Discovering Anomalies to Multiple Normative Patterns in Structural and Numeric Data

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

One of the primary issues with traditional anomaly detection approaches is their inability to handle complex, structural data. One approach to this issue involves the detection of anomalies in data that is represented as a graph. The advantage of graph-based anomaly detection is that the relationships between elements can be analyzed, as opposed to just the data values themselves, for structural oddities in what could be a complex, rich set of information. However, until now, attempts at applying graph-based approaches to anomaly detection have encountered two issues: (1) Numeric values found in the data are not incorporated into the analysis of the structure, which could augment and improve the discovery of anomalies; and (2) The anomalous substructure may not be a deviation of the most prevalent pattern, but deviates from only one of many normative patterns. This paper presents enhancements to existing graph-based anomaly detection techniques that address these two issues and shows experimental results validating the usefulness of these enhancements.

Introduction

Anomaly detection involves the discovery of an unexpected activity or pattern from within normal transactions or data. The ability to discover anomalies is a vital task for a wide range of organizations, such as businesses or national defense agencies, and involves diverse applications, such as fraud detection, intrusion detection, and insider threat detection. Traditionally, methods for discovering anomalies consist of both supervised and unsupervised approaches using techniques such as classification, clustering, nearest neighbors, and statistics [Chandola et al. 2007]. One of the primary issues with these approaches is their inability to handle complex, structural data. One approach to this issue involves the detection of anomalies in data that is represented as a graph.

The advantage of graph-based anomaly detection is that the relationships between elements can be analyzed, as opposed to just the data values themselves, for structural oddities in what could be a complex, rich set of information. However, until now, attempts at applying graph-based approaches to anomaly detection have encountered two issues: (1) Numeric values found in the data are not incorporated into the analysis of the structure, which could augment and improve the discovery of anomalies; and (2) The anomalous substructure may not be a deviation of the most prevalent pattern, but deviates from only one of many normative patterns. This paper presents enhancements to existing graph-based anomaly detection techniques that will address these two concerns.

The existing techniques are based on an approach called Graph-Based Anomaly Detection (GBAD) [Eberle and Holder 2007]. 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 uses 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. However, to date, GBAD treats numeric data no different than string labels and focuses on only one normative pattern when searching for anomalies.

Related Work

Recently there has been an impetus towards analyzing multi-relational data using graph theoretic methods. Not to be confused with the mechanisms for analyzing "spatial" data, graph-based data mining approaches are an attempt at analyzing data that can be represented as a graph (i.e., vertices and edges). Yet, while there has been much written as it pertains to graph-based intrusion detection [Staniford-Chen et al. 1996], very little research has been accomplished in the area of *graph-based anomaly detection*.

In 2003, Noble and Cook used the SUBDUE application to look at the problem of anomaly detection from both the anomalous substructure and anomalous subgraph perspective [Noble and Cook 2003]. They were able to provide measurements of anomalous behavior as it applied to graphs from two different perspectives. *Anomalous substructure* detection dealt with the unusual substructures that were found in an entire graph. In order to distinguish an anomalous substructure from the other substructures, they created a simple measurement whereby the value associated with a substructure indicated a degree of anomaly. They also presented the idea of *anomalous subgraph* detection which dealt with how anomalous a subgraph (i.e., a substructure that is part of a larger graph) was to other subgraphs. The idea was that subgraphs that contained many common substructures were generally less anomalous than subgraphs that contained few common substructures. In addition, they also explored the idea of conditional entropy and data regularity using network intrusion data as well as some artificially created data.

Lin and Chalupsky [Lin and Chalupsky 2003] took the approach of applying what they called rarity measurements to the discovery of unusual links within a graph. The AutoPart system presented a non-parametric approach to finding outliers in graph-based data [Chakrabarti 2004]. Part of this approach was to look for outliers by analyzing how edges that were removed from the overall structure affected the minimum descriptive length (MDL) of the graph [Rissanen 1989]. The idea of entropy was used by Shetty and Adibi [Shetty and Adibi 2005] in their analysis of the famous Enron e-mail data set. Using bipartite graphs, Sun et al. [Sun et al. 2005] presented a model for scoring the normality of nodes as they relate to other nodes. Rattigan and Jensen went after anomalous links using a statistical approach [Rattigan and Jensen 2005].

