Insider Threat Detection Using Graph-Based Approaches

William Eberle
Tennessee Technological University
weberle@tntech.edu

Lawrence Holder
Washington State University
holder@wsu.edu

1. Introduction

Protecting our nation's cyber infrastructure and securing sensitive information are critical challenges for homeland security and require the research, development and deployment of new technologies that can be transitioned into the field for combating cyber security risks. Particular areas of concern are the deliberate and intended actions associated with malicious exploitation, theft or destruction of data, or the compromise of networks, communications or other IT resources, of which the most harmful and difficult to detect threats are those propagated by an insider. However, current efforts to identify unauthorized access to information, such as what is found in document control and management systems, are limited in scope and capabilities.

In order to address this issue, this effort involves performing further research and development on the existing Graph-Based Anomaly Detection (GBAD) system [3]. GBAD discovers anomalous instances of structural patterns in data that represent entities, relationships and actions. 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 minimizes the number of bits needed to describe the input graph after being compressed by the pattern, GBAD implements 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 (or legitimate) transactions, but in fact are structurally different.

As a solution to the problem of insider threat detection, we will apply GBAD to datasets that represent the flow of information between entities, as well as the actions that take place on the information. This research involves the representation of datasets, like a document control and management system, as a graph, enhancement of GBAD's performance levels, and evaluation of GBAD on these datasets. In previous research, GBAD has already achieved over 95% accuracy detecting anomalies in simulated domains, with minimal false positives, on graphs of up to 100,000 vertices.

2. Motivation

Information Technology organizations need mechanisms for detecting possible insider threats that affect their organization's network, systems and information. By applying the approaches implemented within GBAD, an analyst will be able to detect behavior that is attempting to hide illegitimate actions by mimicking legitimate transactions. Specifically, the GBAD approach can be used to address the following security concerns: (1) Potential violations of system security policy by an authorized user; (2) Deliberate and intended actions such as malicious exploitation, theft, or destruction of data; (3) Compromise of networks, communications, or other IT resources; and (4) Differentiation of suspected malicious behavior from normal behavior.

3. Technical Approach

The ability to mine relational data has become important in several domains for detecting various structural patterns. One important area of data mining is anomaly detection, particularly for fraud. The ability to mine data for nefarious behavior is difficult due to the *mimicry* of the perpetrator. If a person or entity is attempting to commit fraud or participate in some sort of illegal activity, they will attempt to convey their actions as close to legitimate actions as possible. 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" [2]. Recently there has been an impetus towards analyzing relational data using graph theoretic methods [1]. Graph-based data mining approaches analyze data that can be represented as a graph (i.e., vertices and edges).

978-0-7695-3568-5/09 $25.00 © 2009 IEEE
DOI 10.1109/CATCH.2009.7

While there are approaches for using graph-based data mining for intrusion detection [4], little work has been done in the area of graph-based anomaly detection, especially for application to areas like document control and management systems.

Take for instance the document flow scenario of an order processing system, as shown in Figure 1. This example would consist of individual transactions where personnel receive, process and possibly pass on documents to other personnel or departments. However, when this information is represented as a graph, possible anomalous actions, for example if the Sales department also passed the Order Acknowledgement to another customer (providing "inside information"), can be considered "additional structure" within the graph that was an unexpected deviation from the normal pattern of document flow.

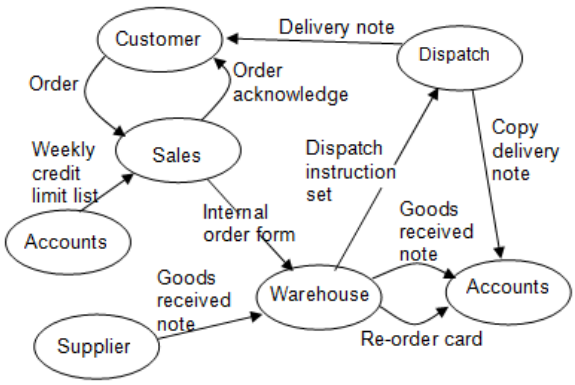

Figure 1 Document management and control.

