Detecting Anomalies in Cargo Shipments Using Graph Properties

William Eberle and Lawrence Holder

Department of Computer Science and Engineering
University of Texas at Arlington
Box 19015, Arlington, TX 76019-0015
{eberle,holder}@cse.uta.edu

Abstract. Detecting anomalies in the structural, or relational, component of data is a new and important challenge, especially in security-related domains. In this paper, we analyze the use of variations in *graph properties* to detect structural anomalies in graphs. Based on several pattern recognition approaches proposed on various domains like the internet, telecommunication call records and social networks, we evaluate the use of these methods for the detection of anomalies in data that is structurally represented as a graph. Our study shows benefits of using graph properties in the analysis of real-world graphs, particularly as it pertains to anomalous activity in *cargo shipments*.

1 Introduction

The ability to mine relational data has become important in several domains (e.g., counter-terrorism), and a graph-based representation of this data has proven useful in detecting various relational, structural patterns [1]. Yet, while detecting anomalies in these domains is also important, less work has been done in detecting anomalies in graph-based data. The purpose of this paper is to present some of the existing work, and then to analyze approaches that can help with the discovery of anomalies. Specifically, we will analyze the use of graph properties as a method for uncovering anomalies in data represented as a graph.

Probably nowhere has the idea of using graph properties to analyze data been applied more than in the area of the web, or the topology of the Internet. Broder et al. [2] presented a graph structure of the web using various graph properties such as distance, proportion and connectedness. They were able to use these properties to represent the interplay of web pages, as well as locate a “spammer”. Using several graph metrics, Jaiswal et al. [3] explored the structure of the internet and were able to provide some insights into a structure’s interconnections. Boykin and Roychowdhury [4] also used graph properties in their analysis of social networks. Using e-mail and address books, they calculated the clustering coefficient of the graphical representation of the data, and used that information to distinguish between spam and desired e-mail. What is interesting in these works is their use of a graph’s properties. The fact is, there has been quite a bit of research in the area of *graph properties* as it pertains to pattern recognition and the defining of a graph’s structure.

However, recently, there have been some efforts to use the properties of a graph in other ways. For instance, Xu and Chen [5] used several graph properties to analyze disparate criminal data. By deriving several graph metrics, they were able to uncover criminal subgroups in data. Some of their work was based upon the work of Klerks [6], as he applied various measures to the detection of criminal organizations.

All of these papers presented just a few of the possible graph properties that could be useful as it pertains to *anomaly detection*. The focus of this paper is to use some of these graph properties, in addition to a few others, to see how they might be used in detecting anomalies in a graph. First, we will define what we consider to be an anomaly as it relates to graphs. Then, we will present the various graph properties that will be used in our tests. This will be followed by a description of the data that will be used, and how the data will be manipulated to introduce anomalies. Then, the results of our examination of the graph properties on synthetic data will be presented.

As a final part of this examination, we will present our results as they are applied to cargo shipments. Using data supplied by U.S. Customs and Border Protection (CBP), we will introduce anomalies based upon scenarios of illegal shipments into the U.S. Over 6 million containers arrive via ships into U.S. ports every year [7]. Due to these high volumes, only 2-4 percent of these shipments can be examined, including those associated with terrorist-related manifests. Analysis of the properties of cargo shipments, represented as graph data, will show the value of this approach.

2 Definition

As was mentioned earlier, the goal of this paper is to present techniques for detecting anomalies in data that can be represented structurally as a graph. While the concept of an anomaly can be rather broad, for our purposes, we define a graph anomaly as a *structural inconsistency*. That is, a graph whose structure was different than *expected*.

A graph *G* is composed of vertices (or nodes) *V*, and edges (or links) *E*. Each vertex can be connected, via an edge, to zero (which means it is isolated) or more other vertices. It can even be connected to itself (called a self-edge). Edges can be undirected, meaning the relationship between the connecting nodes goes either way, or it can be directed, which means that the relationship is one-way.

When representing data as graphs, the data can be defined as the set of expected graphs, anomalous graphs and noisy graphs. An example of this is a set of shipping manifests, where an expected shipment, in this example, is a “bill of lading” that has been inspected (and passed) at all ports of entry. Thus, a defined stream of data would consist of the set of graphs, such that a graph is either expected, meaning it consists of these expected shipments; anomalous, which indicates a graph of shipments that were not expected; or noise, which are shipments that should not have been found in the data stream, like shipping manifests that were incorrectly entered.

