ABSTRACT

An important application in network analysis is the detection of anomalous events in a network time series. These events could merely be times of interest in the network timeline or they could be examples of malicious activity or network malfunction. Once a set of events are identified by the anomaly detection algorithm, a more detailed examination of the graph at these times can reveal important details about the behavior of the network. In this paper we use the score decomposition of the global anomaly score of reported anomalies in several dynamic networks to identify the regions of most anomalous behavior and provide interpretations as to the nature of the anomalous events. We also define a new version of the Graph Edit Distance and Clustering Coefficient statistics which are better at finding the local explanations for anomalous behavior.

1. INTRODUCTION

Many interesting real-world networks can be represented by a dynamic network stream which is a sequence of graphs representing network behavior during discrete time windows. A common task is dynamic network anomaly detection which identifies the graphs in the stream which appear to have unusual behavior. Typically an anomaly detection algorithm generates an anomaly score (either with a network statistic or a model likelihood) for each graph and reports those which have an anomaly score above some threshold [2].

However, it is often not easy to determine why the detection algorithm flagged a particular graph. We may want to find the subgraphs with the most unusual behavior and investigate these further to determine the local explanation for the globally detected anomaly. If the local subgraphs represent a community or have some traits in common this can provide a better interpretation of what caused the anomaly. Also this decomposition aids in visualizations of the graph as it allows us to plot critical subgraphs rather than the entirety of the graph.

We propose to use Anomaly Score Decomposition to identify these critical subgraphs. Anomaly score decomposition breaks the total anomaly score for the graph into the part contributed by each subgraph and the subgraphs with the largest score contribution are reported as the local regions most responsible for the anomaly. For statistics which are summations over the edges, nodes, or triplets of the graph this process is trivial: each component of the summation has an associated contribution to the anomaly score sum. Graph Edit Distance, for example, can be represented as a sum across all pairs of nodes in the graph with each pair that experienced changes contributing some amount to the total edit distance.

However, the nature of the statistic can affect how useful the decomposition proves to be. If the anomaly score is evenly distributed over most of the graph (i.e., it is sensitive to global effects rather than local) the decomposition will not identify any region as particularly responsible and we will lack a meaningful interpretation. Common network statistics such as Graph Edit Distance frequently suffers from this: if the total edges of the network increases but behavior otherwise remains the same each node will have a handful of extra edges and the anomaly score distribution will be uniform.

In order to generate more succinct explanations of anomalies we introduce two new network statistics Mass Shift and Triangle Probability which are insensitive to changes in the total edge weight of the network and more likely to find local anomalies when the total edge weight is not constant.

We demonstrate that performing anomaly score decomposition using the two proposed statistics discovers more interesting and meaningful local anomalies in real world network data than commonly used network statistics.

Recent work on anomaly detection in networks has focused more on visualization of global anomalies and/or description of locally identified anomalies rather than on generating local explanations for global anomalies. [5] outlines a method to compress the network in order to improve global-level visualization. [4] identifies subgraph level anomalies in traffic networks and then generates descriptions by mining a secondary source of information (social media text). [1] develops a joint system for summarization, anomaly detection, and visualization. However, the system identifies node-level anomalies, the network is visualized in full, but then the user can explore the network around the locally discovered anomaly. In contrast to previous systems that identify specific node, edge, or subgraph level anomalies, we consider network statistics that characterize more global anomalies and then focus on methods to explain what parts of the network are most responsible for the detection.

2. GRAPH NOTATION AND STATISTICS

Let the network at time $t$ be represented by $G_t = \{V_t, W_t\}$ where $V_t$ is a set of nodes size $|V_t|$ and $W_t = \{w_{ij,t}, w_{ki,t}, \ldots\}$ is a weighted adjacency matrix with total weight $|W_t|$ representing number of communications between nodes during the time window starting at $t$. 