However, none of these approaches analyze both the structural *and* numeric aspects of a graph representation of data, nor do they search for anomalies that are small deviations from normative patterns.

Graph-Based Anomaly Detection

Definition

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" [Hampton and Levi 2009].

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.

Assumptions

Many of the graph-based anomaly detection approaches up to now have assumed that the data exhibits a power-law distribution [Faloutsos et al. 1999]. The advantage of the approaches presented in this paper is that it does not assume the data consists of a power-law behavior. In fact, no standard distribution model is assumed to exist. All that is required is that the data is *regular*, which in general means that the data is "predictable". While there are many data sets that are not regular in nature, there are also many, such as business processes, that exhibit regular patterns of behavior. After all, that is why companies set up processes in the first place – to establish rules and guidelines for *normal* business activity [Harmon 2007].

In order to address our definition of an anomaly, we make the following assumptions about the data.

Assumption 1: *The majority of a graph consists of a normative pattern, and no more than X% of the normative pattern is altered in the case of an anomaly.*

Since our definition implies that an anomaly constitutes a minor change to the prevalent substructure, we would chose a small percentage (e.g., 10%) to represent the most a substructure would be changed in a fraudulent action.

Assumption 2: *Anomalies consist of one or more modifications, insertions or deletions.*

As was mentioned earlier, there are only three types of changes that can be made to a graph. Therefore, anomalies that consist of structural changes to a graph must consist of one of these types.

Assumption 3: *The normative pattern is connected.*

In the real-world scenarios of business transactions and processes, the entities are typically linked to each other in some way. Certainly, graphs could contain potential anomalies across disconnected substructures, but at this point, we are constraining our research to only connected anomalies.

Approaches

Most 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 (e.g., normal employee transaction versus threatening insider activity).

GBAD (Graph-based Anomaly Detection) [Eberle and Holder 2007] is an *unsupervised* approach, based upon the SUBDUE graph-based knowledge discovery method [Cook and Holder 2000]. Using a greedy beam search and Minimum Description Length (MDL) heuristic [Rissanen 1989], each of the three anomaly detection algorithms in GBAD uses SUBDUE to provide the top 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 all of the possible graph-based anomaly types 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 [Eberle and Holder 2007] for a more detailed description of the actual algorithms.

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 [Noble and Cook 2003], 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. In other words, not only are we providing the anomaly, but we are also presenting the context of that anomaly within the graph.

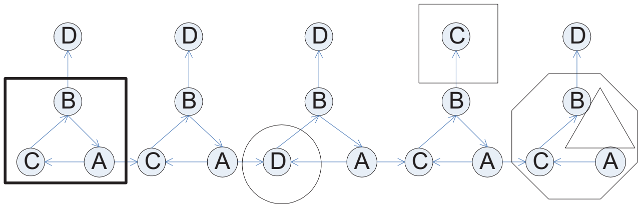

Figure 1. Example graph with normative pattern (bold box) and different types of anomalies (in other shapes).

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 is the resulting anomalous substructure.

Maximum Partial Substructure (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.

Multiple Normative Patterns

One of the issues with this approach is that many data sets, when represented as a graph, consist of multiple 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. For this reason, most telecommunications fraud detection systems use a "profiling" system to distinguish between different customer calling patterns [Cortes and Pregibon 2001]. However, the issue with these sorts of traditional systems is that they are a type of supervised approach because they require a profile of the customer before they can detect anomalies.

The GBAD approach is unsupervised, discovering substructures that are the smallest deviations from the normative pattern (i.e., the substructure that best compresses the graph). However, if we extend GBAD to consider the top N normative substructures, we can then discover other deviations that are potentially more anomalous. This results in the following change to the first step of each of the GBAD algorithms:

> *Find the N normative substructures S_i that have the N smallest values for $DL(S_i)+DL(G|S_i)$.*

where N normative patterns are initially discovered, against which potentially anomalous instances are analyzed.

For example, suppose we have the graph in Figure 2.