3 Graph Properties

A graph can exhibit many properties. When one is dealing with a social network or a computer network, the nodes and their links can vary greatly because of the overall relationship that they represent. Yet, whether it is a graph representing a terrorist network, or a graph representing cargo shipments, they all have graph properties. The important hypothesis that will be considered in this paper is that the structural differences between graphs can be measured using *quantitative* measures.

3.1 Simple

While our initial research examined many of the basic graph properties, only a few of them proved to be insightful as to the structure of a graph for anomaly detection purposes: average shortest path length, density and connectedness.

For the average shortest path length *L*, we used the Floyd-Warshall all-pairs algorithm. Using the algorithm presented in [8], we created an adjacency matrix where the shortest path length between two nodes could be determined. After creating the adjacency matrix, the length of the shortest paths between each connected pair was summed and divided by the total number of node pairs. Thus, this measurement will deviate if the path length between vertices changes.

For a measurement of density, we chose to use a definition that is commonly used when defining social networks [9]. In a social network, entities have clear relationships to other entities, and any disruption of that relationship can affect the social makeup of the network, similar to the way the introduction of an anomaly can disrupt the structural relationship of a set of data. This definition of density is defined as the ratio of the number of actual edges *E* to the maximum possible number of edges (*V*V*): $D = |E| / |V|^2$. Obviously, the insertion, modification or removal of data from a graph alters how compact and interrelated the components may be.

For "connectedness", we used a definition that Broder et al. [2] defined in their paper. They defined a strongly-connected component of a graph as the set of nodes such that for any nodes *u* and *v* in the set, there is a path from *u* to *v*. From that, we defined the "connectedness" of a graph as the set P, that contains all pairs *(u,v)* such that there is a path from *u* to *v* in G, where the cardinality of P is divided by the number of possible pairs (*V*V*): $C = |P| / |V|^2$. This property is a good measurement of the established relationships between entities. If there is an expected amount of connections within a graph (and by connections we mean two vertices that are connected either directly or indirectly), then the severing of a relationship, or the addition of a new relationship, will alter the connectedness measurement.

3.2 Complex

For what we are calling *complex* graph properties, we are going to investigate two measurements. First, there are the *eigenvalues* of a graph. Using an adjacency matrix α (like the one we used for calculating the shortest path lengths in the previous section), the entry $\alpha_{ij} = \alpha_{ji} = 1$ indicates there is a link between *i* and *j*. All other entries are 0. The number λ and the vector *v* represent the eigenvalue and eigenvector of $\alpha v = \lambda v$. The result is multiple eigenvalues (one for each of the number of vertices), however, for our purposes, only the maximum eigenvalue *E* will be used. This is due to the fact that many of the eigenvalues are small (approaching zero), and averaging them would not give us an accurate picture. As it is, looking at just the maximum eigenvalue (as will be shown shortly), provides a useful graph property. This same observation of using the maximum eigenvalue was noted by Chung et al. [11] in their study of eigenvalues as it related to graphs.

Another graph property is the *graph clustering coefficient*. In their work on identifying e-mail "spammers", Boykin and Roychowdhury [4] identified the clustering coefficient for the graph to be the average of the clustering coefficients of each vertex:

$$CC = \frac{1}{|V'|}\sum_{i=1}^{|V'|}\frac{2|E|}{k_i(k_i-1)} \qquad (1)$$

where $|V'|$ is the total number of vertices of degree greater than 1, $|E|$ is the number of edges, and k is the degree. While "spamming" is not necessarily an anomaly, it does convey an unwanted set of data.

4 Data for Anomaly Detection

The following sections describe the data that we will use in our testing, and what types of "anomalous" structures will be introduced into the data.

4.1 Data

The following section describes different structural changes to a graph that could constitute an anomaly in the data. Before we introduce known anomalies into cargo data, we feel it is important to be able to test each of the graph properties on different types of changes so that we can analyze their effectiveness. Thus, in order to control the structure of the graphs, and ultimately have graphs conveying different properties, we will first create various synthetic *random* graphs. We will then apply these same structural changes to some actual cargo shipments.

