# Graph-Based Approaches to Insider Threat Detection

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

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

## ABSTRACT
This work presents the use of *graph-based* approaches to discovering anomalous instances of structural patterns in data that represent entities, relationships and actions. Using the minimum description length (MDL) principle to first identify the normative pattern, the algorithms presented in this paper identify 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, this proposed approach searches for those activities that appear to match normal (or legitimate) transactions, but in fact are structurally different. After briefly presenting the three algorithms, we then show the usefulness of applying these graph theoretic approaches to discovering illegal activity for a simulated insider threat within a passport processing scenario.

## Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Applications - *Data Mining*.

## General Terms
Algorithms.

## Keywords
graph-based anomaly detection, insider threat.

## 1. INTRODUCTION
Protecting our nation's cyber infrastructure and securing sensitive information are critical challenges for both industry and homeland security. One of the primary concerns is 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. We propose to address these challenges by analyzing the relationships between entities in the data.


CSIIRW '09, April 13-15, Oak Ridge, Tennessee, USA


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 insider threat detection. 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 participate in some sort of illegal activity, they will attempt to convey their actions as close to legitimate actions as possible. Recent reports have indicated that approximately 6% of revenues are lost due to fraud, and almost 60% of those fraud cases involve employees [12]. The Identity Theft Resource Center recently reported that 15.8 percent of security breaches so far in 2008 have come from insiders, up from 6 percent in 2007 [1]. Various insider activities such as violations of system security policy by an authorized user, deliberate and intended actions such as malicious exploitation, theft, or destruction of data, the compromise of networks, communications, or other IT resources, and the difficulty in differentiating suspected malicious behavior from normal behavior, have threatened our nation's security. Organizations responsible for the protection of their company's valuable resources require the ability to mine and detect internal transactions for possible insider threats. Yet, most organizations spend considerable resources protecting their networks and information from the outside world, with little effort being applied to the threats from within.

Graph-based data mining approaches analyze data that can be represented as a graph (i.e., vertices and edges). While there are approaches for using graph-based data mining for intrusion detection [2], little work has been done in the area of *graph-based anomaly detection*, especially for application to business processes, such as in document control and management systems.

## 2. GBAD APPROACH
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 general *categories of anomalies*: insertions, modifications and deletions. Insertions would constitute the presence of an unexpected vertex or edge. Modifications would consist of an unexpected label on a vertex or edge. Deletions would constitute the unexpected absence of a vertex or edge.

## 2.1 Algorithms

GBAD (Graph-based Anomaly Detection) [13] is an *unsupervised* approach, based upon the SUBDUE graph-based knowledge discovery method [5]. Using a greedy beam search and Minimum Description Length (MDL) heuristic [6], each of the three anomaly detection algorithms in GBAD uses SUBDUE to find the best 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 the following:

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

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.

We have developed three separate algorithms: GBAD-MDL, GBAD-P and GBAD-MPS. Each of these approaches is intended to discover one of the possible graph-based anomaly categories as set forth earlier. The following is a brief summary of each of the algorithms, along with some simple business process examples to help explain their usage. The reader should refer to [4] for a more detailed description of the actual algorithms.

### *2.1.1 Information Theoretic Algorithm(GBAD-MDL)*

The GBAD-MDL algorithm uses a Minimum Description Length (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 – or more precisely, are *modifications* to the normative pattern. In Noble and Cook's work on graph-based anomaly detection [7], they present an example similar to the one shown in Figure 1.

Running the GBAD-MDL algorithm on this example results in the (circled) anomalous substructure. With Noble and Cook's approach, the D vertex is shown to be the anomaly. While correct, the importance of the GBAD approach is that a larger picture is provided regarding its associated substructure (i.e., the other three vertices A, B and D). Thus, not only are we providing the anomaly, but we are also presenting the context of that anomaly within the graph.

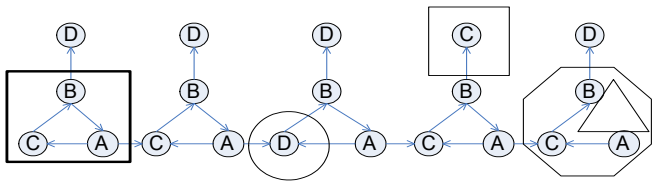


**Figure 1. Example with normative pattern (bold box) and different anomalies.**

### *2.1.2 Probabilistic Algorithm (GBAD-P)*

The GBAD-P algorithm 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, or *insertions*, to the normative substructure with the lowest probability. The difference between the algorithms is that GBAD-MDL is looking at instances of substructures with the same characteristics (e.g., size), whereas GBAD-P is examining the probability of extensions to the normative pattern to determine if there is an instance that includes edges and vertices that are probabilistically less likely than other possible extensions.

Take the same example shown in Figure 1. After one iteration, the instance shown in the **bold** box is one of the instances of the best substructure. Then, on the second iteration, extensions are evaluated, and the instance in the regular box (on top) is the resulting anomaly. However, again, it is important to note that the GBAD approach will report the entire instance as anomalous, not just the anomalous edge and vertex, providing a better context for analytical purposes.

