3D Anomaly Bar Visualization for Large-scale Network

VAST 2012 Mini Challenge #1

Tao Zhang* Qi Liao† Lei Shi‡

ABSTRACT

In this VAST challenge, log data coming from locations all over the Bank of Money facilities that contains close one million IP addresses. Since the geography of the data plays an important role for potential anomalies detection, we present a particular visualization solution based on Google Earth that can provide measure and deal with geo-spatial data. By mapping three important attributes, i.e., number of connections, policy status and activity flag, into 3D bars on top of physical locations (coordinates), anomaly distribution and trends can be efficiently visualized and analyzed. The general KML file generator can be extended for further analysis on other GIS systems.

1 INTRODUCTION

The VAST mini-challenge one data embodies a whole enterprise network with both network traffic log and geographic information. The main goal for our visualization solution is to create large-scale cyber situation awareness.

From this point of view, two problems become the main impediments to detect operational changes outside of the norm. First, at this time, geographic factors could be not only the phenomenons but also the reasons of anomalies which need to be well-represented for the analytic detection. Therefore, how to represent with the geo-information is the critical part of the task. In addition, big volume of the data could be another challenge for human beings visual analysis. There are nearly 900,000 IP addresses and over 4000 physical locations. What to represent dominates the efficiency of analyzing anomaly in such an enormous network.

We propose a useful method to simplify visual representations by data aggregation and represent them by 3D bars together with position information. The visualization solution is based on an existing geographic information system (GIS), such as Google Earth, as visual platform which has extensive built-in features. We use one of its geometry support – polygon, e.g., 3D bar module as our basic visual item. In the virtual planet of Google Earth, 3D bars' latitude, longitude and altitude corresponds to every item's unique location and the sum value it contains. Furthermore, the timestamp of each bar decides whether it will be visible or hidden during a animation on Google Earth. We designed a particular data grouping method which will be discussed in the following section. We implemented it into a KML file generator by python. In our solution, investigators could make conclusions by clearly observing from both static big pictures and dynamic moving trends.

2 DATA PROCESSING

We use regions as the aggregation granularity. While rendering at finer granularity is useful when focusing on a problem source, using branch, for example, will generate more than 4000 3D bars, which will hardly be displayed and recognized by human beings. Grouping by branch will be an appropriate way to represent more details when studying only one region. By using BMT time as another dimension, we obtain a time series of status distributions, one per 15 minutes.

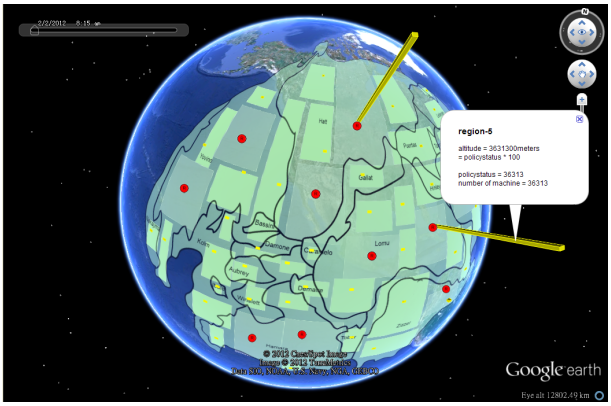

Figure 1: Policy status scenario (at 2/2/2012 8:15am)

We only show one bar per region, corresponding to either the connection, activity or policy sum. The calculations for the three attributes (numConnections, policyStatus, activityFlag) are as follow. A mean numConnection is used by dividing the sum value by the regions machine number. The policyStatus have 5 values which indicate how serious the policy deviating from the machine is undergoing, from 1 (normal) to 5 (very dangerous). In order to emphasize the abnormalities, the policyStatus value is minus by 1 before summed together, so that the normal machines policy (1) will not be counted. The activityFlag attribute have 5 values. Value 1 means working normally, value 2 - 4 mean different abnormal activities on one machine. Since the 2-4 values worth investigating, the value of 1 is counted as 0 and all the other value as 1. After this calculation, the summed value will let us know how many abnormal machines are in the region.

3 3D ANOMALY BAR VISUALIZATION

For every region, we choose its headquarter's latitude and longitude as the bar's position. We define each bar's altitude by the sum value it contains. Some GIS such as Google Earth has various elements for one object such as polygon, linked icon and so on. We choose the 3D bar representation because the extra height dimension makes the spatial data distribution more intuitive than a flat 2D color-coding. Besides, we use distinct color of the bars to distinguish different concerned attributes. Investigators could view single or multi scenarios by clicking the file filter in Google Earth. Animated visualization is used to connecting the dots between timelines.

We leverage a KML file generator written by python, to collect each regions information (location, attributes, time and so on) and generate KML files. These KML files are general enough and can be viewed by Google Earth as well as other GIS systems.

*e-mail: zhang3t@cmich.edu, Central Michigan University.
†e-mail: qi.liao@cmich.edu, Central Michigan University.
‡e-mail: shil@ios.ac.cn, Chinese Academy of Sciences.