DOI: 10.1145/1235
Graph Edit Distance \cite{DBLP:journals/ssets/PeleD03} is defined as the number of edge/node additions or deletions required to transform a graph into the graph seen in the previous time interval. It is defined as:

\[
GED(G_1, G_2) = |V_1| + |V_2| - 2|V_1 \cap V_2| + \sum_{ij \in V_1 \cup V_2} \text{abs}(w_{ij,1} - w_{ij,2})
\]

and the decomposition for any node pair \( i, j \) is simply

\[
\text{abs}(w_{ij,1} - w_{ij,2}) + 2 - I_{i \in V_1 \cap V_2} - I_{j \in V_1 \cap V_2}.
\]

Barrat clustering \cite{barrat2004efficient} is a weighted clustering coefficient statistic measuring the transitivity of a weighted network. It is defined as:

\[
CB(G_i) = \frac{1}{|V_i|(|V_i| - 1)w_i} \sum_{j,k} w_{ij} + w_{jk} - a_{ij,1}a_{ik,1}a_{jk,1}
\]

where \( s_i = \sum_j w_{ij,1}, k_i = \sum_j a_{ij,1} \) and \( a_{ij} = I[w_{ij} > 0] \).

We also define two statistics of our own, Mass Shift and Triangle Probability, which are analogues of GED and Barrat clustering respectively.

Mass Shift is defined as:

\[
MS(G_1, G_2) = \frac{1}{Z_{G_1}} \sum_{ij \in V_1} \left( \frac{w_{ij,1}}{\overline{w}_1} - \frac{w_{ij,2}}{\overline{w}_2} \right)^2
\]

where \( \overline{w}_1 = \sum_{ij \in V_1} \frac{w_{ij,1}}{|V_1|}, V_1 \cap V_2 \), and \( Z_{V_1} = \binom{|V_1|}{3} \).

The intuition behind the statistic is that by normalizing by the \( \overline{w}_1 \) term we are insensitive to changes in the total edge weight \( |V_1| \).

Unlike GED the Mass Shift scores pairs of nodes based on whether the edge weight between those nodes relative to the expected edge weight has changed. In addition taking the squared difference of node weights amplifies the effect of large changes in edge weight and leads to a less uniform distribution to the anomaly score.

Triangle Probability is defined as:

\[
TP(G_i) = \frac{1}{Z} \sum_{ijk \in V_i} \left( \frac{w_{ijk}}{\overline{w}_1} \right)^3
\]

where \( Z = \binom{|V_i|}{3} \).

Similar to the relationship between GED and Mass Shift, Triangle Probability is a measure of transitivity which incorporates the expected edge weights. By dividing by the expected weight of an edge rather than the total degree of each node, triangle probability is more sensitive to triangles with large edge weights, while Barrat clustering divides by node degree making it sensitive to triangles made up of low degree nodes.

3. LOCAL ANOMALY SCORE DECOMPOSITION

After applying the anomaly detection algorithm and obtaining a set of graphs flagged as anomalous, each graph is broken into node pairs which are ranked according to their anomaly score contribution. The top ranked pairs along with their local neighborhoods are then plotted and investigated for more details. The full procedure is summarized in the table below.

In the score decomposition method, \( K \) is a parameter that controls for the percentage of the total anomaly score explained. In a worst-case scenario where the anomaly score is uniformly distributed \( K \% \) of the node pairs will be reported, while if the distribution is more nuanced a very small subgraph may explain \( K \% \) of the anomaly score. The log score is used as it is easier to select an effective \( K \%)\ threshold in log space. We used a value of \( K = 20\% \) for these experiments.

4. EXPERIMENTAL ANALYSIS

We performed our decomposition on the anomalies found in two real-world datasets. The first dataset is the Enron communication data, a subset of e-mail communications from prominent figures of the Enron corporation (150 individuals, 47088 total messages) with a time step width of one week. The second is a Facebook network subset made up of postings to the walls of students in the 2007-2008 school year (444829 individuals, 4171383 total messages) from a particular university and has a time step width of one day.

Figures 1(a) and 2(b) show the local subgraphs reported by the network statistics in the Enron dataset. Subfigures (a) and (c) show activity in the time step immediately prior to the anomaly while subfigures (b) and (d) show the subgraph during the anomaly. The nodes/edges colored red are the ones identified by the decomposition as contributing significantly to the anomaly, while the ones colored black are the immediate neighborhood. The thickness of each edge represents the total number of communications between each pair of nodes during the time interval represented by the graph.

Figure 1(a), (b) shows an unusually large amount of communication between Senior Vice President Richard Shapiro and Government Relations Executive Jeff Dasovich immediately before the approaches Skillings about resigning as CEO. Having a succinct representation of the anomaly during this time allows for easily picking out the individuals most prominent in the event.