### *2.1.3 Maximum Partial Substructure Algorithm (GBAD-MPS)*

The GBAD-MPS algorithm again uses the MDL approach to discover the best substructure in a graph, then it examines all of the instances of parent (or ancestral) substructures that are missing various edges and vertices (i.e., *deletions*). 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 best substructure). Thus, the instance with the lowest cost transformation is considered the anomaly, as it is closest (maximum) to the best substructure without being included on the best substructure's instance list. If more than one instance have the same value, the frequency of the instance's structure will be used to break the tie if possible.

Suppose we take one of the instances of the normative pattern (outlined by an octagon in Figure 1), and remove its edge between the B and A vertices (shown in the triangle). Running GBAD-MPS on the modified graph results in the discovery of an anomalous substructure similar to the normative pattern, but missing the removed edge.

# 3. INSIDER THREAT SCENARIO

In order to demonstrate the potential effectiveness of GBAD for detecting insider threats, we simulated a passport processing scenario that was motivated by two real-world sources of information. One source is the incidents reported in the CERT Insider Threat documents [8][9][10] that involve privacy violations in a government identification card processing organization and fraud in an insurance claim processing organization. The other model we used is based on the process flow associated with a passport application [11]. The outline of this process flow, depicted in Figure 2, is as follows:

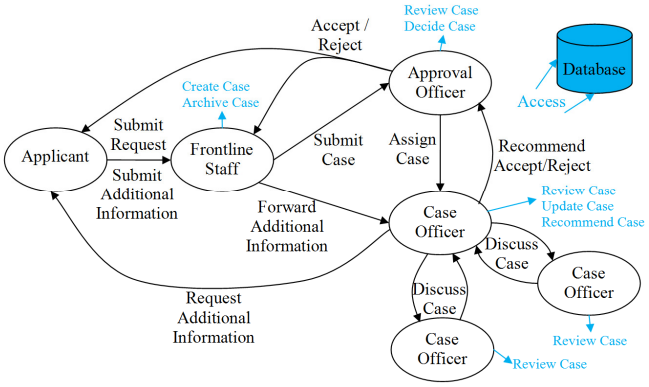


**Figure 2. Information flow in claim scenario.**

1. The applicant submits a request to the frontline staff of the organization.
2. The frontline staff creates a case in the organization's database and then submits the case to the approval officer.
3. The approval officer reviews the case in the database and then assigns the case to one of the case officers. By default, there are three case officers in this organization.
4. The assigned case officer reviews the case. The assigned case officer may request additional information from the applicant, which is submitted to the frontline staff and then forwarded to the assigned case officer. The assigned case officer updates the case in the database based on this new information. The assigned case officer may also discuss the case with one or more of the other case officers, who may review the case in the database in order to comment on the case. Ultimately, the assigned case officer will recommend to accept or reject the case. This recommendation is recorded in the database and sent to the approval officer.
5. Upon receiving the recommendation from the assigned case officer, the approval officer will make a final decision to accept or reject the case. This decision is recorded in the database and sent to both the frontline staff and the applicant.
6. Finally, upon receiving the final decision, the frontline staff archives the case in the database.

There are several scenarios where potential insider threat anomalies might occur, including:

1. Frontline staff performing a Review Case on the database (e.g., invasion of privacy).
2. Frontline staff submits case directly to a case officer (bypassing the approval officer).
3. Frontline staff recommends or decides case.
4. Approval officer overrides accept/reject recommendation from assigned case officer.
5. Unassigned case officer updates or recommends case.
6. Applicant communicates with the approval officer or a case officer.
7. Unassigned case officer communicates with applicant.
8. Database access from an external source or after hours.

Representing the processing of 1,000 passport applications, we generated a graph of approximately 5,000 vertices and 13,000 edges, and proceeded to replicate the scenarios described above.

For scenarios 1, 3 and 6, while the GBAD-MDL and GBAD-MPS algorithms do not discover any anomalous structures, GBAD-P is able to successfully discover the single anomalous cases out of 1,000 where staff is violating the process. For scenario 2, the GBAD-MPS algorithm successfully discovers all three instances where the frontline staffer did not submit the case to the approval officer.

For Scenario 4, we randomly modified three examples by changing the recommendation that the "CaseOfficer" sends to the "ApprovalOfficer". This scenario tests GBAD's ability to handle *multiple normative patterns*. Potentially, there are two types of prevalent patterns in this type of data: (1) The ApprovalOfficer and CaseOfficer both accept a passport application, and (2) The ApprovalOfficer and CaseOfficer both reject an application. Therefore, potentially anomalous scenarios could exist where the ApprovalOfficer overrides the accept/reject recommendation from the assigned CaseOfficer. We generated a graph consisting of these two normative patterns, although these patterns were not among the top-ranked most normative substructures. We then randomly inserted an anomalous instance of the first type (case officer accepts, approval officer rejects) and two anomalous instances of the second type (case officer rejects, approval officer accepts). Configuring the GBAD-P algorithm to analyze the top $N$ normative patterns, where $N$ is set arbitrarily to 20, all three anomalous examples are reported as the most anomalous. Other experiments showed that the size of $N$ was not important. For instance, in this example, when we increase $N$ to 100, the top three anomalies reported are still the same ones. In addition, no other substructures are reported as anomalous along with these top three anomalies (i.e., no false positives).

