Detecting Employee Leaks Using Badge and Network IP Traffic

William Eberle* Tennessee Tech University
Lawrence Holder† Washington State University
Jeffrey Graves‡ Tennessee Tech University

ABSTRACT

In order to detect an embassy employee leaking information to the outside world, we used the Graph-Based Anomaly Detection (GBAD) tool to focus the visualization on interesting structural anomalies. GBAD discovers anomalous instances of structural patterns in data, where the data represents entities, relationships and actions in graph form. In the provided data set, we analyze the proximity and network traffic logs in an attempt to locate possible instances of an insider threat. Through GBAD, we are able to discover anomalies to the normative structure of employee movements and activities in a fictional embassy.

Keywords: graph-based anomaly detection, insider threat

Index Terms: I.2.6 [Learning]: Knowledge Acquisition; I.2.4 [Artificial Intelligence]: Knowledge Representation Formalisms and Methods; I.5.1 [Pattern Recognition]: Models - Structural

1 INTRODUCTION

As part of the IEEE Symposium on Visual Analytics Science and Technology (VAST) for 2009, three mini-challenges and one grand challenge have been posted as part of their annual contest. Each of the mini-challenges consists of various aspects of a fictional insider threat, based upon the leaking of information. The goal of these challenges is to allow contestants to apply various visual analysis techniques so as to discover the spy and their associated actions.

In response to one of the mini-challenges, we chose to analyze the badge and network IP traffic. The proxy data set is comprised of employee "badge swipes" during the month of January in 2008, and the IP log consists of all network activity to and from the facility. The goal of this mini-challenge is to answer two questions about this data: (1) What computers did the spy use to send the sensitive information, and (2) Characterize the patterns of suspicious behavior of computer use.

2 DISCUSSION

In order to analyze the badge and network traffic, we used the Graph-Based Anomaly Detection (GBAD) tool to focus the visualization on interesting structural anomalies [3]; GBAD discovers anomalous instances of structural patterns in data, where the data represents entities, relationships and actions in graph form. Input to GBAD is a labeled graph in which entities are represented by labeled vertices and relationships or actions are represented by labeled edges between entities. GBAD embodies novel algorithms for identifying the three possible changes to a graph: modifications, insertions and deletions. Each algorithm discovers those substructures that match the closest to the normative pattern without matching exactly. As a result, GBAD is looking for those activities that appear to match normal patterns, but in fact are structurally different. GBAD uses the SUBDUE graph-based data mining system [2] as the engine for discovering the normative pattern in a graph. It is our hypothesis that such a system can discover knowledge in a graph representation of the badge and network traffic data that will (1) show the normal structure of the employee movements and network activity, and (2) show anomalies in employee behavior, indicating a possible insider threat.

In order to answer the challenge, we decided to focus on the movements and locations of the employees, along with their connections to the network. Based upon all of the information that was provided with the challenge, we made the following assumptions about this particular data set:

- Any employee can piggyback from one area to another, as long as someone else will open the door for them; nobody is required to use their badge.
- No employee used a computer that was not assigned to him or her, for fear of discovery (or termination).
- No employee spent the night at the embassy. Any activity in the embassy without record of entry is a sign of piggybacking.

Starting with these simple assumptions, we created graphs based upon the movement of employees between areas (outside, building, classified) and the number of connections that were made by the employee each time they were in the building, where vertices represented locations and network connections, and edges indicated order of movements.

The graph topology was designed manually, as the choice of an appropriate graph topology is domain dependent. For this mini-challenge, our graphs consisted of subgraphs that represented employee movements for a particular day. Each subgraph contained a "backbone" of **movement** vertices. Attached to the **movement** vertices were two vertices representing where the person started and ended (i.e., outside, building, classified). The edges were labeled **start** and **end**. If network traffic was sent before the person moved again, a **network** vertex was created and linked to the **movement** vertex via a **sends** edge. The **network** vertex was also linked to a vertex with a numerical label, representing how many messages that were sent before the next movement occurred. Also attached to a **movement** vertex via a **time** edge was a vertex representing the time reported in the proximity log (e.g., early_morning 0:00-7:59, morning 8:00-11:59, after_noon 12:00-16:59, evening 17:00-20:59, night 21:00-23:59). A numerical vertex representing the hour was also connected to the time vertex via an **hour** edge.

In the example shown in Figure 1, a person entered the building in the **early_morning** between 7AM and 8AM. The person sent 2 network messages and then moved into the classified area in the **morning** between 8AM and 9AM. The person then left the classified area in the **morning** between 9AM and 10AM.

