Analyzing Catalano/Vidro Social Structure Using GBAD

William Eberle *
Tennessee Technological University

Lawrence Holder **
Washington State University

ABSTRACT

In order to analyze the Paraiso cell-phone traffic, specifically the Catalano/Vidro social structure, 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. 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. Using the minimum description length (MDL) principle to identify the normative pattern that minimized the number of bits needed to describe the input graph after being compressed by the pattern, 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. Through GBAD, we are able to discover anomalies to the normative structure of the VAST Catalano/Vidro social network.

KEYWORDS: graph-based anomaly detection, social structure.

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 2008, four 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 movement, based upon the ideology of the fictitious Paraiso Movement. The goal of these challenges is to allow contestants to apply various visual analysis techniques so as to uncover patterns and anomalies in the data.

In response to this challenge, we chose to analyze the Cell Phone Calls on Isla del Sueno. This data set is comprised of 400 cell phone calls over a ten day period in June of 2006. The goal of this mini-challenge is to answer two questions about this data set: (1) What is the Catalano/Vidro social network as reflected in the calling data, and (2) Characterize the changes in the Catalano/Vidro social structure over the ten day period.

In order to analyze the cell-phone call records, we used our Graph-Based Anomaly Detection (GBAD) system [2]. GBAD takes a graph-representation of data and applies three algorithms that analyze the graph for structural anomalies. Each of these algorithms is applied after the normative graph structure has been discovered. It is our hypothesis that such a system can discover knowledge in a graph representation of the Cell Phone Calls (social network) data that will (1) show the normal social structure of the Catalano's and Vidro's, (2) show when the social structure has been broken, and (3) show anomalies in the social behaviour of this group, indicating possible breaches to their "inner circle". These normative patterns and anomalies can be highlighted in most visualization tools; in our case, we used GraphViz [4].

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

2 THEORY

GBAD is an *unsupervised* approach, based upon the SUBDUE graph-based knowledge discovery system [1]. Using a greedy beam search and Minimum Description Length (MDL) heuristic, each of the three GBAD anomaly detection algorithms uses SUBDUE to find the normative pattern in an input graph. In our implementation, the MDL approach is used to determine the best pattern as the one that minimizes the following:

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

where G is the entire graph, S is the substructure pattern, $DL(G|S)$ is the description length of G after compressing it using S, and $DL(S)$ is the description length of the substructure.

Within GBAD, we have developed three separate algorithms: GBAD-MDL, GBAD-P and GBAD-MPS. Each of these approaches is intended to discover all possible structural graph-based anomalies [3].

2.1 Information Theoretic Algorithm (GBAD-MDL)

The GBAD-MDL algorithm uses the MDL heuristic to discover the normative (best compressing) pattern in a graph, and then subsequently examines all of the instances for similar patterns. Using an inexact matching approach, the GBAD-MDL algorithm reports those instances that are the "closest" (without matching exactly) in structure to the normative pattern.

2.2 Probabilistic Algorithm (GBAD-P)

The GBAD-P algorithm also uses the MDL evaluation technique to discover the normative pattern 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 subtle difference between the two algorithms is that GBAD-MDL 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 possible extensions.

2.3 Maximum Partial Substructure (GBAD-MPS)

The GBAD-MPS algorithm again uses the MDL approach to discover the normative pattern in a graph, then it examines all of the instances of parent (or ancestral) substructures that are missing various edges and vertices. The value associated with the parent instances represents the cost of transformation (i.e., how much change would have to take place for the instance to match the

normative pattern substructure). Thus, the instance with the lowest-cost transformation (if more than one instance have the same value, the frequency of the instance is used to break the tie if possible) is considered the anomaly, as it is closest (maximum) to the normative pattern substructure without being included on the substructure's instance list.

3 DISCUSSION

In order to answer the challenge, we decided to focus on the social interactions by creating a graph of the social network that indicated for a particular day, who called who. Based upon all of the information that was provided with the challenge, we made the following assumptions about this particular data set:

- The person with an ID of 200 is Ferdinando Catalano.
- Anyone that Ferdinando Catalano calls (or that calls him) is in his "inner circle".
- The person with an ID of 5 is Estaban Catalano, Ferdinando's brother, as he is called the most frequently.
- The person with an ID of 1 is David Vidro, as he talks the most to the others that Ferdinando talks to.