For each of the tests below, we want to make sure that we create (a) enough samples to be statistically valid, and (b) comparison samples that are of the same number of vertices and edges (size). For the latter criterion, we knew that if we vary the number of edges and vertices in the input graphs, the difference in many of the graph properties will be too volatile, making it difficult to compare. So, for each of our tests, we will randomly change the connections, while keeping the same sizes.

Generation of the anomalous graphs will be handled the same way. Anomalous graphs will be grouped by the size of their associated non-anomalous graph, where the size of the anomaly within the graph is based upon the size of the graph. In other words, the smaller the graph, the smaller the anomaly. For obvious reasons, we do not want to bias the results by inserting a large substructure into a small graph.

It should also be noted that not only could the size (or number of vertices) of an anomaly bias our results, but the number of connections (or edges) could also have the same effect. So, to keep the baseline tests on equal footing, we will keep the density of the graphs relatively the same: a ratio of approximately 4 edges for every 3 vertices. This ratio was chosen for two reasons. First, the computational complexity of some of the calculations increases as the number of connections is increased.

Second, the time it takes to generate the random graphs is also adversely affected time-wise as the number of edges is increased.

For those anomalies that are randomly inserted into the structure, we want to convey two ideas. First, our view of an anomaly is one of something that wants to be hidden. In other words, if an anomaly were of a malicious nature, the perpetrator would probably want to alter the structure of the data as little as possible, so as to remain elusive. Second, the anomaly will be of similar structure, but perhaps not perfect. Therefore, each of the inserted anomalies has the same connection strategy to the rest of its associated graph.

In order to further validate our results, and provide real-world usefulness to these experiments, we will also create graphs from actual cargo shipments supplied by the CBP. After constructing a graphical representation of the shipments, we will then introduce anomalies that represent illegal cargo.

4.2 Structural Changes

Anomalies, whether malicious or not, can take many forms when it comes to the structure of the data. Data can be added, removed, or altered in its relationship to its surrounding environment. As we mentioned in the previous section, the structure of the non-anomalous and anomalous graphs will be kept as similar as possible. While we will control the size of the graphs so that we can repeat tests and make statistically valid observations, the structure of the synthetic graphs will be random. In other words, we will specify the number of vertices and edges when we generate the graphs, but we will let our random graph generator determine the order of the edges.

For the synthetic data used in this paper, we are going to randomly interject the following structural changes:

- Adding a substructure (of one or mores edges and vertices)
- Removing a substructure (of one or more edges and vertices)
- Moving one or more edges
- Adding an *isolated* substructure (i.e., *not* connected to the rest of the graph)

These are all the possible structural changes that can be introduced into a graph.

To illustrate these synthetic changes, Fig. 1 shows an example of one of the synthetic graphs with a substructure anomaly (shown in bold), as visualized using AT&T's GraphViz program. For cargo data, the structural changes that will be applied are indicative of the possible patterns that represent illegal shipments (drug smuggling, arms dealing, etc.). Each of the same types of structural changes that are applied on the synthetic data can also be found in the graphs that contain the illegal cargo containers.

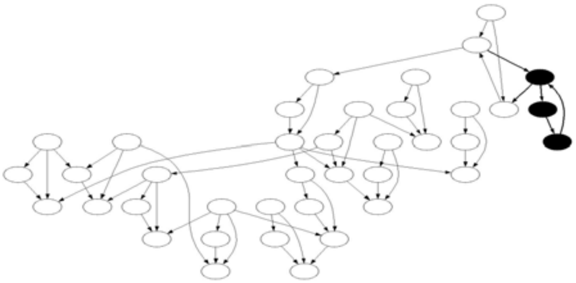

Fig. 1. Example of random graph with inserted anomaly.

5 Synthetic Results

For each of our tests, we created 6 different graph size types consisting of approximately 35, 100, 400, 1000, and 2000 vertices, and another being a dense graph of about 100 vertices and 1000 edges. For each of these increment sizes, we created 30 non-anomalous graphs. We then generated 30 anomalous graphs for each of the four structural changes discussed in the previous section, for each of the graph types.

