Graph-Based Anomaly Detection Applied to Homeland Security Cargo Screening

William Eberle
Department of Computer Science
Tennessee Technological University
Cookeville, TN USA
weberle@tntech.edu

Lawrence Holder
School of Electrical Engineering & CS
Washington State University
Pullman, WA USA
holder@wsu.edu

Beverly Massengill
Department of Computer Science
Tennessee Technological University
Cookeville, TN USA
bamassengi21@students.tntech.edu

Abstract

Protecting our nation's ports is a critical challenge for *homeland security* and requires the research, development and deployment of new technologies that will allow for the efficient securing of shipments entering this country. Most approaches look only at statistical irregularities in the attributes of the cargo, and not at the relationships of this cargo to others. However, anomalies detected in these relationships could add to the suspicion of the cargo, and therefore improve the accuracy with which we detect suspicious cargo. This paper proposes an improvement in our ability to detect suspicious cargo bound for the U.S. through a *graph-based anomaly detection* approach. Using anonymized data received from the Department of Homeland Security, we demonstrate the effectiveness of our approach and its usefulness to a homeland security analyst who is tasked with uncovering illegal and potentially dangerous cargo shipments.

Introduction

In August of 2011, two Chinese nationals, on trial in New York City, were found guilty of illegally trafficking in counterfeit perfumes [Department of Justice 2011]. On September 12, 2011, Customs and Border Protection (CBP) officers at the port of Newark seized a shipment of 10,740 noncompliant High Intensity Discharge (HID) conversion kits – a potential safety threat to Americans, according to the CBP's Commercial Targeting and Analysis Center (CTAC) [Byrd 2011]. And these are just a few of the publicly reported stories dealing with the shipping of illegal and potentially harmful goods into the U.S. Protecting our nation's ports is a critical challenge for homeland security and requires the research, development and deployment of new technologies that will allow for the efficient securing of shipments entering this country. Somewhere between 11 and 15 million containers arrive via ships into U.S. ports every year [Homeland Security News Wire 2010]. Thousands of suspicious cargos, whether they are illegal or dangerous, are examined by port authorities every day. Due to the volume, strategic decisions must be made as to which cargo should be inspected, and which cargo will pass customs without incident. This is a daunting task that requires advanced analytical capabilities to maximize effectiveness and minimize false searches.

When analyzing shipping manifests for anomalous activity, current approaches might look for numerical or statistical deviations in the values associated with container size, quantity, or hazardous material codes. In addition, most anomaly detection methods use a supervised approach, requiring labeled data in advance (e.g., illegal versus legitimate) in order to train their system.

The overall target of the proposed work is the improvement of our ability to detect suspicious cargo bound for the United States. However, the techniques developed under this project are also applicable to mass transit and border security (both cargo and traveler screening). The proposed techniques can improve security in these venues by augmenting existing screening approaches with further evidence supplied by the analysis of structural relationships among the entities involved (e.g., affiliations between people, shippers, institutions, etc.). Our graph-based approach allows the integration of data from multiple sources (e.g., cargo manifest and shipper information) into one graph. Such graphs can be built for each instance of cargo, and these graphs can be analyzed to detect both normative patterns of behavior and deviations to these normative patterns. Current approaches typically look only at statistical irregularities in the attributes of the cargo, but not at the relationships of this cargo to others. Anomalies detected in these relationships can add to the suspicion of the cargo, and therefore improve the accuracy with which we detect suspicious cargo. Again, similar scenarios exist for the application of this technology to the screening of cargo and travelers in mass transit and border settings.

The following section discusses recent work in outlier or anomaly detection, particularly with respect to homeland security and specifically analyzing cargo shipments. This is followed by a discussion of our graph-based anomaly

detection approach and a discussion of actual cargo shipping data. We then present empirical results on the data, demonstrating the performance of our approach both in terms of accuracy as well as running times. While we evaluate our anomaly detection approach only on one cargo dataset in this paper, this dataset is already a collection of several types of data about the cargo, and the success of the approach will indicate its applicability to a graph constructed from multiple, diverse data sources. We then conclude the paper with some final observations and a brief discussion of our future work.