Starting with these simple assumptions, and a graph that consisted of vertices for each unique ID with links between the vertices if there was a conversation on that day, we were able to create a simple visualization of Ferdinando's "inner circle" social network structure (or Catalano/Vidro social structure) over the 10 days that data was generated

3.1 Normative Pattern and Anomalies

Figure 1 was rendered using AT&T's graph visualization program GraphViz [4]. This visualization shows the graph structure of interactions between people in 200's (i.e., Ferdinando Catalano's) inner circle (i.e., 200, 1, 2, 3, 5, 97 and 137), the normative pattern within the graph, and the anomalous patterns in terms of the normative pattern.

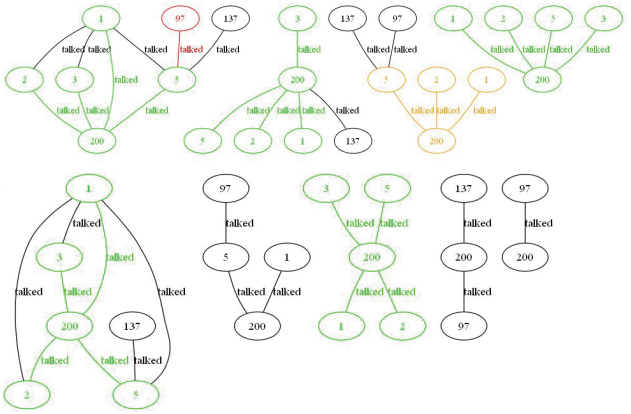

Figure 1 Ferdinando/Catalano's social structure with associated normative pattern and anomalies.

In Figure 1, the structure in green indicates the normative pattern that was discovered in this graph. The substructure in red indicates an anomaly that was discovered by the GBAD-P algorithm, which analyzes a graph for anomalous extensions to the normative pattern. In this case, the fact that 5 called 97 was anomalous, when compared to other instances of what was the normal social structure. The substructure in orange indicates an anomaly that was discovered by the GBAD-MPS algorithm, which analyzes a graph for missing structure. In this case, the fact that 200 did not talk to 3 on that day is considered anomalous.

3.2 Observations

Looking at the visualization shown in Figure 1 of the Catalano/Vidro calling-history, we are able to make several interesting observations about his social network:

- Notice that there are only 9 substructures in the graph. This is due to the fact that on Day 8, nobody in 200's inner circle talked to each other. In other words, there were no calls between 1, 2, 3, 5, 97, 137 or 200 on that day.
- Catalano/Vidro's "normative" social pattern only occurs on Days 1, 2, 4, 5 and 7.
- Nobody from the "normative inner circle" (i.e., 1, 2, 3, 5 and 200), communicates with anyone else in the normative circle after Day 7. Could it be that Ferdinando sent them to the United States at this point?
- 200 communicates with both 97 and 137 on Day 9, and just 97 on Day 10.
- 200 is involved in an "inner-circle" conversation on every day (except Day 8).

We also played with several other variants of the graph, including the "directedness" of the graph. While we chose an undirected graph for all of the results shown above (because we considered a conversation between two people to be a two-way communication), we also looked at a directed version of the graph, where the edge between two vertices was directed going from the person who called to the person who was being called. When we did that, we noticed that 97 and 137 are never called by 1, 2, 3 and 5 – and they only call 5 and 200.

4 CONCLUSION

The advantage of this approach in terms of this contest is that we are able to show how a simpler visualization technique, combined with graph mining tools, can help an analyst focus their search for relevant information in what can be a fairly complex network of communications. Traditional data mining approaches involve the probabilities and distributions of data values, while a graph-based approach such as this can discover differences in data when structure and relationships are represented as nodes and links.

While the cell-phone calls were made over a ten-day period, as was shown, we chose to represent each day as an individual sub-graph. In the future, we will investigate the representation of dynamic graphs which would then allow us to indicate a change in the structure over time.

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

[2] William Eberle and Lawrence Holder. "Anomaly Detection in Data Represented as Graphs". Intelligent Data Analysis, An International Journal, Volume 11(6), 2007.

[3] William Eberle and Lawrence Holder. "Mining for Structural Anomalies in Graph-based Data". International Conference on Data Mining (DMIN), June 2007.

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