Applying Graph-Based Anomaly Detection Approaches to the Discovery of Insider Threats

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

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

***Abstract*—The ability to mine data represented as a graph has become important in several domains for detecting various structural patterns. One important area of data mining is anomaly detection, but little work has been done in terms of detecting anomalies in graph-based data. In this paper we present graph-based approaches to uncovering anomalies in applications containing information representing possible insider threat activity: e-mail, cell-phone calls, and order processing.**

Keywords- anomaly detection; minimum description length; insider threat

I. INTRODUCTION

The ability to mine structurally complex data has become the focus of many initiatives, ranging from business process analysis to cyber-security. Since September 11, 2001, there has been an increasing emphasis on applicable methods for analyzing everything from bank transactions to network traffic, as our nation scours individual communications for possible illegal or terrorist activity.

One particular domain that has garnered much interest is e-mail traffic. As the New York Times reported, it "... is a potential treasure trove for investigators monitoring suspected terrorists and other criminals…" [4]. Up until recently, researchers have struggled with being able to obtain corporate e-mail due to the obvious restrictions placed on releasing what could be sensitive information. However, with the Federal Energy Regulatory Commission publication of the e-mail associated with the infamous Enron Corporation, researchers now have access to a rich data set of correspondences between management, lawyers and traders, many of whom were directly involved in the scandal. One of the areas that this paper explores is the detection of anomalies in the structural information that can be found in the flow of e-mail traffic.

Another domain that has been the subject of data mining activities involves the analysis of phone calls. Organizations such as the National Security Agency (NSA) have spent the last several years collecting suspicious phone calls and storing them in a database [6]. The significance of being able to peruse phone call information lies in the fact that an analyst can see who called whom, when they talked, for how long they talked, and the location of both parties. In the case of cell-phone calls, one can also ascertain the specific global position of two entities. Such information has been formative to not only general data mining research, but more specifically, research in diverse areas such as marketing, terrorist monitoring, and social network analysis. This paper will also explore the detection of anomalies in the structural data that can be found in phone traffic.

Recently there has been an impetus towards analyzing multi-relational data using graph-theoretic methods. In partial response to some of the current issues associated with *graph-based anomaly detection*, and the application to the domains discussed, we have developed three novel algorithms for analyzing graph substructures for the purpose of uncovering three types of graph-based anomalies: modifications, insertions and deletions. In this paper, we define what we consider to be an anomaly as it relates to graphs. Then, we present results when applying our anomaly detection approaches to e-mails, cell-phone traffic, and order processing information.

II. GRAPH-BASED ANOMALIES

The idea behind the approach used in this work is to find anomalies in graph-based data where the anomalous substructure in a graph is part of (or attached to or missing from) a *normative substructure*.

Definition: *A graph substructure S' is anomalous if it is not isomorphic to the graph's 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*. The importance of this definition lies in its relationship to any deceptive practices that are intended to illegally obtain or hide information. 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" [3].

There are three *categories of graph 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 deal with the unexpected absence of a vertex or edge. GBAD (Graph-Based Anomaly Detection) [2] is an *unsupervised* approach, based upon the SUBDUE graph-based knowledge discovery method [1]. Using a greedy beam search and Minimum Description Length (MDL) heuristic [5], each of the three anomaly detection algorithms in GBAD uses SUBDUE to find the best substructure, or normative pattern, in an input graph. The MDL approach is used to determine the best substructure as the one that minimizes the following:

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

where *G* is the entire graph, *S* is a 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. In order to

discover each possible anomaly type, we have implemented three algorithms in GBAD, each with the purpose of discovering a specific type of anomaly. For anomalous modifications, our **GBAD-MDL** algorithm uses the MDL heuristic to discover the best substructure in a graph, and then subsequently examines all of the instances of that substructure that "look similar" to that pattern. For anomalous insertions, our **GBAD-P** (probability) 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 pattern, looking for extensions with the lowest probability. Then, for anomalous deletions, our **GBAD-MPS** (maximum partial substructure) algorithm again uses the MDL approach to discover the normative pattern in a graph, and then it examines all instances of *ancestral* substructures that are missing various edges and vertices. The reader should refer to [2] for a more detailed description of the algorithms

III. E-MAIL CORRESPONDENCES

One of the more recent domains that have become publicly available is the data set of e-mails between employees from the Enron Corporation. The Enron e-mail dataset consists of not only messages, but also employee information such as their full name and work title. By limiting our graph to the Enron employees and their correspondences, we are able to not only create a "social network", but also discover anomalous behaviors among *classes* of individuals. Thus, we generated graphs based upon the social aspect and company position of employees that start a "chain" of e-mails, where a chain consists of the originating e-mail and any subsequent replies or forwards to that corresponding e-mail. Each graph consists of the substructures shown in Figure 1.