5.1 Density (D)

For the smaller graphs, the density of the graph lessens when an anomalous substructure is connected to existing vertices in the graph. This makes sense, as the number of actual vertices and edges would increase, while the number of *possible* pairs would increase even more, resulting in a wider deviation, and hence a lower density. This also explains why the density of the graphs that contain the *isolated* substructure is less, due to the fact that they contain unconnected vertices.

The anomalous graphs as a result of the removal of a substructure result in a wide deviation in the density measurement. Since the removal was random, and the graph was randomly generated, it is not surprising that the results were not consistent. However, the average density of the anomalous graphs was not too different from the average density of the non-anomalous graphs. Fig. 2 shows this deviation, while Fig. 3 shows that the average density value does not vary much from the expected density.

For the larger graphs, the density of graphs that had a substructure removed varies even more. At first we can hypothesize that, as was mentioned earlier, the variance in density is just dependent on the randomness of the change. However, the density variance could be attributable to the size of the substructure that was removed. Later we will try varying the size of the anomaly in proportion to the size of the entire graph and analyze the results. The average results for all of the runs are shown in Table 1.

While the properties of the dense graphs (10 to 1 ratio of edges to vertices) with the inserted anomalies show the same behavior, it is interesting that the density of the graphs with the removed substructure does not show the same erratic deviation. What this tells us is that the denser the graph, the less the anomaly of a removed substructure can be observed.

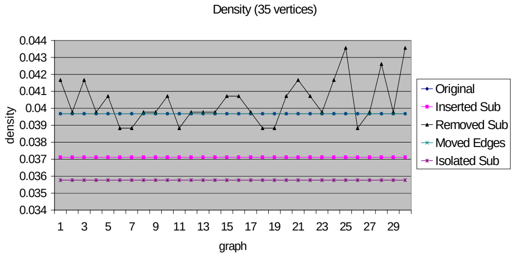

Fig. 2. Density plot of different anomalies.

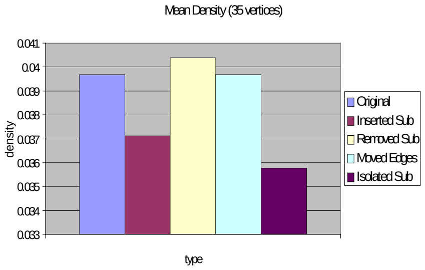

Fig. 3. Mean density of different anomalies.

Table 1. Average density (and standard deviation). (Note: Deviations < .0001 are shown as 0.)

Graph Size (V)	Original	Inserted Sub	Removed Sub	Moved Edges	Isolated Sub
35	0.03968 (0)	0.03711 (0)	0.04037 (.001)	0.03968 (0)	0.03576 (0)
100	0.01306 (0)	0.01282(0)	0.01307 (0)	0.01306 (0)	0.01242 (0)
400	0.00307 (0)	0.00305 (0)	0.00307 (0)	0.00307 (0)	0.00303 (0)
1000	0.00113 (0)	0.00113 (0)	0.00113 (0)	0.00113 (0)	0.00112 (0)
2000	0.00068 (0)	0.00068 (0)	0.00068 (0)	0.00068 (0)	0.00068 (0)
100/1000	0.11324 (.0008)	0.09987 (.0007)	0.11316 (.0008)	0.11324 (.0008)	0.09969 (.0007)

In summary, what we observe is that the graph property of density is visibly affected by the insertion of a substructure, and therefore could be used as a mechanism for detecting those types of anomalies. A possible scenario of such a structural anomaly could be found in a calling network, where terrorists are now calling people that they do not normally call. It is possible that one could discover this anomaly by analyzing the density of the graphical representation of the calls.

5.2 Connectedness (C)

For small graphs, the connectedness of the graphs varies for each of the different types of anomalies. However, while the variation for some anomaly types is not significant, there is some deviation. Similar to density, the insertion and isolation anomalies result in lower values. And, in addition, the insertion of an isolated substructure has an even greater variation on the measurement (see Fig. 4). As we use larger and larger graphs, it appears that a higher deviation is being exhibited. But, when the average of the connectedness values is examined, only the insertion of a substructure (isolated or not) actually has lower connectedness values.