Related Work

Homeland security research, particularly for analysis of suspicious cargo, has brought about a diverse set of applications. In 2008, Cardoso et al. presented an anomaly detection approach called SMART (Standalone Multiple Anomaly Recognition Technique) to determine the presence of contraband in trucks and cargo containers [Cardoso 2008]. The advantage of their approach is that it is a non-intrusive inspection of images. Using spectral decomposition analysis techniques, they hypothesized that by differentiating the common background of the image, the anomalies would represent contraband concealed in the cargo. Also in 2008, Swaney et al. developed an approach called FORELL that uses an intelligent agent to learn rules for detection of anomalies in images created by high energy x-rays [Swaney 2008]. The advantage is that this approach automatically creates anomaly detection rules, and is supposedly able to identify objects it has never seen before. However, both of these approaches require images of the cargo, requiring someone to either scan the cargo or gather relevant photos.

Agovic et al. proposed an approach for detecting anomalous cargo based on sensor readings at truck weight stations [Agovic 2007]. Due to the amount of noise present in this type of sensor data, they propose manifold embedding methods for feature representation, with the purpose of removing anomalous points away from the normal regions of the data.

In Ling et al's work, they study the movement of barges loaded with hazardous material [Ling 2011]. In their study, they develop a prototype system called TRACC in order to identify potential security threats by predicting and detecting anomalies in collected real-time data.

There has also been quite a bit of work dealing with outlier detection in domains that have similar characteristics to the tracking and analysis of cargo shipments. In particular, there has been much research involving *spatial data mining*, which involved the discovery of nontrivial patterns in spatial datasets [Janeja and Adam 2008]. Work by [Sun and Chawla 2004][Lu et al. 2003][Kou et al. 2006] propose *neighborhood-based anomaly detection* approaches to discover spatial outliers using various different types of distance measures in order to determine the deviation of a spatial object. Outliers are detected in spatial neighborhoods based on the impact of spatial attributes such as location, area, and contour. However, in all of these approaches, the spatial neighborhood does not account for spatial autocorrelation and heterogeneity in combination. Another approach involving *scan statistics* tests if a process is purely random or if any unusual groupings in the form of a scan window can be detected, indicating a non-random cause [Janeja and Adam 2008]. The LS^3 approach proposes a linear semantic based scan statistic which accounts for variability in the data [Janeja and Atluri 2005]. However, this approach is limited to linear paths and linear windows. Perhaps the most widely used approach is called the spatial scan statistic where a spatial region is scanned with a circular window of varying size [Kulldorff 1997]. However, this approach does not identify irregularly shaped windows, and it does not take heterogeneity into consideration.

Graph-based anomaly detection provides an *unsupervised* approach for analyzing the *structure* and *relationships* in a graph-based representation of data. While most traditional approaches analyze the statistical properties associated with data points, a graph-based approach will analyze relationships through the structure of the data represented as nodes and links.

Graph-Based Anomaly Detection

The ability to mine data for nefarious behavior is difficult due to the *mimicry* of the perpetrator. If a person or entity is attempting to commit fraud or participate in some sort of illegal activity, they will attempt to convey their actions as close to legitimate actions as possible. For instance, the United Nations Office on Drugs and Crime states the first fundamental law of money laundering as "The more successful money-laundering apparatus is in imitating the patterns and behavior of legitimate transactions, the less the likelihood of it being exposed" [Hampton and Levi 1999]. Recently there has been an impetus towards analyzing relational data using graph theoretic methods [Holder and Cook 2007]. The advantage of graph-based anomaly detection is that the relationships between elements can be analyzed, as opposed to just the data values themselves, for structural oddities in what could be a complex, rich set of information.

