Discovering Suspicious Behavior Using Graph-Based Approach

Lenin Mookiah, William Eberle
Department of Computer Science,
Tennessee Technological University, Cookeville, Tennessee.
lmookiah42@students.tntech.edu and weberle@tntech.edu

Larry Holder
School of Electrical Engineering and Computer Science,
Washington State University, Pullman Washington
holder@wsu.edu

Abstract

The ability to discover illicit behavior in complex, heterogeneous data is a daunting problem. In the VAST 2014 competition, one of the challenges involves identifying for local law enforcement which employees are involved and where the police should be concentrating their efforts. This involves using diverse data such as GPS locations and credit card transactions. Traditional anomaly detection methods have difficulties handling this type of heterogeneous data, where there are movements and relationships between entities. One approach to helping with this problem is a graph-based approach. The primary advantage of a graph-based approach lies in its potential to handle rich contextual information. In this paper, we apply a graph-based anomaly detection approach for discovering suspicious employees and their actions as a tool for aiding a potential criminal investigation.

Introduction

For the Visual Analytics Science and Technology (VAST) 2014 challenge, contestants are asked to aide law enforcement in the fictional settings of Kronos and Tethys (VAST, 2014). They are investigating the mysterious disappearance of employees from a fictional company called GAStech. Employees of GAStech have company cars for daily usage of both personal and business affairs. Those who do not have company cars have the ability to check out company trucks for business use, but these trucks cannot be used for personal business. Employees prefer company cars because they are generally much higher quality than the cars they would be able to afford otherwise. However, GAStech does not trust their employees. Hence GAStech has installed geospatial tracking software in the vehicles without the employees' knowledge. The vehicles are tracked periodically as long as they are moving. This vehicle tracking data is available to law enforcement to support their investigation. Data is only available for the two weeks prior to their disappearance. Unfortunately, data is not available for the day the GAStech employees went missing. Additionally, law enforcement has personal and business credit and debit card transactions data for the local GAStech employees for the two weeks preceding their disappearance. Many of the employees also use loyalty cards for discount purchases. In addition, the data has limitations as a result of missing, conflicting, and imperfect information that would complicate recommendations for further investigation.

The VAST 2014 competition poses 3 separate mini-challenges as well as one grand-challenge. For the purposes of this work, we focused our efforts on the second mini-challenge, which focuses on movements and transactions amongst the employees. Mini-challenge 1 is primarily a natural language processing task, and mini-challenge 3 is dealing with streaming textual and audio blogs – neither of which is conducive to applying a graph-based approach. With mini-challenge 2, the task is to identify potentially malicious patterns to assist law enforcement. Our approach to handling the problem is to analyze the structure of the transactions and movements of individuals represented as a graph. In this paper, we apply a graph-based anomaly detection approach towards the discovery of suspicious employees and geographic locations.

Related Work

There are quite a few directions researchers have taken on finding unusual and malicious activities. These directions include data visualization, predicting suspicious events such as terrorism and crimes, studying factors behind unusual patterns, and insider threat.

Some research work has studied near future prediction of interested events in specific locations. Liao et al. (2010) proposes a novel Bayesian based prediction model. The authors use the model to predict the accurate location of the next crime scene in a serial crime. Results are effective with predicting three out of four crime locations (i.e., accuracy of 75%).

Gerber (2014) employs the social network Twitter to predict crime. The author collected GPS-tagged tweets (n = 60,876) mentioning crime events between January 1, 2013 and March 31, 2013. The author uses Latent Dirichlet Allocation (LDA) for learning topics and related terms from tweets and eschews deep semantic analysis in favor of shallower analysis via topic modeling. Out of 25 crime

types, 19 show improvement in the Area Under Curve (AUC) method after incorporating topics from tweets along with density-based estimation.

Michalak and Korczak (2011) apply graph mining techniques for the detection of suspicious banking transactions. The proposed method uses fuzzy numbers to represent transaction amounts involved in money laundering event. Evaluating the proposed method on its false-positive rate provides impressive results of not classifying any legal transactions as illegal.

Hot spots are the most popular visualization technique. "Hot spots" are defined as areas which have higher concentrations of crime events. In the prediction task, a regression model is the most widely studied. Leong and Chan (2013) have studied the trend of using web-based crime mapping from the 100 highest GDP cities of the world. The authors conclude that a web-based mapping showing crime visualization is the most common tool. Similarly Shingleton (2012) employs regression models to predict crime trends in Salinas, California. The author uses three methods: Ordinary Least Squares, Poisson Regression, and Violence Prediction using Negative Binomial Regression. He concludes that the ordinary least squares approach is adequate enough to predict three crime types as it is comparable to the regression approaches.

