Detecting Suspicious Behavior Using a Graph-Based Approach

Lenin Mookiah*, William Eberle*, Lawrence Holder
Tennessee Tech University* and Washington State University

Abstract

The ability to discover illicit behaviour 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 they should be concentrating their efforts. One approach to handling this problem is a graph-based approach. In this paper, we present a graph-based anomaly detection approach for discovering suspicious employees and geographic locations.

Keywords: graph-based anomaly detection, knowledge discovery

Index Terms: H.2.8 [Database Management]: Database Applications—Data Mining

1 Introduction

For the VAST 2014 challenge, contestants are asked to aide law enforcement, from the fictional settings of Kronos and Tethys, in assessing the disappearance of employees from a fictional company called GAStech. In the case of Challenge 2, we were tasked with the following: identifying normal employee patterns of behaviour; identifying unusual patterns or events in the data; and reporting any inconsistencies in the data. One approach to handling the problem is to analyze the structure of the transactions and movements of individuals as a graph. The ability to discover illicit behaviour in complex, heterogeneous data is a daunting problem. In this paper, we present a graph-based anomaly detection approach for discovering suspicious employees and geographic locations.

2 Graph-Based Anomaly Detection

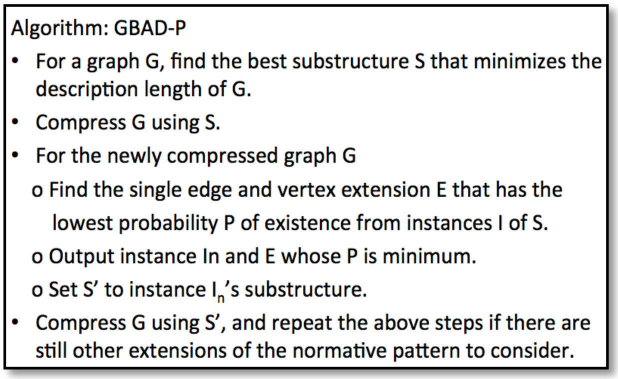
For analyzing this data, we chose to use a graph-based anomaly detection approach called GBAD. GBAD uses a definition of anomalousness based upon the theory that a person or entity that is attempting to commit an unusual, or illegal, action would do so by attempting to imitate known behaviours – thus concealing their true intentions. Based on this definition, an anomaly would not be random. GBAD uses three different algorithms for discovering anomalous graph substructures.

2.1 GBAD-MDL and GBAD-MPS

The GBAD-MDL algorithm uses the minimum description length principle (MDL) [1] [2] to determine the normative substructure, and from that substructure, find other substructures that while structurally similar (i.e., have the same number of vertices and edges), have some label modifications that are within an acceptable level of change. The GBAD-MPS algorithm also determines the best substructure as the one that minimizes the description length of a graph, but instead of looking at label changes, it looks for edges and vertices that are missing. We discovered that the GBAD-MDL and GBAD-MPS were not very helpful in finding anomalies for this particular VAST challenge.

2.2 GBAD-P

The GBAD-P algorithm also 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 difference between the this approach and the GBAD-MDL approach is that it is looking at instances of substructures with the same characteristics (i.e., size, degree, etc.), whereas GBAD-P is examining the probability of extensions to the normative pattern to determine if there is an instance that when extended beyond its normative structure is traversing edges and vertices that are probabilistically less likely than other extended instances. Figure 1 shows the GBAD-P algorithm.



```
Algorithm: GBAD-P
• For a graph G, find the best substructure S that minimizes the
  description length of G.
• Compress G using S.
• For the newly compressed graph G
    o Find the single edge and vertex extension E that has the
      lowest probability P of existence from instances I of S.
    o Output instance In and E whose P is minimum.
    o Set S' to instance $I_n$'s substructure.
• Compress G using S', and repeat the above steps if there are
  still other extensions of the normative pattern to consider.
```

Figure 1: GBAD-P.

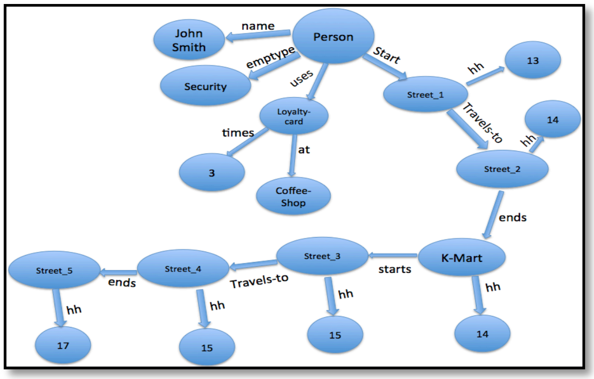


Figure 2: Graph-Topology.

3 Discussion

The following are what we discovered using GBAD.

*lmookiah42@students.tntech.edu; weberle@tntech.edu holder@wsu.edu

IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE

3.1 Pattern

Using the graph topology shown in Figure 2 as input to GBAD using the GBAD-P algorithm, we were able to discover the normative (best) substructure and any anomalous substructures for each employee. One of the more interesting things we notice 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".