Novel algorithms have been developed for analyzing graph structures for the purpose of uncovering all three types of graph-based anomalies: modifications, insertions and deletions [Eberle and Holder 2007]. The idea behind these approaches is to find anomalies in graph-based data

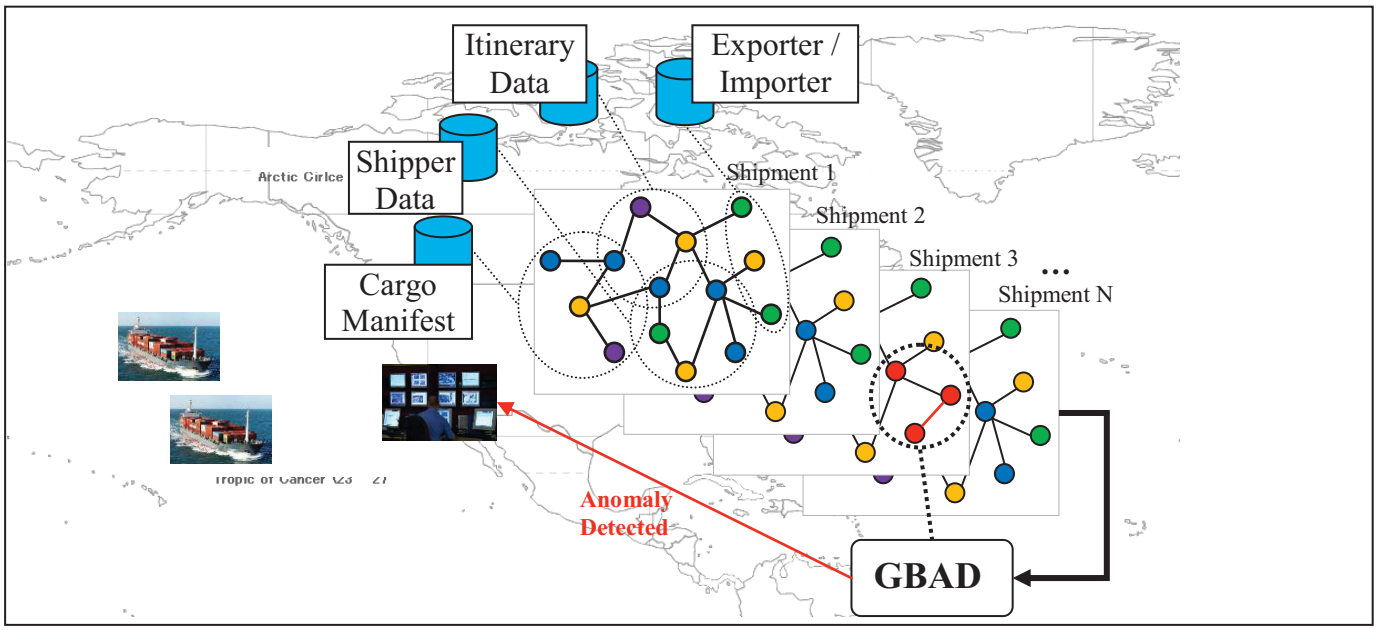

Figure 1. Detecting anomalies in shipping manifests using GBAD.