Young et al. (2013) study insider threats in computer activities of 5500 employees using domain knowledge. The author studies the characteristics of features that help to better find anomalies using multiple methods. These methods are evaluated using methodologies that include area under the curve, lift curve, and average lift.

Graph Based Approach

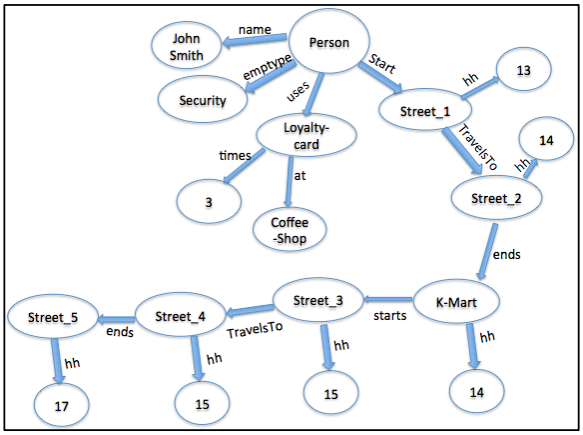

Figure 1: Graph Topology

The core idea behind the approach used in this work is to discover anomalies in data represented as a graph, where the anomalous substructure in a graph is deviation from a normative sub-structure.

Definition: *A change of X % is the cost of transforming subgraph S into an isomorphism of subgraph S'.*

Definition: *A graph substructure S is anomalous if it is not isomorphic to the graphs normative substructure S, but is isomorphic to S' within X %.*

X signifies the percentage of vertices and edges that would need to be changed in order for *S* to be isomorphic to *S'*.

Graph-based Anomaly Detection (GBAD) is an unsupervised approach, which is based on the SUBDUE graph-based knowledge discovery method (Cook and Holder 1994). There are three general categories of structural anomalies: insertions, modifications and deletions. Insertions constitute the presence of an unexpected vertex or edge. Modifications consist of an unexpected label on a vertex or edge. Deletions constitute the unexpected absence of a vertex or edge. Each of the above situations is handled by three separate algorithms: GBAD-MDL, GBAD-P and GBAD-MPS.

Primarily using a greedy beam search and Minimum Descriptive Length (MDL) heuristics (Rissanen 1989), all three algorithms use SUBDUE to find the best substructure or normative pattern from an input graph. With the GBAD approach, the goal is to find the best substructure that would minimize the following objective function:

$$M(S,G) = DL(G \mid S) + DL(S)$$

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.

The GBAD-P algorithm uses the MDL evaluation technique to discover the best substructure in a graph, but instead of examining all instances for similarity, this approach examines all extensions to the normative substructure (pattern), looking for extensions with the lowest probability. The GBAD-MPS algorithm also uses the MDL approach to discover the best substructure in a graph. Subsequently it examines all of the instances of parent (or ancestral) substructures that are missing various edges and vertices (i.e., *deletions*). The GBAD-MDL approach uses the MDL heuristic to discover the best substructure in a graph (i.e., the substructure that compromises the graph the most), and then subsequently examines all of the instances of the substructure that "look similar" to that pattern. It is this ability to examine vertices and edges that could represent movements and transactions, where unusual deviations from the norm could constitute potential illegal activities, make the application of the GBAD approach to this challenge interesting.

For more information regarding the GBAD algorithms, the reader should refer to (Eberle and Holder, 2007).

Data Set

The VAST data set consists of the activities of 44 employees. 9 of these employees do not have GPS car recordings available. Employee activities include GPS car recordings, credit card transactions, and loyalty card transactions for a period of two weeks. Table 1 provides statistics regarding the data set. Table 2 shows a sample of the employee profiles. Table 3 shows a sample of GPS movement for various cars. Table 4 shows a sample of the type of credit card information that is available.