Subfigures (c) and (d), on the other hand, show graph edit distance flagging a huge portion of the network rather than some subgraph. While this does represent an event (the Congressional hearings) there is no interpretation of the event other than that there were many messages being sent at that time.

Figure 2 shows the local anomalies reported by the two transitivity scores triangle probability and Barrat clustering. Subfigures (a) and (b) show the triangular communications occurring between members of the Enron legal department which was occurring during the price-fixing strategy in California.

Barrat clustering identifies the legal department in subfigures (c) and (d) just like triangle probability but does so at a time with relatively low communication. Barrat clustering normalizes by node degree which makes it more likely to report triangles with less weight as long as the nodes of the triangle have few communications outside the triangle.

Figures 3 and 4 show the local subgraphs found in the Facebook dataset. Figure 3 subfigures (a) and (b) show a Mass Shift anomaly that we named “race to 2k posts;” at this time a pair of individuals noticed they were close in on two thousand posts on their walls and decided to reach that goal in one night. The result is a massively
higher amount of communication than was typical between the two in prior time steps.

Graph edit distance, by contrast, identifies no coherent local structure in (c) and (d). It is likely that this event signifies a global increase in communication rather than a change in the distribution of messages. As the additional edges were distributed throughout the network, when looking for subgraphs that generated the most anomaly score the majority of the network has similar scores so a random chunk of the network is reported as the responsible.

Figure 4 show the cumulative anomaly score distributions across all network edges for Mass Shift and Graph Edit Distance during the anomalies they detected in the Facebook network. (a) shows the Mass Shift distribution; here the score is extremely skewed with a handful of edges composing the vast majority of the total anomaly score. This makes selecting the most anomalous local structure extremely simple. (b) shows the Graph Edit Distance distribution; unlike Mass Shift, the distribution is almost uniform across the edges as most edges have a weight only slightly higher or lower than in the previous timestep. Even if the edges with the most anomaly score contribution are selected the resulting subgraph only provides slightly more anomaly score than a random set of edges, preventing the reporting of a succinct local anomaly.

Figure 5 (b) shows the communications occurring during the 2007 Open Championship golf tournament while (a) is the day before. The three individuals with the most communication were arguing about the odds that Tiger Woods would win the tournament. As the Facebook data is composed of wall postings, it is very rare to have significant amounts of communication in a triangular pattern such as this.

Figure 5 (c), (d) is the structure found by Barrat clustering; as before it finds a set of triangular communication with relatively low weights with only a handful of messages going between the members of the triangle. When examining the content of these messages they were found to be small talk and have no associated event or anomaly.

5. CONCLUSIONS

We have proposed an anomaly score decomposition approach to explain how an anomaly can be attributed to the regions of a graph. Learning the identities of the nodes primarily responsible for the
anomaly or the content of their communications can shed light on the exact nature of the anomaly that was detected. However it is important that the statistics used have an anomaly score distribution which is not uniform as otherwise it is impossible to identify the most critical regions of the anomalous network. In this work we proposed the Mass Shift and Triangle Probability statistics which are better at identifying novel explanations for events than Graph Edit Distance or Barrat Clustering.

6. REFERENCES


Figure 2: (a),(b): Subgraph responsible for most of the triangle probability anomaly in the Enron network at the weeks of May 1 and May 8, 2000 respectively. (c),(d): Subgraph responsible for most of the Barrat clustering anomaly in the Enron network at the weeks of November 1 and November 8, 1999 respectively.
Figure 3: (a), (b): Subgraph responsible for most of the mass shift anomaly in the Facebook network at (a) June 1 and (b) June 2, 2007 respectively. (c), (d): Subgraph responsible for most of the graph edit distance anomaly in the Facebook network at October 15 and October 16, 2007 respectively.
Figure 4: (a) Cumulative Mass Shift anomaly score distribution across all network edges in detected Facebook anomaly. (b) Cumulative Graph Edit Distance anomaly score distribution across all network edges in detected Facebook anomaly.

Figure 5: (a), (b): Subgraph responsible for most of the triangle probability anomaly in the Facebook network at July 20 and July 21, 2007 respectively. (c), (d): Subgraph responsible for most of the Barrat clustering anomaly in the Facebook network at May 17 and May 18, 2007 respectively.