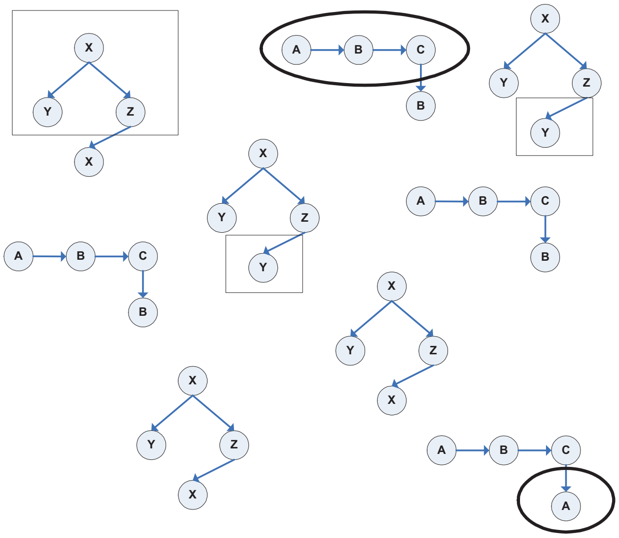

Figure 2. Example of multiple normative patterns.

In Figure 2, the best normative pattern consists of the substructure outlined in the big box. Then, using that normative pattern, GBAD would report the two anomalous substructures shown in the small boxes. However there is another normative pattern which is the second best substructure in the graph, shown outlined with an ellipse (in **bold**). From that normative pattern, a more anomalous substructure is discovered (shown in a smaller ellipse, also in **bold**), as the probability of an extension to an *A* vertex is rarer than the previously reported anomalous extensions (*Y*) associated with the first normative pattern.

In the next section, we will show a real-world example of this scenario and the algorithmic change to GBAD.

Numeric Distribution

While GBAD provides for the *structural* analysis of complex data sets, another one of the issues with this approach is the lack of analysis regarding the numeric values that are present in certain data. GBAD has had success discovering anomalies regarding the *relationships* between data entities [Eberle and Holder 2007], including differences between node and link labels, but sometimes the distances between actual entity *values* needs to be considered. Take for instance the simple example shown in Figure 3.

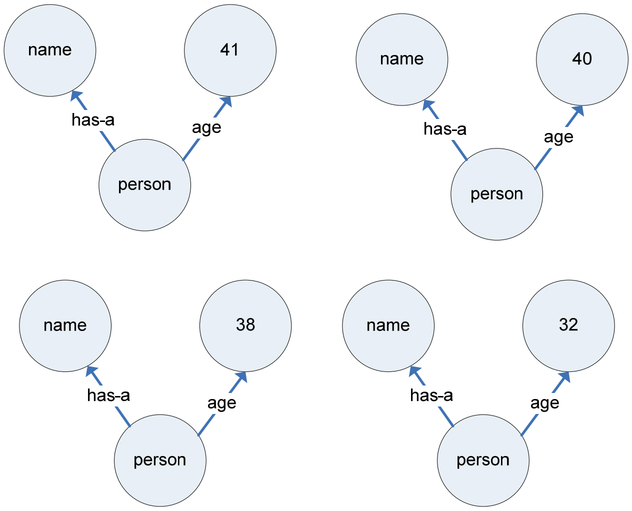

Figure 3. Example of vertices labeled with numeric values.

In Figure 3, each person has a name and an age. Running GBAD on this simple graph results in the reporting of 4 vertices as equally anomalous. While each person has an age, because their ages have different values, they are each viewed as being structurally different.

Currently, GBAD-P calculates the probability of the existence of an edge and/or vertex as:

P(attribute=value) = P(attribute exists)

where *P(attribute exists)* is in terms of the probability that it exists as an extension of the normative pattern. However, when we implement the following change to the GBAD-P algorithm:

*P(attribute=value) = P(attribute=value | attribute exists) * P(attribute exists)*

where the probability of the data is calculated as the probability of the value, given that the attribute even exists, times the probability that it exists. Calculating the mean and standard deviation for all attribute values, we can generate *P(attribute=value | attribute exists)* by using a Gaussian distribution:

$$\rho(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

where σ is the standard deviation and μ is the mean.