For scenario 5, we randomly inserted into two examples the situation where a "CaseOfficer" recommends to accept a case for which they were not assigned. In this scenario, GBAD-MDL does not report any anomalies, while both GBAD-MPS and GBAD-P each discover both anomalous instances. GBAD-MPS discovers the anomalies because the "CaseOfficer" has assigned himself to the case without any corresponding recommendation back to the "ApprovalOfficer" or "Database", while GBAD-P uncovers the extra "CaseOfficer" and his unauthorized assignment to the case. Figure 3 shows the normative pattern and the anomalous structures from one of these examples. Also, while not shown, this same structural anomaly can be found in scenario 7. Scenario 7 consists of an extra edge going from the unauthorized "CaseOfficer" node to the "Customer" node, and as such is only different from Scenario 5 by the label on the edge and the targeted node.

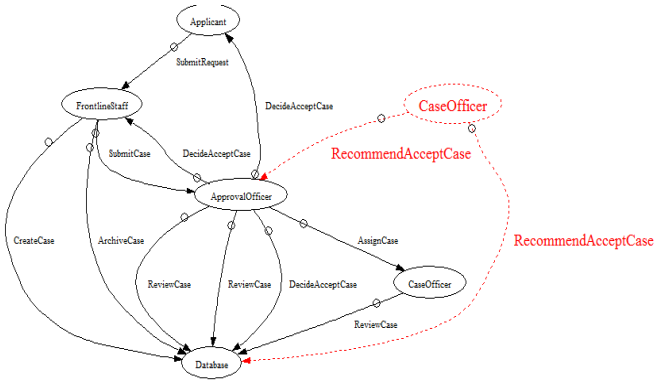


**Figure 3. Graph of Scenario 5, showing the unauthorized CaseOfficer's handling of a case.**

Finally, for scenario 8, we represented time in the graph as the number of hours since midnight, and we enhanced GBAD to use a simple statistical analysis of numerical attributes as part of its evaluation of the graph structure. In this case, we randomly inserted two anomalies into the graph, and the GBAD-P algorithm was able to successfully discover both anomalies where access to the company database was during unexpected hours, with no false positives reported. While the structure was the same, the time information (represented as a number), provides extra information that aides in the insider threat detection. Also, it is important to note that no false positives are reported with this scenario.

## 4. FUTURE WORK

In the future, we are going to continue researching other numeric analysis approaches that can be incorporated into the structural analysis so as to further delineate "anomalousness". In addition, we will analyze our ability to discover an anomaly involving two different numeric attributes that individually are not anomalous, but together are rare. We will also investigate the limitations involved with analyzing multiple normative patterns, including how well this approach scales with the size of the graph, number of normative patterns, and size of the normative patterns. In addition, we are exploring the incorporation of traditional data mining approaches as additional quantifiers to determining anomalousness, as well as applying graph-theoretic algorithms to dynamic graphs that are changing over time.

## 5. ACKNOWLEDGMENTS

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.

## 6. REFERENCES

[1] Foley, L., "ITRC Beach Meter Reaches 342, to Date", *Reuters*, June 30, 2008.

[2] S. Staniford-Chen et al.., "GrIDS – A Graph Based Intrusion Detection System for Large Network," *Proceedings of the 19th National Information Systems Security Conference*, 1996.

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

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

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

[7] Noble, C. and Cook, D. *Graph-Based Anomaly Detection*. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 631-636, 2003.

[8] Randazzo, M., Keeney, M., Kowalski, E., Cappelli, D. and Moore, A.. "Insider Threat Study: Illicit Cyber Activity in the Banking and Finance Sector," http://www.cert.org/insider_threat/, 2004.

[9] Kowalski, E., Cappelli, D. and Moore, A.. "Insider Threat Study: Illicit Cyber Activity in the Information Technology and Telecommunications Sector," http://www.cert.org/insider_threat/, 2008.

[10] Kowalski, E., Conway, T., Keverline, S., Williams, M., Cappelli, D. and Moore, A.. "Insider Threat Study: Illicit Cyber Activity in the Government Sector," http://www.cert.org/insider_threat/, 2008.

[11] Chun, A. "An AI framework for the automatic assessment of e-government forms," *AI Magazine*, Volume 29, Spring 2008.

[12] *2006 AFCE Report to the Nation on Occupational Fraud & Abuse*, Association of Certified Fraud Examiners, 2006.

[13] Eberle, W. and Holder, L, "Mining for Structural Anomalies in Graph-Based Data," *International Conference on Data Mining*. June, 2007