It is also noted that the same behavior is found in the dense graphs. So, not only is the connectedness of a graph not affected by its density, but connectedness also uncovers the same two anomalies (connected and isolated substructure insertions) as the density measurement. The results indicate that density and connectedness could be used in conjunction for the scenario of a terrorist call network.

5.3 Clustering Coefficient (CC)

For the small graphs, the measurements from the isolated anomalous substructures and the anomaly of moved edges are the only ones that show any significant changes. While it makes sense that the insertion of an isolated substructure would affect a graph’s clustering, the variance because of the moved edges is significant due to the way the deviation changes. As the graphs get larger, the distribution still holds. However, as the size of the graphs grows, the coefficient of the graphs with moved edges increases significantly. For instance, as shown in Fig. 5a, when the size of the graph gets to 2000 vertices, the mean clustering coefficient when edges are moved is almost twice as much as the coefficient found in the original graph.

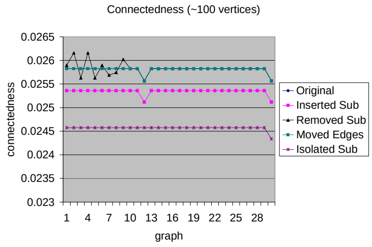

Fig. 4. Connectedness

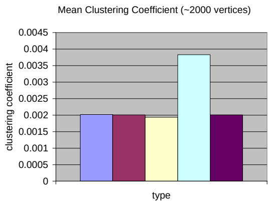

Fig. 5a. Coefficient.

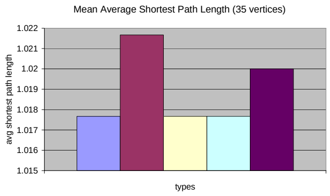

Fig. 5b. Avg shortest path.

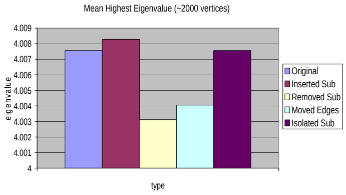

Fig. 5c. Mean highest eigenvalue.

The clustering coefficient is the first of these measurements to be significantly affected by a graph's density. While the ability to detect edges that had been moved is clear in graphs of varying sizes, when the graphs became denser, the indications shift to inserted isolated substructures. The possibility of discovering moved edges is encouraging. Perhaps the clustering coefficient can be used in the analysis of air traffic data, whereby flight patterns are initiated that are not the normal expected routes for particular types of aircraft, which might indicate possible terrorist activity.

5.4 Average Shortest Path Length (L)

Looking at the small graphs, similar to the clustering coefficient, the distribution of values for the average shortest path lengths appears to be similar across each of the types of anomalies. However, unlike the clustering coefficient, the average shows a deviation when an anomalous substructure is inserted, whether it is isolated or not, that is not readily apparent until it is plotted as a histogram (Fig. 5b).

Yet, as the size of the graph increases, the values come together. For our examples of graphs with 400 vertices, the difference in values between the different types is not even noticeable. This trend continues with the rest of our larger tests.

Clearly, this is the weakest of the measurements. While it does show us something on very small graphs (for inserted anomalies), it is perhaps the measurement of the moved-edges anomaly on dense graphs that could be the most significant. This makes us also think of another scenario whereby analysis of a medical network of people that are treated for a disease could help in mitigating a possible outbreak. A disease that is spreading (edges) via seemingly unrelated people (vertices) might show up in the structural change of varying medical reports.

5.5 Highest Eigenvalue (E)

Up until now, we have been unable to find a measurement that aids in the uncovering of substructures that have been removed. However, the average of the highest eigenvalues is different when it comes to the anomalous removal of a substructure. As the graphs get larger (as well as for dense graphs), not only is the difference in anomalous removals significant, but now the moving of edges shows some interesting perturbations. Since the eigenvalues are tied to the vertices and their relationships (edges), and we are looking for the highest eigenvalue, it is interesting that these two anomalies are discovered using this measurement. As shown in Fig. 5c, it appears that the removal of substructures greatly affects the overall structure (as does the moving of edges). If the eigenvalue is a true indicator of a graph's structural make-up, it can indeed be a good measurement for structures that have been altered.