We have developed novel algorithms for analyzing graph substructures for the purpose of uncovering all three types of graph-based anomalies: modifications, insertions and deletions. The idea behind our approaches 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 non-anomalous substructure, or the *normative substructure*. This definition of an anomaly is unique in the arena of graph-based anomaly detection, as well as non-graph-based anomaly detection. Some anomaly detection methods use a supervised approach, which requires some sort of baseline of information from which comparisons or training can be performed. In general, if one has an idea what is normal behavior, deviations from that behavior could constitute an anomaly. However, the issue with those approaches is that one has to have the data in advance in order to train the system, and the data has to already be labeled (i.e., fraudulent versus legitimate). Other anomaly detection methods, such as clustering, use an unsupervised approach. Objects that fall outside a cluster (outliers), usually within a specified deviation, are considered candidate anomalies. However, they are usually based upon statistical evaluations and do not take into account relational information. Our work up until now has resulted in the development of three algorithms, which we have implemented using a tool called GBAD (Graph-Based Anomaly Detection). GBAD is an *unsupervised* approach, based upon the SUBDUE graph-based knowledge discovery system [5]. Using a greedy beam search and Minimum Description Length (MDL) heuristic, each of the three anomaly detection algorithms uses SUBDUE to discover the most prevalent substructure, or normative pattern, in an input graph. In our implementation, the MDL approach is used to determine the best substructure(s) as the one that minimizes $M(S,G) = DL(G|S) + DL(S)$, 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.

In order to discover each of the possible anomaly types, we have implemented three algorithms in GBAD, each with the purpose of discovering a specific type of anomaly. For anomalous graph 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. Using an inexact matching algorithm, the result will be those instances that are the "closest" (without matching exactly) in structure to the normative pattern (i.e., compresses the graph the most), where there is a tradeoff between the cost of transforming the instance to match the pattern and the frequency with which the instance occurs. Since cost of transformation and frequency are independent variables, multiplying their values together results in a combinatory value; the lower the value, the more anomalous the structure. It is these inexact matching instances that are analyzed as anomalies.

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. 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 including edges and vertices that are probabilistically less likely than other possible extensions.

Finally, for anomalous graph deletions, our GBAD-MPS (maximum partial substructure) 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 substructure). Thus, the instance with the lowest cost of transformation and lowest frequency is considered the anomaly, as it is closest to the normative substructure without being an instance of this substructure.

4. Experiments

We initially identified the following datasets for use in this effort. These datasets were selected based on their representation of information flow and their relevance to the task of insider threat detection.

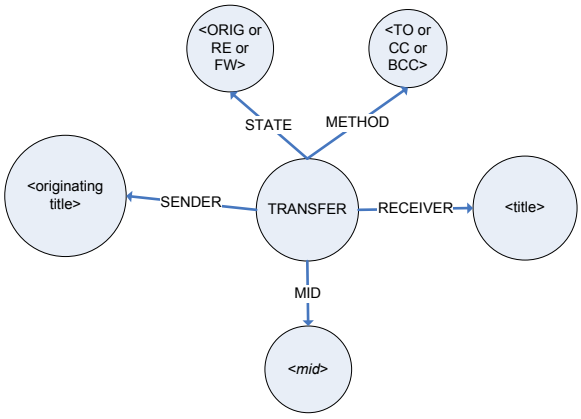

Figure 2 Graph representation of e-mail.

E-mail

The Enron e-mail dataset is a collection of e-mails from employees of Enron prior to the collapse of the corporation in 2001 (www.isi.edu/~adibi/Enron/Enron.htm). Looking at how messages normally flow between employees, we can analyze whether or not a particular chain of e-mails is anomalous or normal. After experimenting with several graph representations of e-mail flow, we settled on a graph structure that focused on the transfer of information between employees, including their company title and the method in which they received the e-mail (see Figure 2). From this representation, GBAD has been able to find several anomalies in the e-mail traffic.

