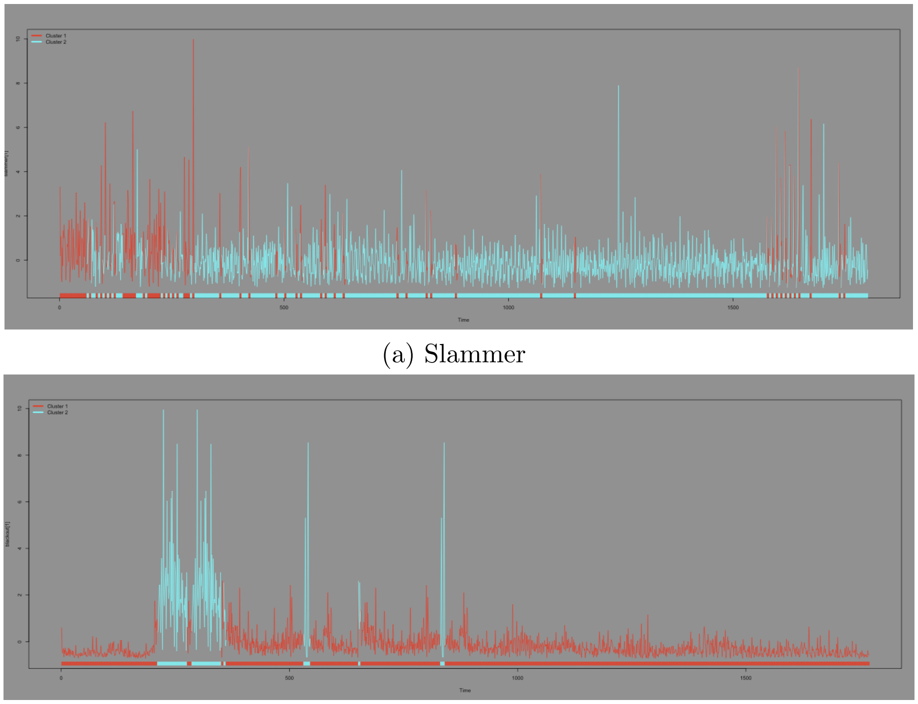
(b) Blackout

Figure 6: Clustering results obtained by K-means on two BGP datasets for $K = 2$. One of the four original time series features is shown. Colors indicate cluster assignment. The colored bars below the plots are projections of the cluster assignment onto the time axis.

and Iris datasets for supervised learning were obtained from the UCI repository [2] and were comprised of three, eight and three classes, respectively. By applying the heuristics, three clusters were detected in Wine and Ecoli data, while two clusters were detected within Iris dataset. Even though classification was not the primary goal of our work, we were able to evaluate classification accuracies (Table 2) of the discovered cluster structures. All points within a cluster were assigned the majority class label. Several classes in Ecoli and Iris data were mixed together in the space of the original features, which resulted in lower classification accuracies.

Table 2: Classification accuracies of the cluster structures detected in three datasets

Data	K-means	I_2
Wine	96.62	97.19
Ecoli	75.3	74.7
Iris	66.7	66.7



(a) Slammer

(b) Blackout

Figure 7: Clustering results obtained by K-means on event portions of two BGP datasets for $K = 2$. One of the four original time series features is shown. Colors indicate cluster assignment. The colored bars below the plots are projections of the cluster assignment onto the time axis.

Wifi (802.11) dataset was provided by the authors of [3], where it was used in the experiments on detection of spoofing attacks in wireless networks. We used four subsets of Wifi dataset with one, two, three and four clusters, respectively. In all four cases, the correct number of clusters was detected using the heuristics described in this work.