Using the same simple example shown in Figure 3, the probability P(x) that each *age* edge exists is 0.25. The mean of the *age* value is 37.75 and the standard deviation is 4.03. When applying this revised probability *P'*, GBAD-P is able to correctly identify that while the structures are the same, with edges labeled "age", the associated vertex with a labeled age of "32", results in the lowest probability, *P'(x),* and thus the greater "anomalousness" (i.e. closer to zero):

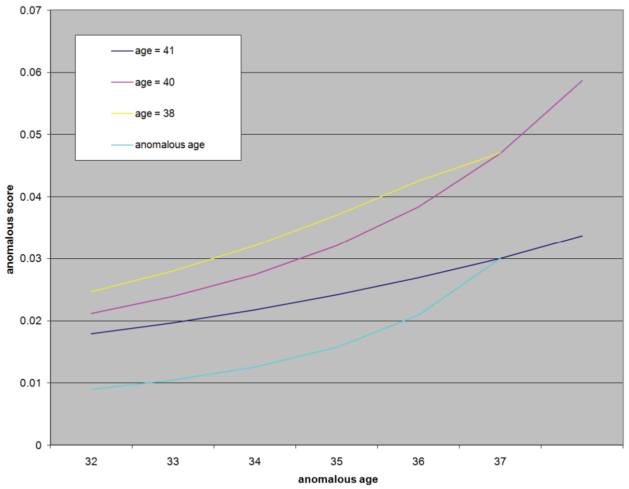

Figure 4. Numeric deviation effecting anomalousness.

P'(age=41) = 0.017876	P'(age=38) = 0.024694
P'(age=40) = 0.021173	P'(age=32) = 0.008946

In addition, further experimentation with using a Gaussian probability metric along with the structural anomalous metric indicates that any numeric value less than one standard deviation results in the anomaly not being reported as anomalous. For example, Figure 4 shows how the anomalousness lessens as the numeric value gets closer to the mean, where eventually the originally anomalous vertex is just as anomalous as another vertex, and is even removed from consideration as the other vertex then becomes more anomalous.

In the next section, we will show a more complex real-world example of this scenario and the algorithmic change to GBAD.

Experimental Results

Business Process

First, we simulated a passport application document processing scenario based upon the process flow depicted in Figure 5.We generated a graph representing the processing of 1,000 passport applications, consisting of approximately 5,000 vertices and 13,000 edges. 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.

For our testing, we used a tool called OMNeT++ [OMNeT] to generate a graph consisting of these two normative patterns, although these patterns were not among the top-ranked most normative substructures. We then had the tool randomly insert 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). Applying the GBAD algorithms to this graph results in the anomalous instance(s) associated with only one of the normative patterns to be discovered. However, when we modify the GBAD-P algorithm (which was the only algorithm to discover an anomalous instance) 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).

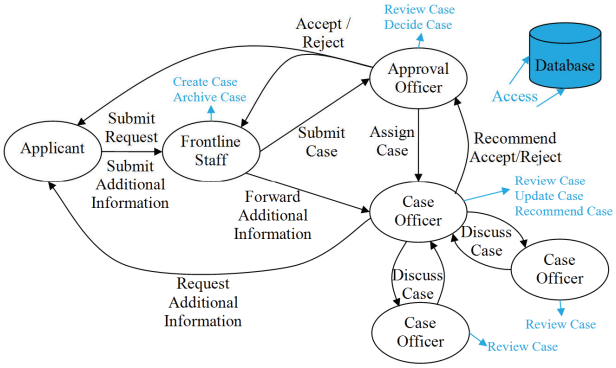

Figure 5. Depiction of application processing.

Financial Transactions

We then created a more complex graph that consists of a bank transactions scenario. In this case, the graph consists of 10 bank accounts, where each account consists of two deposits and two withdrawals. Then one extra deposit was inserted into 3 different accounts, with 2 of the deposits being closer to the mean than the other deposit. The graph consists of vertices labeled "account", "deposit", and "withdrawal", edges labeled "transaction" and "amount", and vertices with dollar values (e.g., "2000.0"), similar to what is shown in Figure 6.