For example, we have found several suspicious occurrences of executives communicating directly to low-level employees about financial data. While the normative pattern of e-mails involves transfers between non-management employees, there are two anomalous situations involving CEOs. In one case, a Trader abnormally sends an e-mail regarding "Financial Disclosure of $1.2 Billion Equity Adjustment" to a President. The anomalies detected by GBAD in the Enron data indicate the ability to find anomalous communications between individuals in an organization. While not all such communication is evidence of a threat, the uniqueness warrants further investigation that may uncover threat activity. A similar approach can be applied to the scenarios recently documented by CERT [7], in which employees wishing to do harm initiate unusual communications to other employees or to entities external to the organization. GBAD has the potential of identifying such anomalies.

Cell-Phone Traffic

Each year the IEEE Symposium on Visual Analytics Science and Technology (VAST) provides a challenge to identify patterns of interest in data. The 2008 challenge (www.cs.umd.edu/hcil/VASTchallenge08/) includes information on cell-phone calls made over a 10-day period, during which an event of interest occurs. Based on a simple graph representation of the cell-phone traffic, the application of GBAD to this VAST data allowed us to detect the main social network inherent in the cell-phone traffic, as well as anomalies to this network over a 10-day period (Figure 3).

From this effort, we submitted our results to the 2008 VAST challenge, and a paper describing our approach will be included in the associated NIST website (http://vac.nist.gov/). The results for both the Enron and VAST datasets are indicative of the general application of GBAD to information flow data, where anomalies to the normal flow can reveal insider threats.

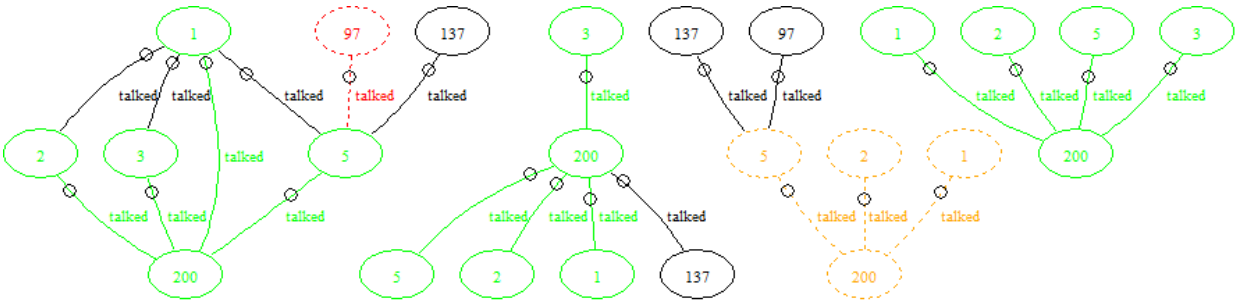

Figure 3 Social network (cell phone calls) with associated normative pattern and anomalies.

Process Model Simulator

We are also using the OMNeT++ public-domain discrete event simulator (www.omnetpp.org) as a platform to simulate flow data based on business process models. OMNeT++ is an extendible C++ library that offers maximum flexibility in representing business processes of interest to the sponsor of this research. We have implemented a simple order-processing model (see Figure 4) within OMNeT++, where an anomaly occurs when the sales department forwards an order outside the organization (e.g., to a competing customer). When executed on the model, OMNeT++ produces information that is converted to graph form for input to GBAD.

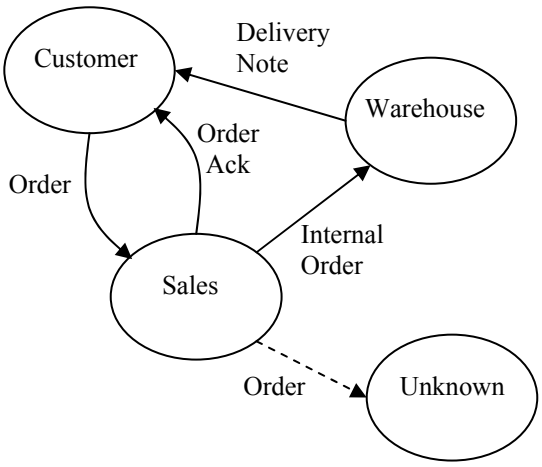

Figure 4: Depiction of information flow during an order fulfillment process (dashed anomalous edge)

In preliminary testing, GBAD was able to correctly identify a single anomalous order within a graph representing 1000 orders.