A graph input file for GBAD is an ASCII text file that defines the vertices using sequential numbering and the edges using numbered vertices. A C program was used to process the proxLog.csv and IPLog3.5.csv files and output a graph file for use with GBAD. An example (partial) graph input file, created using this method, looks like the following:

```
v 1 movement
v 2 building
e 1 2 start
```

Once the graph files are created, GBAD is used to obtain (1) the normative pattern discovered in the specified graph input file

*e-mail: weberle@tntech.edu
†e-mail: holder@wsu.edu
‡e-mail: jagraves21@tntech.edu

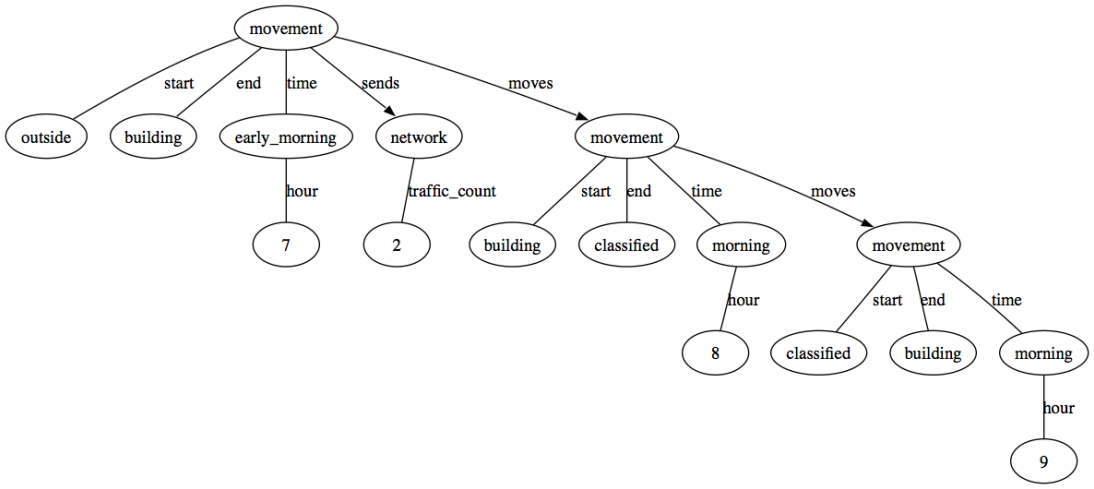

Figure 1: Example subgraph.

and (2) the top-N most anomalous patterns. The graph input file and discovered patterns are then converted to the dot format and visualized in GraphViz [1].

We initially created one graph of all employee activity for all days and were able to discover the normative pattern for all employees across all days. Figure 2 shows a visualization of the normative pattern. After uncovering the normative pattern, GBAD then uses three algorithms to discover all of the possible structural changes that can exist in a graph (i.e., modification, deletions, and insertions). Both the process to discover the normative pattern and the anomalies is done automatically with a single run of GBAD.

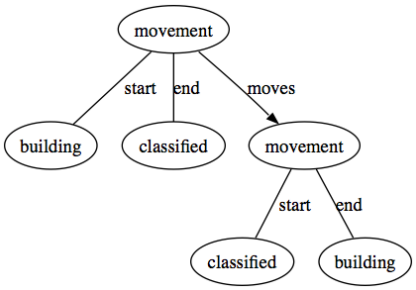

Figure 2: Normative pattern.

3 Observations

In order to determine which employee was the insider threat, we manually ranked our observations based upon which employees were involved in the following types of attributes: piggybacking, movement, network activity, time of day. Based upon these criteria, we suspected that employee number 38 was involved because of patterns of behavior such as:

- Piggybacking into the classified area.
- Found sending network traffic with no record of entry.
- Weekend activity.
- Large number of network connections.
- Activity at unusual times of the day.

Figure 3 shows an example of one of the anomalous instances reported by GBAD for this employee. We noticed some other interesting observations about other employees. It seemed like employee 12 liked to work late, sometimes close to midnight. It was somewhat common for employee 8 to exit the building in the middle of the day after making network connections, and return later in the day. Employee 26 moved around the embassy significantly more than other employees.

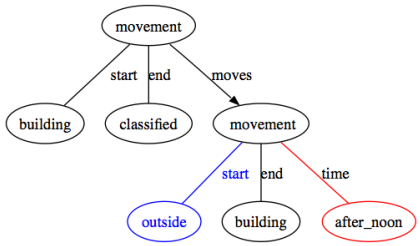

Figure 3: Example of unusual movement by employee at an abnormal time for that employee.

4 Conclusion

In the end, while we were able to discover various anomalies in the data, our assumption that the spy would not use someone else's computer turned out to be a wrong assumption. As a result, none of the employees we suspected turned out to be the spy, although two of the suspected employees did have their computers compromised by the spy. GBAD can be used to detect anomalies of possible insider threat activity in a graph representation of data that captures relational information. For this challenge, we missed detecting the actual spy not only because of our incorrect assumption, but also because we did not search for computers that were being used when their owners were not in the area. A possible solution could be to have multiple graph representations, where the detection process is accomplished in increments. First, we could detect the suspicious network activity, then we could determine which employees were involved, and third we could then analyze a timeline of events in order to discover the one employee (or employees) who could have accessed the computers.

Acknowledgements

This material is based upon work supported by the Department of Homeland Security under Contract No. N66001-08-C-2030. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of Homeland Security.

References

[1] AT&T GraphViz, www.graphviz.org.

[2] 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.

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