In order to represent the data as a graph, we chose the graph topology shown in Figure 1 for each employee. For building the graph, we map the car GPS data at a given time for each employee to its corresponding street GPS co-ordinates, allowing us to identify the corresponding street name. Additionally we use credit card time stamp. Our intuition behind choosing this particular graph topology is that the suspicious employees at certain times do unusual (or anomalous) activities, and this representation will allow us to better understand both normal and anomalous movements. In addition, if we treat each employee both globally (i.e., as a single graph of all employees), individually (i.e., where each graph represents all of the actions and movements of a specific employee), and temporally (i.e., where each graph represents all movements and actions for a specific day), we can potentially uncover anomalies from different perspectives.

Number of Employee	44
Number of Car GPS	35
Total number of GPS records	685,169
Total Number of credit card transactions	1,491
Total Number of loyalty card transactions	1,393

Table 1: Data Statistics

Last Name	First Name	Car ID	Current Employment Type	Current Employment Title
Alcazar	Lucas	1	Information Technology	IT Helpdesk
Azada	Lars	2	Engineering	Engineer
Balas	Felix	3	Engineering	Engineer
Barranco	Ingrid	4	Executive	SVP/CFO

Table 2: Sample Employees' Profile

Time Stamp	Car ID	Latitude	Longitude
1/6/14 6:28	35	36.0762253	24.87468932
1/6/14 6:28	35	36.07622006	24.87459598
1/6/14 6:28	35	36.07621062	24.87444293

Table 3. Sample Car GPS data

Data Preparation

The provided database had some missing information, which can be attributed to the deletion of old records or changes in the site structure over time. Each graph input file contains the complete activity of an employee, divided into subgraph instances representing each day. Hence, each GBAD input file of an employee will have one instance for each of the 14 days. The average size of each graph input file is 1,703 vertices and 1,407 edges.

Time Stamp	Location	Price	First Name	Last Name
1/6/14 7:28	Brew've Been Served	11.34	Edvard	Vann
1/6/14 7:34	Hallowed Grounds	52.22	Hideki	Cocinaro
1/6/14 7:35	Brew've Been Served	8.33	Stenig	Fusil
1/6/14 7:36	Hallowed Grounds	16.72	Birgitta	Frente

Table 4. Sample Credit Card Transaction

Experiments

For analyzing this data, we chose to use a graph-based anomaly detection approach called GBAD. GBAD uses a definition of anomalousness based upon the theory that a person or entity that is attempting to commit an unusual, or illegal, action would do so by attempting to imitate known behaviors – thus concealing their true intentions. Based on this definition, an anomaly would not be random. As described earlier, GBAD uses three different algorithms for discovering anomalous graph substructures. Initial analysis using the graph input files leads us to focus our discoveries using the GBAD-P algorithm (i.e., anomalous insertions into the graph), as the GBAD-MDL and GBAD-MPS algorithms do not successfully discover any anomalous substructures on this data.

Using graph input files comprised of the graph topology shown in Figure 1, GBAD reports the normative (best) substructure and any anomalous substructures for each employee.

Ground Truth

Our goal is to identify suspicious patterns in the data. Additionally, we need to identify people involved whom can be further investigated by a law enforcement agency. Figure 2 shows the map with locations of various important places such as the office, safe houses and stores.

After the VAST competition, we were provided with the following ground truth in the data (i.e., the known anomalies of interest):

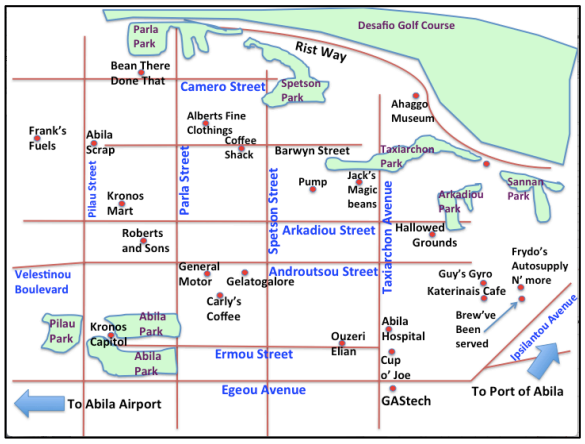

Figure 2: Map showing important locations such as office, safe houses and stores.

1. Inga Ferro (Car ID: 13), Loreto Bodrogi (Car ID: 15), Hennie Osvaldo (Car ID: 21), Minke Mies (Car ID: 24) do several practice drives from the office to various safe houses. The practice runs to the safe houses take place during the day with three of the safe houses in the eastern part of the city and two in the western part. These events happen subsequently after the following event (i.e., event 2).