where the anomalous substructure in a graph is part of (or attached to or missing from) a non-anomalous substructure, or the *normative substructure*. This definition of an anomaly is unique in the arena of graph-based anomaly detection, as well as non-graph-based anomaly detection. Most anomaly detection methods use a supervised approach, which requires some sort of baseline of information from which comparisons or training can be performed. In general, if one has an idea what is normal behavior, deviations from that behavior could constitute an anomaly. However, the issue with those approaches is that one has to have the data in advance in order to train the system, and the data has to already be labeled (e.g., fraudulent versus legitimate). Work up until now has resulted in the development of three algorithms (*mdl*, *mps*, *prob*), which are implemented within a system called GBAD (Graph-Based Anomaly Detection). GBAD is an *unsupervised* approach, based upon the SUBDUE graph-based knowledge discovery system [Cook and Holder 1998]. Using a greedy beam search and Minimum Description Length (MDL) heuristic [Rissanen 1989], each of the three anomaly detection algorithms uses SUBDUE to discover the most prevalent substructure, or normative pattern, in an input graph. The MDL approach is used to determine the best substructure(s) as the one that minimizes $M(S,G) = DL(G|S) + DL(S)$, where G is the entire graph, S is the substructure, $DL(G|S)$ is the description length of G after compressing it using S, and $DL(S)$ is the description length of the substructure. Description length (DL) is a measure of the minimum number of bits necessary to represent the given information (in this case, a graph). Figure 1 depicts the scenario we envision for the use of GBAD to detect anomalies in cargo data. Each cargo is represented as a graph, which contains information about the cargo (e.g., contents, source, destination, ports visited, etc.) as well as relationships between this information and other relevant information (e.g., links between the shipper and other institutions). Past shipments can be represented in this way, and GBAD can find normative patterns in this historic data. Current shipments, also represented as a graph, can be searched for near misses (i.e., anomalies) to these normative patterns. The presence of these anomalies in a cargo's graph raises the level of suspicion in a cargo screening system.

Modified GBAD

One of the issues that we expected to deal with is the size of the cargo shipment information. A graph representation of just 500 shipments could result in tens-of-thousands of nodes and edges. While a previous implementation of GBAD (called GBAD-MDL) demonstrated successes at discovering interesting patterns [Eberle et al. 2011], it suffers from the subgraph isomorphism issue. Despite the integration of several novel heuristics, only minor performance improvements were demonstrated.

One graph-based knowledge discovery approach that has shown to be expedient without losing any accuracy can be found in the many *frequent subgraph miners*. Perhaps the most effective approaches have been the ones that convert graphs to a string in canonical form, and then perform a canonical-string based graph match. Existing approaches such as GASTON [Nijsson and Kok 2004], gSpan [Yan and Han 2002], GBI [Matsuda et al. 2002] and Grew

[Kuramochi and Karypis 2004] use canonical approaches to return frequent substructures in a database that are represented as a graph. In addition, each of these approaches has demonstrated significant improvements in processing time when applied to real-world data sets. Yet, while these approaches have been implemented in some commercial applications for graph-based knowledge discovery, there have been no attempts to incorporate the efficiency of these algorithms into an *anomaly detection* framework.

So, to verify the potential effectiveness of implementing anomaly detection algorithms into a frequent subgraph mining approach, we implemented the GBAD algorithms into the GASTON framework, and called this new approach GBAD-FSM. The main idea behind the GASTON algorithm is an *Apriori* approach whereby prior knowledge of frequent item-set properties is used to discover those substructures that are frequent. This well-known property provides a reduction in the search space, which can then be used to improve the performance for determining which substructures have an anomalous match. Initial experiments on synthetic data sets have demonstrated significant improvements in running times, but now we need to apply them to real-world cargo shipping data that exhibits different characteristics from the synthetic graphs we have used so far.

```
v 1 “Shipment”
v 2 “Hong Kong”
v 3 “Long Beach”
v 4 “Sacramento”
v 5 “Pacific Shippers”
v 6 “Bigmart”
v 7 “Pacific Imports”
v 8 “0”
v 9 “0”
v 10 “1”
v 11 “2007-10-10”
v 12 “N”
v 13 “N”
v 14 “118.41795”
v 15 “1”
u 1 2 “Port of Lading”
u 1 3 ”Port of Unlading”
u 1 4 “Port of Entry”
u 1 5 “Shipper”
u 1 6 “Consignee”
u 1 7 “Importer”
u 1 8 “Bill Rule 71”
u 1 9 “Bill Rule 72”
u 1 10 “Bill Rule 85”
u 1 11 “Date of Arrival”
u 1 12 “Narcotics Violation”
u 1 13 “Terror Proxy”
u 1 14 “ATS Score”
u 1 15 “Shipment ID”
```

Figure 2. Example of shipping transaction represented in graph input file.