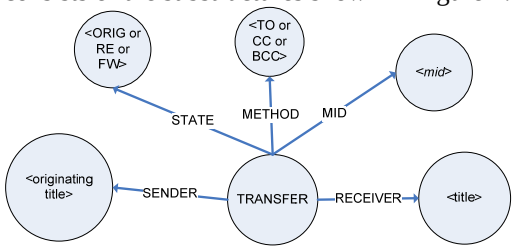

Figure 1. Graph substructure of e-mail data set.

In this representation, a graph consists of individual, disconnected substructures that represent the "flow" of each e-mail that originates from someone with a specified employment title (e.g., Director). An e-mail can be sent by one or more TRANSFERs to one or more individuals with varying employment titles (represented by a directional arrow to show who sent the message to whom), and can either be sent back (as a reply or forward) to the <originating title>, or forwarded/replied on to other <title> entities. There is no limit to the number of times a message can be replied/forwarded.

There are many different employee titles within Enron (i.e., Managers, Directors, CEOs, etc.), and each of the GBAD algorithms were able to show different structural anomalies in the chains of e-mails that originated along people's company titles. For instance, running GBAD on the graph that consists of e-mails originating from Directors, the anomalous instance shown in Figure 2 is discovered.

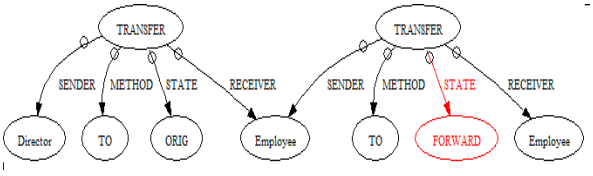

Figure 2. Anomalous instance (portion) of e-mail being forwarded.

This anomalous instance consists of a message being sent from a Director to an Employee (i.e., non-management personnel), that was then forwarded to another non-management Employee. What is interesting about this anomaly is that the data set consists of many e-mails that are sent "TO" "Employee"s from "Director"s, but this is the only situation where the Employee FORWARDed the e-mail onto another "Employee", who was not privy to the original e-mail. Specifically, the e-mail started with Hyatt (director) regarding "Oasis Dairy Farms Judgement", who sent it to Watson (employee), who then forwarded it to Blair (employee).

While applying GBAD-MPS and GBAD-P to the graph of e-mails originating from personnel with the title of "Trader" does not produce any significant anomalies, the GBAD-MDL algorithm does produce two anomalous instances. Figure 3 shows two situations where a Trader was involved in a chain of e-mails that resulted in correspondences to a CEO and a President (respectively) that were not normal.

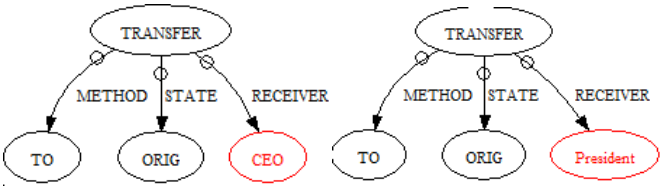

Figure 3. Anomalous instances of e-mails to a CEO and to a President.

In terms of the first anomalous instance, shown in Figure 3, from an e-mail entitled "Financial Disclosure of $1.2 Billion Equity Adjustment", there are only 4 e-mails that are sent to CEOs. But, this is the only example of an e-mail being sent TO a CEO - the other 3 e-mails are CCed to the CEO. In the case of the second anomalous instance shown in Figure 3, an e-mail entitled "Fastow Rumor", this is the only time that an e-mail is sent by a Trader to a President.

IV. CELL-PHONE TRAFFIC

As part of the 2008 IEEE Symposium on Visual Analytics Science and Technology (VAST), four mini-challenges and one grand challenge were posted as part of their annual contest (http://www.cs.umd.edu/hcil/VASTchallenge08/). While the goal of the challenge is to target new visual analytics approaches, it is still possible to apply these graph-based anomaly detection algorithms to the same data sets. One of the data sets consists of cell-phone traffic between inhabitants of the fictitious island of Isla Del Sueño. The data consists of 9,834 cell-phone calls between 400 people over a 10-day period. The challenge is to describe the social network of a religious group headed by Ferdinando Cattalano and how it changes over 10 days. The graph of the cell-phone traffic is represented as shown in Figure 4.

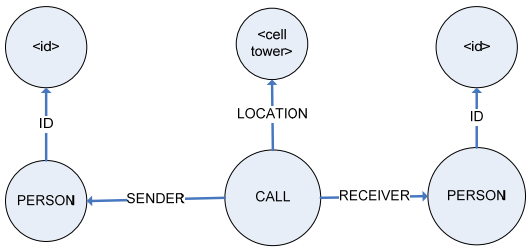

Figure 4. Graph of a cell-phone call from the VAST dataset.