2. Loreto Bodrogi (Car ID: 15), Isia Vann (Car ID: 16), Hennie Osvaldo (Car ID: 21), Minke Mies (Car ID: 24) carry out surveillance on the homes of GASTech executives (i.e., executive houses) in 3 hour shifts on January 6, 7, 9, 10, 11, 13, and 14 in the middle of the night when nobody else is driving.

3. Minke Mies steals the credit card from Lucas Alcazar (a help desk worker). Lucas still has his transactions as planned up until close-of-business on January 13th. Then he makes no card purchases till January 16, when his replacement credit card arrives. He still has loyalty data but no credit card data. On January 13th, Minke makes a couple of small test purchases online, followed by buying some gas. These take place while the actual owner of the card, Alcazar, is demonstrably elsewhere. Then Mies makes a large amount purchase ($10,000) at Frydo's autosupply n' more. No other purchases are subsequently made on the Lucas Alcazar's card.

Additionally, the following are anomalous but benign patterns provided by the VAST competition committee:

4. Employees Isande Borrasca (Car ID: 7) and Brand Tempesta (Car ID: 33) are having an affair. Multiple times they go to a hotel over lunch, usually leaving and returning at times offset from each other. Happens on days: January 8, 10, 14, and 17.

5. Employee Bertrand Ovan (Car ID: 29) cruising around town on January 11.

6. Kanon Herrero (Car ID: 22 ; Badging office) and Elsa Orilla (Drill Tech) (Car ID: 28) are dating. They go to lunch together every day. On January 18, they go to several locations together. Kanon usually pays so Elsa will not have corresponding card data for those occasions. Elsa drives on January 9, 14, and 17, while Kanon drives on January 6, 7, 8, 10, 13, 15, and 16.

7. Executives have a practice golf session on January 12th then golf again with Sanjorge on January 19th.

8. Employee Lucas Alcazar (Car ID: 1) works after hours on January 6, 7, 8, 15, 16, and 17.

9. A large party for all the engineers and IT is held on January 10. They all go to one house for the party around the same time.

10. IT group manager Linnea Bergen takes the IT group out for lunch on January 17 so she has a high cost lunch, with high loyalty points. The participants carpooled in a non-GPS car so there are no GPS records.

11. Axel Cazas (Car ID: 9) has a spotty GPS with frequent gaps in his data.

GBAD

Our goal is to identify suspicious patterns in the data set using GBAD. Additionally, we need to identify people involved who can be further investigated by a law enforcement agency. One of the more interesting things we notice from the output of GBAD occurs around the path of "Rist Way" (near Chostus Hotel) and around "Spetson Park". In particular, the suspect employees spend time passing through "niovis st" and "exadakitiou way" in "Rist Way" and at some streets around "Spetson Park". Streets involved are "niovis st", "exadakitiou way", "n estos st", "n utmana st", "n ketallinias st", "n ithakis st", "n oddisseos st".

In general, the following is a summary of our observations based upon the discovered suspicious events:

- Activity at unusual times of the day.
- Involves streets far away from the employees' main office GAStech.

Potential suspects are Cazar Gustav, Calzas Axel, Balas Felix, Vann Isia (from Ground Truth 2), Osvaldo Hennie (from Ground Truth 1 and 2), Onda Marin, Dedos Lidelse, Vann Edvard, Tempestad Brand,and Mies Minke (from Ground Truth 1, 2 and 3). Patterns happen between the evening of January 10 and January 11, and most of the suspects are from the department of engineering and security. The GBAD-P algorithm uses the hypothesis of capturing inserted edges and/or vertices (e.g., unexpected