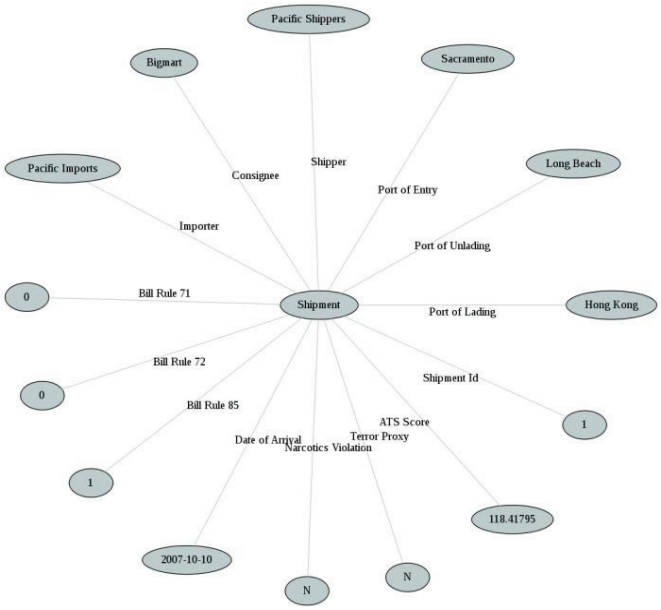

Figure 3. Graph representation of a shipment.

Empirical Results

The following experiments were performed using anonymized shipping data obtained from the DHS. The data were originally in an *SQL* script written to insert 550 shipment transactions into a database. To extract the data, we parsed the script, and placed the database headers as edge labels, and each database entry as a vertex connected to a central vertex, which we labeled “Shipment”. This process created an input graph file of 550 subgraphs, one for each row in the database. An example transaction from the resulting graph input can be seen in Figure 2, and a graph representation of the input can be seen in Figure 3, where each line represents either a vertex (v) or an undirected edge (u), and its corresponding vertex number(s) and labels. In our graph representation, each shipment is represented in its own, disconnected subgraph. The following experiments were performed on a Debian 5.0, 64-bit server with 8GB of RAM.

Performance Time

To evaluate GBAD run times, the input file which contained 550 shipment transactions was duplicated to produce multiple input files, each time doubling the size of the former. These files were then analyzed by the *mdl*, *mps*, and *prob* algorithms using the modified GBAD approach we are calling GBAD-FSM. The running time for each algorithm on each input size was recorded in seconds in Table 1.

Transactions	mdl	mps	prob
550	9.21	0.26	0.61
1100	10.08	0.62	1.53
2200	11.55	1.74	4.62
4400	14.93	5.44	15.4
8800	22.24	19.15	54.8
17600	40.87	70.13	210.62
35200	95.6	274.33	814.88
70400	262.1	1325.02	3251.69
140800	770	6191.95	13205.33
281600	2499.32	27676.96	54368.24

Table 1. Running times in seconds for GBAD-FSM using *mdl*, *mps*, and *prob* algorithms on graphs of increasing size.

Figure 4 shows visual representations of the algorithms' running times. As can be seen in the tables and figures, the running time of the *mdl* algorithm is fairly linear, while both the *mps* and *prob* algorithms appear to grow exponentially after about 140,800 shipments. In 2010, 7,579 oceangoing vessels made 62,747 calls at U.S. ports [AAPA 2010]. While exponential running times are generally not desirable, the expected number of shipments should never come close to these experimental numbers.