Table 2. Positive measurements.

Size (V)	Inserted Sub	Removed Sub	Moved Edges	Isolated Sub
35	*(D) (C) (L) (E)*	*(E)*	*(CC)*	*(D) (C) (CC) (L)*
100	*(D) (C)*	*(E)*	*(CC)*	*(D) (C) (CC)*
400	*(D) (C)*	*(E)*	*(CC)*	*(D) (C) (CC)*
1000	*(D) (C)*	*(E)*	*(CC)*	*(D) (C) (CC)*
2000	*(D) (C)*	*(E)*	*(CC) (E)*	*(D) (C) (CC)*
100/1000	*(D) (C)*	*(E)*	*(L) (E)*	*(D) (C) (CC)*

The eigenvalue could be useful in the analysis of financial data whereby the methods by which money is handled changes when a fraudulent scheme is taking place, for instance, as part of a money-laundering scheme. Instead of all of the normal steps in a money transaction, certain parts of the “process” are not present. In other words, they are removed from the structural representation of the transactions.

5.6 Summary

Table 2 represents a summary of each of the graph types versus each of the different types of anomalies in terms of what measurements show promising results. It is apparent from this synopsis that there is not one graph property measurement that will work for *all* of the graph types and anomalies (that are presented here). There are some graph property measurements that work for all graph types for a particular anomaly, and even some anomalies can be detected with multiple graph property measurements. But, the question is, can we define a metric by which we can empirically choose what graph properties should be used to detect an anomaly?

If we are looking for anomalies that have been inserted into the data (and not isolated), the graph properties of density *D* and connectedness *C* both appear to be useful. We can say that, $A_1= f(d_1+c_1/d_2+c_2)$, where the anomaly A_1 is a function of the ratio between the expected density d_1 and the expected connectedness c_1 and the actual density d_2 and the actual connectedness c_2. The function *f* returns the positive ratio of the two sets of values, as a positive power function (i.e., the absolute value of the ratio subtracted from 1.0), such that $0 < f(x) \leq 1.0$ where the closer *f(x)* is to 1.0, the more anomalous the graph.

For detecting anomalies where data has been removed, the eigenvalue *E* measurement is the only metric we attempted that is useful. So, our second anomaly measurement is simply: $A_2 = f(e_1/e_2)$, where A_2 is a function of the differential between the expected eigenvalue and the actual eigenvalue.

For the anomalies associated with moved edges, there does not appear to be a clear-cut choice across all graph types. We can use the clustering coefficient *CC* for all but the denser graphs: $A_3 = f(cc_1/cc_2)$,where the anomaly A_3 is a function of the differential between the expected clustering coefficient cc_1 and the actual clustering coefficient cc_2. For dense graphs, we can use the eigenvalue *E* and the average shortest path length *L*: $A_4 = f(e_1 + l_1 / e_2 + l_2)$,where the anomaly A_4 is a function of the differential between the expected eigenvalue e_1 and the expected path length l_1 and the actual eigenvalue e_2 and the actual length l_2.

Finally, while we could use A_1 as a function for finding inserted substructures that are isolated from the rest of the graph, we can actually make a better measurement of anomalousness by adding the clustering coefficient: $A_5 = f(d_1 + c_1 + cc_1 /d_2 + c_2 + cc_2)$,where the anomaly A_5 is a function of the differential between the expected

Table 3. Anomalous scores.

Graph Size	A1 (insertions)	A2 (removal)	A3 (moved)	A5 (insertions – isolated)
35	.0632323825	.0489112512	.1244572353	.0527049953
100	.0181850386	.0021727594	.4077198879	.0512737839
400	.013549854	.003954027	.1833891237	.0145506724
1000	.0065715021	.0033349632	.0021704178	.0077651433
2000	.0036705668	.0010489317	.8998137548	.0047484545
100/1000	.1173940532	.0077862738	**(A4)**.0075303372	.0931024849

density d_1, the expected connectedness c_1 and the expected clustering coefficient cc_1, and the actual density d_2, connectedness c_2 and clustering coefficient cc_2.