Applying the GBAD algorithms to this information results in several structural anomalies within the data, especially when particular individuals are analyzed in terms of their calling patterns. For instance, when we look at the calling patterns of individuals who correspond with Ferdinando Cattalano, one notices several anomalous substructures, including some who contact Ferdinando on days that are out of the ordinary, and even some individuals who call others outside of their normal chain of cell-phone calls. In addition, GBAD was applied to a graph of the *social network* of phone usage, yielding additional anomalous behavior between targeted persons. From these results, we are able to determine members of the normative social network surrounding Ferdinando and when the network begins to break down.

V. ORDER PROCESSING

We have also evaluated the GBAD approach for identifying anomalies on information flows – data that passes between people or entities. In order to do this, we are using the OMNeT++ event simulator (www.omnetpp.org) to model information flows, generate flow data, represent the data in graph form, and then analyze the graphs using GBAD. This process has two main benefits. First, we can model different flows with known structure and anomalies, which allows us to easily verify GBAD's ability to detect these anomalies. Second, the OMNeT++ framework can be used to model real business processes, or other types of information flows, to further evaluate the real-world applicability of the GBAD approach. Here we give a brief introduction of this process on a simple order-fulfillment example.

Consider the order-fulfillment process depicted in Figure 5. The process is initiated by the Customer placing an Order, which is sent to the Sales department. The Sales department sends an Order Acknowledgement back to the Customer and sends an Internal Order to the Warehouse. Once the Warehouse ships the order, they send a Delivery Note to the Customer. One possible anomaly could be when someone in the Sales department sends the Order to an Unknown entity, perhaps to leak insider information to a competitor.

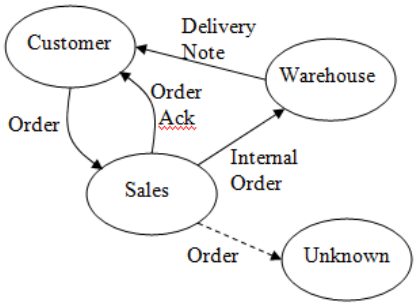

Figure 5. Depiction of information flow during an order fulfillment process.

To simulate the information flow associated with order processing using OMNeT++, each node in the process is defined as a module in the Network Description (NED) language. Modules (like OrderProcess) can consist of sub-modules and their interconnections. The actual function of each module (how it processes messages) is defined in C++. For instance, we implemented a handleMessage method of the Sales module, which waits for Order messages. After receiving an Order message, the Sales module waits 10-60 seconds and then sends an Order Acknowledgement message to the Customer module, sends an Internal Order message to the Warehouse module, and with a Bernoulli probability of 0.001 (as defined in the OMNet++ initialization file) sends an Order message to the Unknown module (depicted in Figure 5).

In addition to logging information produced by OMNeT++, we are able to specify GBAD-related messages to be printed from each module describing order-related messages as they are sent and received by the modules. It is this information we use to construct graphs of the information flow. A utility program called "o2g" converts the GBAD-enhanced OMNeT++ simulation output into the graph input format required by our GBAD implementation. GBAD then finds the normative patterns and anomalies in the graph. For the experiment depicted in Figure 5, representing the processing flow of 1,000 orders, we generated a graph of approximately 3,000 vertices and 4,000 edges. From this graph, GBAD is able to successfully discover, with no false-positives, the anomalous (dashed) edge and vertex shown.

VI. CONCLUSIONS AND FUTURE WORK

Using the MDL principle and probabilistic approaches, we have been able to successfully discover anomalies in graphs and patterns of varying sizes with minimal to no false positives. Results from running the GBAD algorithms on e-mail, cell-phone traffic and business processes show how these graph-theoretic approaches can be used to identify insider threats. Some future directions that we are exploring include the incorporation of traditional data mining approaches as additional quantifiers to determining anomalousness, as well as applying graph-theoretic algorithms to dynamic graphs that change over time. In addition, using the OMNet++ example, we can create limitless numbers and varieties of simulations modeling business processes, network traffic, email flows, etc. These can then be used to evaluate GBAD systematically and on models of real-world processes.

REFERENCES

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

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

[3] Hampton, M. and Levi, M. *Fast spinning into oblivion? Recent developments in money-laundering policies and offshore finance centres*. Third World Quartely, Vol. 20, Num 3, June 1999, pp. 645-656, 1999.

[4] Kolata, G. "Enron Offers an Unlikely Boost to E-Mail Surveillance", May 22, 2005, www.nytimes.com.

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

[6] Cauley, L. "NSA has massive database of Americans' phone calls", USA Today, May 11, 2006.