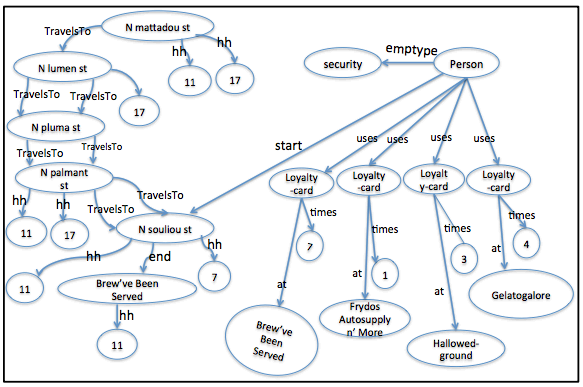

Figure 3: Normative Pattern of Osvaldo Hennie

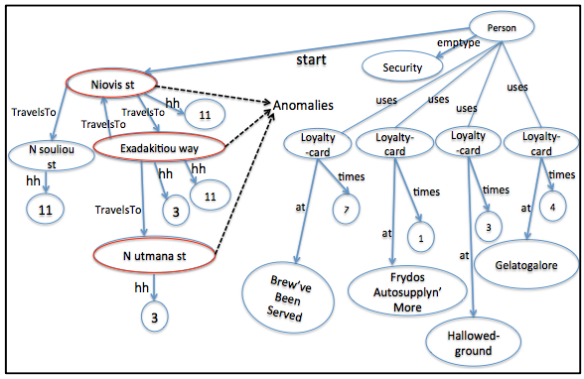

Figure 4: Anomalies for suspect Osvaldo Hennie

actions by an individual) that are probabilistically less likely to occur.

Using GBAD, we found suspicious events occurring between the evenings of January 10 and January 11. The following are some examples of these reported suspicious events:

Event 1 - Osvaldo Hennie on January 11 in the afternoon spent around 6+ hours at "n utmana st 3600 3698" (near "Spetson Park") passing via "niovis st" and "exadakitiou way" (confirmed in Ground Truth 1 and 2).

Event 2 - Vann Isia on January 10 late at night passed through "exadakitiou way", and spent 3 hours that night between "n utmana st 3700 3798" (confirmed in Ground Truth 2).

Event 3 - Tempestad Brand on January 10 passed through "exadakitiou way" and "niovis st 2700 2798" late at night and spent 4 hours at "n ketallinias st 4600 4650" (near "Spetson Park"). Around midnight Tempestad Brand spent time between "niovis st" and "exadakitiou way" (Ground Truth 9).

Event 4 - Vann Edvard passed through "niovis st" and "exadakitiou way". On January 10 Vann Edvard spent 4 hours between "exadakitiou way" and "n estos st 3600 3698" (near Spetson Park).

Event 5 - Onda Marin passed through "niovis st" and "exadakitiou way" on January 10 around midnight and spent approximately 4 hours between "exadakitiou way" and "n estos st 3600 3698" (close to Spetson Park – Ground Truth 9).

Event 6 - Mies Minke passed through "niovis st" and "exadakitiou way" at night, and spent approximately 3 hours between "n ithakis st 3700 3848" and "n oddisseos st 3600 3698" (confirmed in Ground Truth 1 and 2).

Event 7 - Balas Felix spent 5 hours around midnight at "n ketallinias st 4600 4650" (near "Spetson Park" – Ground Truth 9).

Event 8 - Calzas Axel on January 10 around midnight spent 4 hours at "n ketallinias st 4600 4650" (near Spetson Park – Ground Truth 9).

Event 9 - Dedos Lidelse on day January 10 spent 4 hours at night at "n ketallinias st" and passed through "niovis st" (Ground Truth 9).

Taking employee Osvaldo Hennie as an example (Event 1), Figure 3 and Figure 4 show the normative and unusual patterns associated with his actions. This pattern is indicative of at least 8 employees who are moving around locations of "Spetson Park" and "Chostus Hotel" which are away from their office (or) their regular eating-places such as Guy's Gyros.

For a considerable number of employees, GPS recordings were either incomplete or missing for given days. These missing or imperfect data could pose as challenges on finding anomalies. For example employee Herrero Kanon was missing data for January 06 with no GPS locations recorded. For January 07 there are fewer GPS recordings than usual; only morning and evening data were recorded. One would assume that the employee should have been seen to at least travel from home to work. However, the employee was not found with any suspicious patterns between January 10 and January 11. Similarly, employee Osvaldo Hennie is missing data for the day January 12. Since the most important suspicious event we look for is between January 10 and January 11, we still conclude the employee as suspicious.

Analysis

After the competition was complete, we received a response from the VAST committee letting us know that we had successfully identified the main actors in this crime scenario. However, while we were able to successfully narrow-down our main suspect to Mies Minke (Event 6), we did discover some weaknesses in our approach.