To show the usefulness of each of these measurements, Table 3 shows the result of applying the above formulas to the results presented previously in this paper. While some of the values shown are greater than 0 than others, they still indicate the presence of an anomaly, and their deviations can be attributed to the size of the anomaly versus the size of the graph. When we test increasing the size of the anomalies, for each additional anomalous vertex and edge, the result is a score that linearly grows towards 1.0. One of our future goals will be to further refine these measurements to clearly differentiate between different types and sizes of anomalies.

6 Cargo Results

In addition to the randomly generated synthetic data presented above, we also ran the algorithms on actual cargo data. This particular data set consists of cargo shipments that represent imported items from foreign countries to the U.S. In order to keep the size of the graphs similar to the random graphs created earlier, we converted the data into approximately 50 shipments per graph, which translated to about 1100 vertices and 1300. This again provides us with enough samples to be statistically valid, as well as comparison samples that are of the same number of vertices and edges.

The anomalies that we introduced into the cargo data consists of two scenarios. The first anomaly is derived from a press release issued by the U.S. Customs Service where almost a ton of marijuana was seized at a port in Florida [12]. This anomaly represents drug smuggling, whereby the perpetrators attempt to smuggle the contraband into the U.S. without disclosing some financial information about the shipment. Also, an extra port was traversed in-route. In other words, while the shipment looked for the most part like containers of toys, food, and bicycles from Jamaica, there were a couple of structural changes that might not have been noticed otherwise. Fig. 6 shows a graphical representation of a shipment (as a substructure in the entire graph) that contains the anomaly. For space reasons, only this small substructure of the entire graph can be shown. The entire graph contains several of these substructures linked by common nodes, such as ports, carriers, etc.

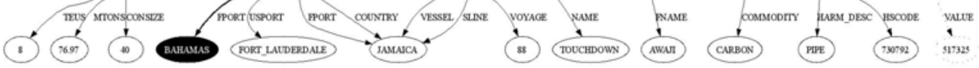

Fig. 6. Graph representation of cargo shipment containing the anomaly, with an insertion in bold and removals represented as dotted lines.

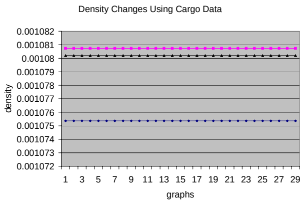

Fig. 7a. Density changes.

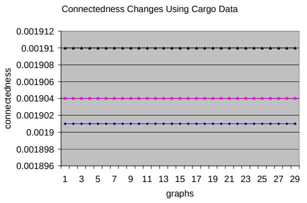

Fig. 7b. Connectedness changes.

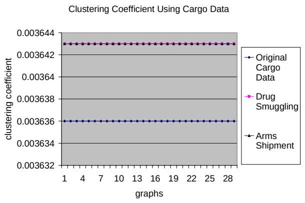

Fig. 7c. Coefficient changes.

The second anomaly is derived from a news story where the CBP became suspicious of a shipment that was heading to El Salvador via Portland [13]. This anomaly is an arms shipment where again, the shippers are attempting to hide the true contents of their containers. Similar to the first anomaly, there are certain manifest information not consistent with other similar (but legal) shipments. In addition, the original port of departure (in this case, China) is removed from the manifest. Again, these are all structural changes in the graph representation of the cargo data.

For both of these anomalies, there are no significant deviations displayed using the average shortest path or eigenvalue metrics. However, there are visible differences for the density, connectedness and clustering coefficient measurements (as shown in Figs. 7a, 7b and 7c). Despite only minor modifications to the expected graph structure (a small substructure removed and another one inserted for the first anomaly, and a few smaller substructures removed along with an edge in the second anomaly), these measurements are able to individually show the anomalies.

In the previous section, we presented combined measurements that provided us with a more comprehensive metric. In this case, A_5 is clearly the desired metric as it represents a measure of density, connectedness and the clustering coefficient – the three that individually show anomalous behavior. Applying A_5 to the cargo data and these anomalies, we get a value of .003 for the smuggling anomaly and .004 for the arms anomaly – similar to what was seen with the synthetic data of 2000 vertices.

7 Future Work

The graph properties presented in this paper are just a handful of the measurements of a graph that we can use. Some of the more interesting ones that were not addressed in this paper are: *rich club connectivity*, *node coreness*, *joint degree distribution*, *average neighbor connectivity*, *entropy*, *power laws* and *interestingness* [14]. Each of these has shown usefulness in the comparative study of Internet topologies [10].