The following is a summary of our observations based upon the discovered suspicious events:

- Activity at unusual time of the day.
- Involves streets far away from work location. Employee activity in office at times not his/her regular work-time.

3.2 Suspects and Events

Potential suspects are Cazar Gustav, Calzas Axel, Balas Felix, Vann Isia, Osvaldo Hennie, Onda Marin, Dedos Lidelse, Vann Edvard, Tempestad Brand, and Mies Minke. Patterns happen between the evening of 01/10 and 01/11, and most of the suspects are from the department of engineering and security.

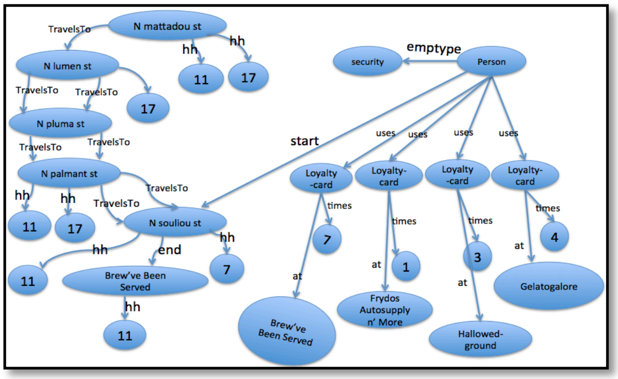


Figure 3: Normative Pattern for employee "Osvaldo Hennie".

Suspicious events were observed to occur between the evenings of 01/10 and 01/11. We present our findings of 10 suspicious events:

Event 1 - Osvaldo Hennie on 01/11 in the afternoon spent around 6+ hours at "n utmana st 3600 3698" (near "Spetson Park") passing via "niovis st" and "exadakitiou way".

Event 2 - Vann Isia on 01/10 late at night passed through "exadakitiou way", and spent 3 hours that night between "n utmana st 3700 3798".

Event 3 - Tempestad Brand on 01/10 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".

Event 4 - Vann Edvard passed through "niovis st" and "exadakitiou way". On 01/10 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 01/10 around midnight and spent approximately 4 hours between "exadakitiou way" and "n estos st 3600 3698" (close to Spetson Park).

Event 6 - Cazar Gustav spent around 5 hours at "n ketallinias st 4600 4650" (near Spetson Park).

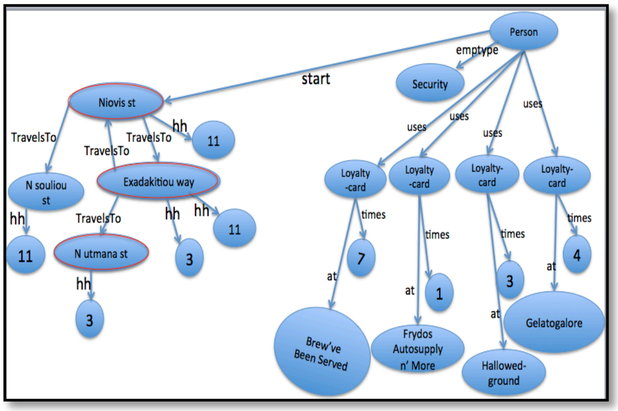


Figure 4: Unusual event for employee "Osvaldo Hennie".

Event 7 - 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".

Event 8 - Balas Felix spent 5 hours around midnight at "n ketallinias st 4600 4650" (near "Spetson Park").

Event 9 - Calzas Axel on 01/10 around midnight spent 4 hours at "n ketallinias st 4600 4650" (near Spetson Park).

Event 10 - Dedos Lidelse on day 01/10 spent 4 hours at night at "n ketallinias st" and passed through "niovis st".

Figure 3 and Figure 4 show normative and unusual patterns for employee Osvaldo Hennie. The patterns identified are significant because at least 8 employees are moving around locations of "Spetson Park" and "Chostus Hotel" which are away from their office (or) their regular eating-places.

For a considerable number of employees, GPS recordings were either incomplete or missing for given days. For example employee Herrero Kanon was missing data for 01/06 with no GPS locations recorded. For 01/07 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 01/10 and 01/11. Similarly, employee Osvaldo Hennie is missing data for the day 01/12. Since the most important suspicious event we look for is between 01/10 and 01/11, we still conclude the employee as suspicious.

4 CONCLUSION

In this work, we have investigated the problem of detecting suspicious behaviour using a graph-based approach. GBAD was able to discover the normative patterns for each employee as well as the unusual events using its "probabilistic" algorithm. In summary, our graph-based approach consistently discovered the most unusual patterns for effectively uncovering suspicious events. In future work, we could augment our anomaly score approach with some clustering or ranking algorithm based on our observed criteria. In addition, using pre-processing heuristics or a ranking algorithm on observed criteria, we could reduce the size of the graph, thereby reducing potential computational costs.

REFERENCES

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

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