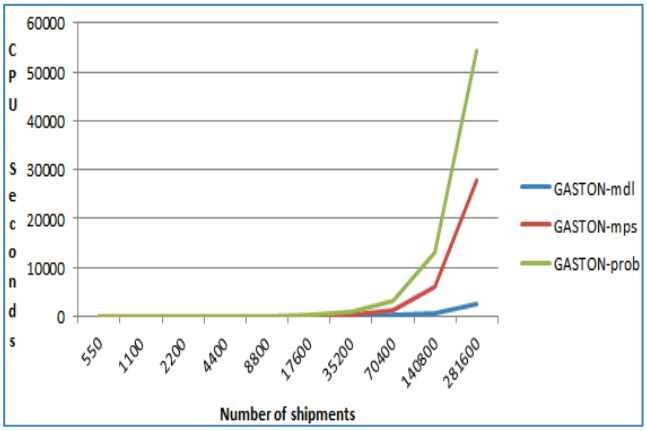

Figure 4. Running times for the three algorithms in GBAD-FSM.

Precision

To test GBAD-FSM's precision, an input file which contained 550 shipment transactions was duplicated to produce nine input files, each twice the size of the former. These files were then seeded with three different anomalies: an insertion, a deletion, and a label change. The modified input files were then checked by the *mdl*, *mps*, and *prob* algorithms using the GBAD-FSM approach. As expected, each algorithm produced different results. The *prob* algorithm discovered the insertions, *mps* discovered the deletions as well as the label changes, and *mdl* discovered just the label change.

TP/FP			
Transactions	mdl	mps	prob
550	1/0	2/102	1/0
1100	1/0	2/202	1/0
2200	1/0	2/402	1/0
4400	1/0	2/802	1/0
8800	1/0	2/1602	1/0
17600	1/0	2/3202	1/0
35200	1/0	2/6402	1/0
70400	1/0	2/12802	1/0
140800	1/0	2/25602	1/0

Table 2. The true positives / false positives ratio for each detection algorithm for GBAD-FSM.

False-Positive Rate			
Transactions	mdl	mps	prob
550	0	0.185	0
1100	0	0.184	0
2200	0	0.183	0
4400	0	0.182	0
8800	0	0.182	0
17600	0	0.182	0
35200	0	0.182	0
70400	0	0.182	0
140800	0	0.182	0

Table 3. The false positive rate (the number of false positives divided by the total number of transactions) for GBAD-FSM.

Table 2 shows the true positive to false positive ratio, and Table 3 shows the false positive rate (defined as the number of false positives divided by the total number of transactions) for each algorithm. For this experiment, false positives were defined as results which were not necessarily incorrect, but were unrelated to the inserted anomalies. Only the *mps* algorithm reported false positives, while *mdl* and *prob* always reported the anomaly without any false positives. In this case, the false positives are attributes with only two values that are fairly evenly split between positive and negative examples. Why the *mps* algorithm reports this non-anomalous deviation is a focus of future work.

Real-World Example

In a real-world example, seemingly legitimate boxes of laundry detergent were found to be counterfeit because of slight anomalies in the shipments' details [Lacitis 2011].

Specifically, the cargo had been weighed in kilograms while most shipments of a similar type had been weighed in pounds, and the weight value contained a comma where most shipments would have a decimal point. To determine whether GBAD-FSM could detect anomalies of this kind, the input file containing 550 shipment transactions (3,840 vertices and 3,584 edges) was modified so that all numerical attributes had decimal points, but no commas, and an attribute, "Unit of Measure" was always defined as pounds. Then one transaction was changed to reflect the comma and kilogram anomalies. The modified input was processed by all the detection algorithms, and the anomaly was discovered by the *mdl* algorithm in 1.79 seconds.

Conclusions and Future Work

We have demonstrated some of the capabilities of a graph-based approach to anomaly detection in the domain of homeland security cargo screening. Using a frequent subgraph mining approach to deal with the scalability needed in a real-world domain, we have shown that not only can we maintain detection accuracies, but we can also significantly speed-up the detection times – both valuable criteria in any tool used by a homeland security analyst. Future work will continue to focus on the *scalability* of this approach. While we have demonstrated some successes on large graphs, there are many real-world domains where the number of nodes and edges are on the order of billions or even trillions. One way to address this issue is to handle the data as a *stream*, thereby building the graph as data "comes in", searching for normative patterns and anomalies. In addition, we will continue to analyze the effectiveness of this approach using data sources that are integrated together. While this paper only deals with data from a single source, future work with agencies such as the DHS will involve the incorporation of data from various sources into a single graph input file.