First, in our analysis with the output of GBAD, we assumed the strict definition of a suspect as the one who performs activities that deviate from the normative pattern, which resulted in a few false positives. For example, Inga Ferro made more frequent visits to stores such as FX

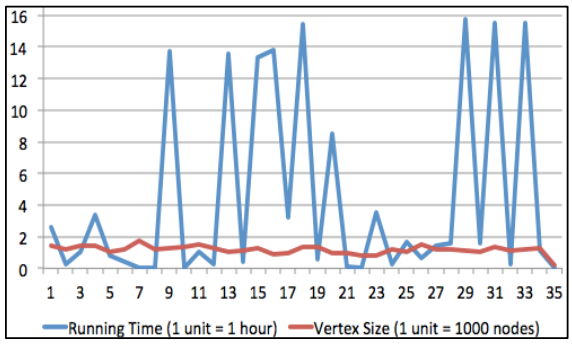

Figure 5: Performance of GBAD-P

compared to other suspects. In other words, Inga Ferros normative pattern has the substructure of anomalous patterns found associated with other suspects. Second, our approach is able to effectively answer questions such as "who", "when", "where" aspects of the challenge, but could not explain the "why" aspect. The GBAD approach only calculates a score of anomalousness for a particular substructure, with no background to explain. Third, we did discover a high false positive rate of around 66 %. This is primarily due to the benign party event mentioned earlier, that while an anomalous event, would not have been of interest to the authorities. However, if the authorities know about the party, and as such do not want any events (in this case, events 3, 5, 7, 8, and 9) surrounding that party to be factored into their investigation, removal of these related events reduces our false positive rate to 25 %. Fourth, the main suspect Mies Minke performed a transaction of $10,000. Again, GBAD is designed to discover interesting substructures – not unusual statistical anomalies in numerical values. In the future, modifying GBAD to handle *attributed* graphs would allow us to better understand the *context* of suspicious patterns.

Conclusions and Future Work

In this work, we applied a graph-based approach to analyzing heterogeneous data for the discovery of relevant patterns. Our approach was able to effectively find unusual patterns of interest that could lead to further investigation by law enforcement. In the future, we will investigate extending the graph-based anomaly detection approach to include a cumulative score calculation. This will allow us to better handle the statistical nature of some of the data (like credit card transactions) in addition to the existing structural capabilities of a graph-based approach. Along with calculating a cumulative score, a ranking function could be explored, whereby we can better distinguish different reported anomalies. In addition, we realized that the overall context or plot could not be understood with our graph-based approach. Supplementing our approach with some interactive visualization techniques, along with cumulative metrics, could provide better insights to an analyst. We are also currently working on the scalability of graph-based approaches to pattern discovery and anomaly detection. Figure 5 shows the running times of the GBAD-P algorithm with x-axis representing Car ID of employees. Unfortunately, the average running time of GBAD-P in our experiment is 6.13 hours. One possible way to reduce the running time in this scenario could be to reduce the graph size in a pre-processing stage using some ranking or heuristic algorithm based upon knowledge of the domain.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Nos. 1318913 and 1318957.

References

Cook, D. J. and Holder, L. B. 1994. Substructure discovery using minimum description length and background knowledge. *Journal of Artificial Intelligence Researc*h, 1:231–255.

Eberle, W. and Holder, L. B. 2007. Anomaly detection in data represented as graphs. *Intelligent Data Analysis*, 11(6):663–689.

Gerber, M. S. 2014. Predicting crime using twitter and kernel density estimation. *Decision Support Systems* 61:115–125.

Leong, K. and Chan, S. C. 2013. A content analysis of web-based crime mapping in the world's top 100 highest GDP cities. *Crime Prevention & Community Safety* 15(1):1–22.

Liao, R., Wang, X., Li, L., and Qin, Z. 2010. A novel serial crime prediction model based on bayesian learning theory. *In Machine Learning and Cybernetics (ICMLC), 2010 International Conference* on, volume 4, 1757–1762. IEEE.

Michalak, K. and Korczak, J. 2011. Graph mining approach to suspicious transaction detection. *2011 Federated Conference on Computer Science and Information Systems* (FedCSIS) 69–75.

Shingleton, J. S. 2012. Crime trend prediction using regression models for Salinas, California. *Technical report, DTIC Document*.

http://vacommunity.org/VAST+Challenge+2014

Young, T. W., Goldberg, G. H., Memory, A., Sartain, F. J., and Senator, E. T. 2013. Use of Domain Knowledge to Detect Insider Threats in Computer Activities, *IEEE Security and Privacy Workshops*, pp. 60-67.