5. Future Work

The next step with the OMNeT++ simulator is to identify incidents from the recent CERT Insider Threat reports (http://www.cert.org/insider_threat/). There are several CERT insider threat documents that detail potential business process models and real insider threats that have occurred in such businesses. We have targeted a document access scenario based on incidents published in the CERT reports [7] and a general model described by Chun in [6]. We will extend this model to include database access processes and anomalies to these processes based on the patterns from the CERT insider threat incidents. We will then model the process in OMNeT++, generate synthetic graph data for input to GBAD, and evaluate GBAD's performance on various insider threat scenarios within this process model. We will also continue to analyze the Enron and VAST data sets for other types of anomalies, in order to further improve the usefulness of our approach. In addition, we have begun to do experiments with multiple normative patterns. In some cases, the anomalous substructure may not be a deviation of the most prevalent pattern, but instead deviates from only one of many normative patterns. For example, a graph of telephone calls across multiple customers or service providers would contain different calling patterns. The normative "behavior" of one customer would not be representative of another customer's calling pattern.

Experiments with GBAD on simulated datasets have shown an almost 100% discovery rate for each algorithm on graphs of varying sizes, with normative patterns and anomalies of varying sizes. Currently,

the accuracy and running times of the algorithms are relative to the size of the search space. As we increase the size of the search space (and subsequently the amount of memory), the detection accuracy increases along with the running time. In the future, performance enhancements to GBAD will be focused on reducing the time spent in the main computational bottleneck for GBAD: testing if two graphs match (i.e., graph isomorphism). This graph isomorphism check is expensive, but is necessary to find matches in the data to the graphical patterns discovered by GBAD. However, in some cases, a complete check is not required. Identifying and exploiting these cases will improve GBAD's performance. In the case where a complete graph match is required, we have placed limits on the graph match to ensure polynomial running times. Still, further speedups can be achieved by employing fast match for constrained graph types or canonical labeling techniques as used in frequent subgraph discovery [8]. Preliminary results indicate that significant speed-ups can be achieved, an order-of-magnitude speedup in some cases. We will test these enhancements on various types of graphs.

6. Acknowledgements

This material is based upon work supported by the U.S. Department of Homeland Security Science and Technology Directorate 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.

7. References

1. L. Holder and D. Cook. *Mining Graph Data*. John Wiley and Sons, 2007.
2. Hampton, M. and Levi, M. *Fast spinning into oblivion? Recent developments in money-laundering policies and offshore finance centres*. Third World Quarterly, Volume 20, Number 3, June 1999, pp. 645-656, 1999.
3. Eberle, W. and Holder, L. *Mining for Structural Anomalies in Graph-Based Data*. International Conference on Data Mining. June, 2007.
4. Staniford-Chen, S. et al.. *GrIDS – A Graph Based Intrusion Detection System for Large Networks*. Proceedings of the 19th National Information Systems Security Conference, 1996.
5. D. Cook and L. Holder. *Graph-based data mining*. IEEE Intelligent Systems 15(2):32-41, 1998.
6. H.W. Chun. *An AI Framework for the Automatic Assessment of e-Government Forms*, AI Magazine, Vol. 29, No. 1, Spring 2008, pp.41-51.
7. E. Kowalski, D. Cappelli, T. Conway, B. Willke, S. Keverline, A. Moore and M. Williams, "Insider Threat Study: Illicit Cyber Activity in the Government Sector," January 2008. URL: http://www.cert.org/archive/pdf/insiderthreat_gov2008.pdf.
8. X. Yan and J. Han, "gSpan: Graph-based Substructure Pattern Mining," IEEE International Conference on Data Mining, 2002.