We have presented the graph properties that can be used to discover whether or not a graph might have an anomaly. Of further importance is determining *where* the anomaly exists within the anomalous graph. Since we have theorized that the properties of a graph tell us when something is inconsistent about a graph, this same idea can be applied to the *subgraphs* within a graph. There has been a lot of research done on the art of partitioning graphs [15][16]. In order to support our graph property hypothesis by finding the actual anomaly in a graph, we have to be able to "divide up" a graph into smaller sub-graphs, identify which of these sub-graphs have anomalous graph properties, and further divide them until we are left with the anomaly.

8 Conclusions

A graph-based approach to anomaly detection is an untapped area of research. Work up until now has been limited, with most graph-based approaches focusing on finding patterns and looking at social networks or the web. We feel that this approach of using graph properties can be extremely useful in the analysis of data for anomaly detection purposes. Using just a handful of graph properties, we showed that the differences between what were defined as normal graphs and those that were intentionally altered can be shown to have severe property changes. While the changes seem to vary based upon the type of modification that was performed, they can be used in conjunction with each other to paint a better picture of what is occurring, as was shown in the results from the real-world issue of analyzing cargo containers for illegal, and possibly terrorist-related, shipments.

References

1. Holder, L., Cook, D., Coble, J., and Mukherjee, M.: Graph-Based Relational Learning with Application to Security. Fundamenta Informaticae Special Issue on Mining Graphs, Trees and Sequences, Vol. 66, Number 1-2. (2005) 83-101
2. Broder, A. et al.: Graph Structure in the Web. Computer Networks. Vol. 33. (2000) 309-320
3. Jaiswal, S. et al.: Comparing the Structure of Power-law graphs and the Internet AS Graph. Technical Report 04-30, CS Dept, UMass Amherst, May 2004. (2004)
4. Boykin, P. and Roychowdhury, V.: Leveraging Social Networks to Fight Spam. IEEE Computer, April 2005, Vol. 38, Number 4. (2005) 61-67
5. Xu, J. and Chen, H.: CrimeNet Explorer: A Framework for Criminal Network Knowledge Discovery. ACM Transactions on Information Systems (TOIS). Vol. 23. (2005) 201-226
6. Klerks, P.: The Network Paradigm Applied to Criminal Organizations: Theoretical nitpicking of a relevant doctrine for investigators? Recent developments in the Netherlands. Connections, Vol. 24, Number 3. (2001) 53-65
7. U. S. Senate: Cargo Containers: The Next Terrorist Target? Hearing Before the Committee on Government Affairs. March 20, 2003. U.S. Government Printing Office (2003)
8. Cormen, T. et al.: Introduction to Algorithms. McGraw Hill, 2nd Edition. (2001) 629-634
9. Scott, J.: Social Network Analysis: A Handbook. SAGE Publications, Second Edition. (2000) 72-78
10. Mahadevan, P. et al.: Comparative Analysis of the Internet AS-Level Topologies Extracted from Different Data Sources. (2004)
11. Chung, F., Lu, L., and Vu, V.: Eigenvalues of Random Power Law Graphs. Annals of Combinatorics 7, 2003. (2003) 21-33
12. U.S. Customs Service: 1,754 Pounds of Marijuana Seized in Cargo Container at Port Everglades. November 6, 2000. (http://www.cbp.gov/hot-new/pressrel/2000/1106-01.htm)
13. Mae Dey Newsletter: Customs Seizes Weapons. Vol. 23, Issue 4, August/September (2003)
14. Lin S., and Chalupsky, H.: Using Unsupervised Link Discovery Methods to Find Interesting Facts and Connections in Bibliography Datasets. SIGKDD Explorations. Vol. 5, Number 2. (2003)
15. Hogstedt, K. et al.: Graph Cutting Algorithms for Distributed Applications Partitioning. ACM SIGMETRICS, Vol. 28, Issue 4. (2001) 27-29
16. Lang, K.: Finding good nearly balanced cuts in power law graphs. Yahoo! Research Technical Report, YRL-2004-036. (2004)