Again, in order to calculate the probability of the normal distribution, first the mean and standard deviation of all of the amount values are calculated. Applying the GBAD-P

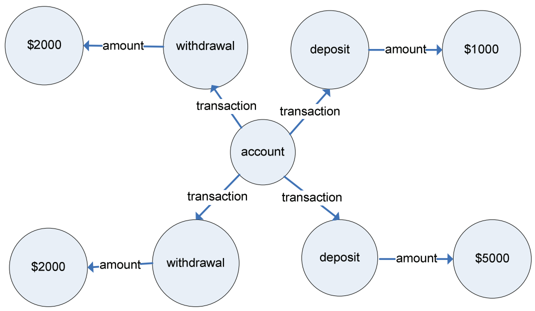

Figure 6. Example of anomalous bank transactions.

algorithm, it first discovers the structural differences inherent in the 3 accounts that contain the extra deposits, then it applies the new Gaussian probability metric to correctly identify the account that contains the deposit with the largest deviation in amount. Also, as was shown in the earlier example, further experimentation with using a Gaussian probability metric on the transaction amount, along with the structural anomalous metric indicates that any value less than one standard deviation results in the anomaly not being reported as anomalous.

What makes this significant from a practical perspective, is that while the value of the anomalous deposit was high ($5000 for this transaction, and $1000 and $2000 for the other two extra deposits), there were actually 11 transactions of this same amount (i.e., out of 43 transactions, over 1/4 of the transactions were at the $5000 level) within this graph. If one were to perform a traditional numerical analysis of this value in terms of all of the deposits (and withdrawals) that were made, the value of $5000 would not have been interesting. However, when combined with the anomaly of the extra structure (i.e., an extra deposit transaction), then it becomes significant.

Conclusions and Future Work

Two of the issues with current graph-based anomaly detection approaches are their inability to use numeric values along with their structural analysis to aide in the discovery of anomalies, and their inability to discover anomalous substructures that are not part of the normative pattern. This paper presents novel graph-based anomaly detection approaches that start to address these two concerns. 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.

Acknowledgement

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

Chakrabarti, D. *AutoPart: Parameter-Free Graph Partitioning and Outlier Detection*. PKDD 2004, 8th European Conference on Principles/Practices of KDD, 112-124, 2004.

Chandola, V., Banerjee, A. and Kumar, V., *Anomaly Detection: A Survey*, Technical Report TR 07-017, University of Minnesota, August 15, 2007.

Chun, A. *An AI framework for the automatic assessment of e-government forms*. AI Magazine, Vol 29, Spring 2008.

Cook, D. and Holder, L., *Graph-based data mining*, IEEE Intelligent Systems 15(2), 32-41, 2000.

Cortes, C. and Pregibon, D. *Signature-based methods for data streams*, Data Mining and Knowledge Discovery 2001, 5(3): 167-182.

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

Faloutsos, M., Faloutsos, P., and Faloutsos, C. *On Power-law Relationships of the Internet Topology*, conference on applications, technologies, architectures, and protocols for computer communications, SIGCOMM, 251-262, 1999.

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.

Harmon, P., *Business Process Change*, Morgan Kaufman, Second Edition, 2007.

Lin, S. and Chalupsky, H. *Unsupervised Link Discovery in Multi-relational Data via Rarity Analysis*. Proc. of 3rd IEEE ICDM Intl. Conf. on Data Mining, 171-178, 2003.

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

OMNeT++. http://www.omnetpp.org/.

Rattigan, M. and Jensen, D. *The case for anomalous link discovery*. ACM SIGKDD Expl. News., 7(2):41--47, 2005.

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

Shetty, J. and Adibi, J. *Discovering Important Nodes through Graph Entropy: The Case of Enron Email Database*. KDD, Proceedings of the 3rd international workshop on link discovery, 74-81, 2005.

Sun, J, Qu, H., Chakrabarti, D. and Faloutsos, C. *Relevance search and anomaly detection in bipartite graphs*. SIGKDD Explorations 7(2), 48-55, 2005.