Acknowledgements

This material is based upon work supported by the DHS under Contract No. HSHQDC-10-C-00212. Any opinions, findings and conclusions expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of Homeland Security.

References

American Association of Port Authorities, 2010, http://aapa-ports.org/Industry/content.cfm?ItemNumber=1022&navItemNumber=901.

Byrd, E. 2011. *CBP Targets Illegal Imports of Auto Headlamps*. CBP.gov, October 6, 2011.

Cook, D. and Holder, L. 1998. *Graph-based data mining*. IEEE Intelligent Systems 15(2):32-41.

Department of Justice. 2011. *Two Chinese Defendants Plead Guilty in Brooklyn, N.Y., to Trafficking in Counterfeit Perfume. New York.* August 5, 2011.

Eberle. W. and Holder, L. 2007. *Anomaly Detection in Data Represented as Graphs*. Intelligent Data Analysis, An International Journal, Volume 11(6).

Eberle, W., Holder, L. and Graves, J. 2011. *Insider Threat Detection Using a Graph-based Approach*. Journal of Applied Security Research, Volume 6, Issue 1, pp. 32-81, January 2011.

Eberle, W. and Holder, L. 2011. *Compression Versus Frequency for Mining Patterns and Anomalies in Graphs*. Conference on Knowledge Discovery and Data Mining (KDD) Mining and Learning with Graphs (MLD), August 2011.

Hampton, M. and Levi, M. 1999. *Fast spinning into oblivion? Recent developments in money-laundering policies and offshore finance centres*. Third World Quarterly, Volume 20, Number 3, pp. 645-656.

Holder, L. and Cook, D. 2007. *Mining Graph Data*. John Wiley and Sons.

Homeland Security News Wire. 2010. *Incentives for private industry, risk-based inspection for cargo containers*. February 22, 2010.

Janeja, V. and Adam, N. 2008. *Homeland Security and Spatial Data Mining*. In Proceedings of Encyclopedia of GIS. 2008, 434-440.

Janeja, V. and Atluri, V. 2005. *A linear semantic scan statistic technique for detecting anomalous windows*. ACM Symposium on Applied Computing.

Kou, Y., Lu. C., Chen, D. 2006. *Spatial weighted outlier detection*. In Proceedings of the Sixth SIAM International Conference on Data Mining, Bethesda, MD, USA 20-22, April, 2006.

Kulldorff, M. 1997. Comm. Statistics – Theory Meth. 26(6), 1481-1496.

Kuramochi, M. and Karypis, G. 2004. *Grew – A Scalable Frequent Subgraph Discovery Algorithm.* IEEE International Conference on Data Mining (ICDM '04).

Lacitis, E. 2011. *Port sleuths on the trail of counterfeit goods*. The Washington Post, January 15, 2011.

Lu, C., Chen, D., Kou, Y. 2003. *Detecting spatial outliers with multiple attributes*. In Fifth IEEE International Conference on Tools with Artificial Intelligence, p. 122.

Matsuda, T., Motoda, H., Yoshida, T. and Washio, T. 2002. *Knowledge Discovery from Structured Data by Beam-Wise Graph-Based Induction*. PRICAI 2002: Trends in Artificial Intelligence. Volume 2417. Pp. 123-141.

Nijssen, S. and Kok, J. 2004. *A Quickstart in Frequent Structure Mining Can Make a Difference*. International Conference on Knowledge Discovery and Data Mining, SIGKDD. pp. 647-652.

Rissanen, J. 1989. *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Company.

Sun, P. and Chawla, S. 2004. *On local spatial outliers*. In Fourth IEEE International Conference on Data Mining, pp. 209-216.

Yan, X. and Han, J. 2002. *gSpan: Graph-Based Substructure Pattern Mining*. Proceedings of International Conference on Data Mining, ICDM, pp. 51-58.