5 Conclusion

Three heuristics for determining an appropriate number of clusters in a dataset were proposed in this work. By applying the heuristics, we were able to correctly identify the number of clusters in synthetic data. We then applied the developed heuristics in the domain of cyber security on datasets from Border Gateway Protocol (BGP) and from a Wifi 802.11 wireless network. Experimental results on BGP data demonstrated that clustering methods are adequate for detection of large-scale anomalous events in the Internet. This is especially encouraging since, unlike supervised approaches, clustering does not require labeled data, obtaining which is very costly in the real world. We were also able to detect the correct number of clusters in four different subsets of Wifi dataset.

Additional preliminary experimental results on Wine, Ecoli and Iris datasets from the UCI machine learning repository indicated that the proposed heuristics are scalable across different application domains.

A situation where more than one heuristic could be applied to a dataset was never encountered during the experiments reported in this work. However, it is conceivable that such situation may arise. Therefore, one direction for future research lies in the development of methods for determining an appropriate number of clusters in cases where multiple heuristics could be applied to a dataset.

Experimental results suggested that heuristics given in Sects. 3.1 and 3.2 are adequate when the number K of clusters is moderate ($K \leq 7$). Investigation of possible extensions of these heuristics for larger values of K constitutes another direction for future work.

Recent studies [6, 5] demonstrated the strength of higher-order path analysis for supervised learning in domains of textual and of computer network security data. As a part of an ongoing research on textual analytics for homeland security applications, we intend to expand on the current work by incorporating higher-order information for clustering textual data.

6 Acknowledgments

This material is based upon work supported by, or in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under contract/grant number W911NF0710459. Co-author William M. Pottenger gratefully acknowledges the help of his Lord and Savior, Jesus Christ, in his life and work.

References

[1] How and why the blackout began in Ohio, August 2004. http://www.nerc.com/docs/docs/blackout/ch5.pdf.

[2] A. Asuncion and D. Newman. UCI machine learning repository, 2007.

[3] Y. Chen, W. Trappe, and R. Martin. Detecting and localizing wireless spoofing attacks. *Sensor, Mesh and Ad Hoc Communications and Networks, 2007. SECON '07. 4th Annual IEEE Communications Society Conference on*, pages 193–202, June 2007.

[4] E. Dimitriadou, S. Dolniar, and A. Weingessel. An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, 67(1):137–159, March 2002.

[5] M. C. Ganiz. *Higher-order path analysis for supervised machine learning.* PhD thesis, Lehigh University, Bethlehem, PA, USA, January 2008.

[6] M. C. Ganiz, S. Kanitkar, M. C. Chuah, and W. M. Pottenger. Detection of interdomain routing anomalies based on higher-order path analysis. In *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*, pages 874–879, Washington, DC, USA, 2006. IEEE Computer Society.

[7] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, December 1985.

[8] A. K. Jain and R. C. Dubes. *Algorithms for clustering data.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[9] 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.*, 35(5):55–66, 2005.

[10] S. P. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theory*, IT-28:129, March 1982.

[11] N. I. Lytkin, C. A. Kulikowski, and I. B. Muchnik. Variance-based criteria for clustering and their application to the analysis of management styles of mutual funds based on time series of daily returns. Technical Report 2008-01, DIMACS, Rutgers University, Piscataway, NJ, USA, February 2008.

[12] G. Milligan and M. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, June 1985.

[13] B. Mirkin. *Clustering For Data Mining: A Data Recovery Approach.* Chapman & Hall/CRC, 2005.

[14] D. Moore, V. Paxson, S. Savage, C. Shannon, S. Staniford, and N. Weaver. Inside the slammer worm. *Security & Privacy, IEEE*, 1(4):33–39, July-Aug. 2003.

[15] U. of Oregon Route Views Project. http://www.routeviews.org.

[16] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[17] C. Shannon and D. Moore. The spread of the witty worm. *Security & Privacy, IEEE*, 2(4):46–50, July-Aug. 2004.

[18] Y. Shim, J. Chung, and I.-C. Choi. A comparison study of cluster validity indices using a nonhierarchical clustering algorithm. *Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on*, 1:199–204, Nov. 2005.