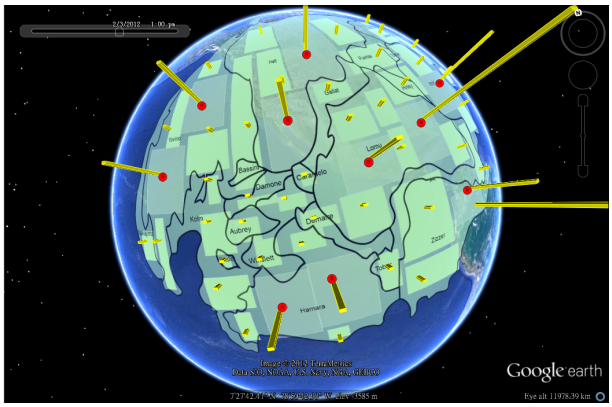

Figure 2: Policy status scenario (at 2/3/2012 1:00pm)

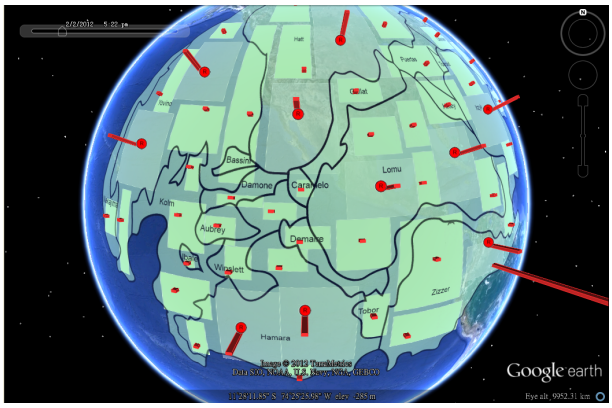

Figure 3: Activity flags scenario (at 2/2/2012 5:22pm)

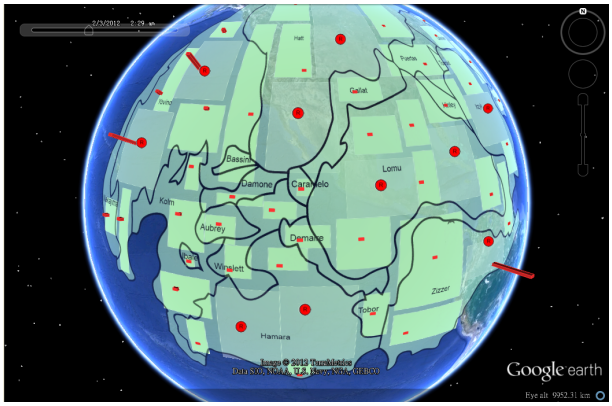

Figure 4: Activity flags scenario (at 2/3/2012 2:29am)

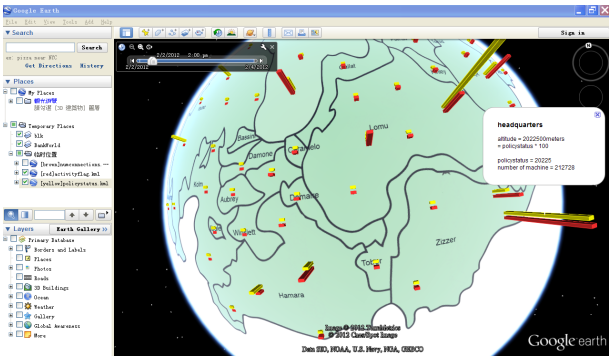

Figure 5: Multi-scenario

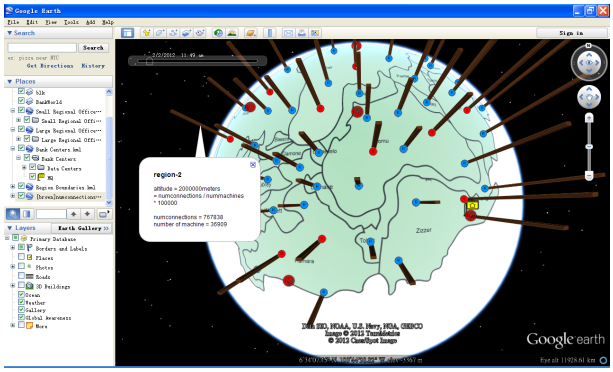

Figure 6: Number of connections scenario

4 Preliminary Results

The section shows snapshots of 3D anomaly bar visualization. As seen from Figure 1 - 5, all machines policy and activity status are aggregated by 50 regions (40 small, 10 large and the headquarter). The yellow 3D bars represent the sum of all machines' policy status in each region, by the bar height, or say, altitude. Each single policy value at a machine is re-scaled by a minus of one, i.e., a value of '1' in the policy status contributes to a zero bar height, larger policy value contributes positive values. The 3D bars in red represent the sum of all machines' activity flags in each region. Each activity flag at a machine with a value of 1 (normal) contributes a zero value, otherwise the activity flags larger than 1 will contribute a 1 value to the summed activity.) In the Figure 6, every brown 3D bar as well as mean numConnection is used by dividing the sum value by the regions machine number.

Through anomaly bar visualization (Figure 1 and 2), the policy status bar's height represents how severe a policy issue a machine has. A higher value might indicate the region is being attacked. The rising patterns of the policy status can be a critical issue for BoM, especially in headquarters. It seems that most policy deviation warnings still exist over time, while the sum of policystatus keeps increasing. But it is only one assumption. Another possibility might be the policyStatuss increase happens in all large regional offices, because the large regional offices are common targets for various intrusion attempts.

In Figure 3 and 4, activityflag bar's height represents the number of the abnormal machines. The high frequency of abnormal situation might because of the higher usage of the machine in the large regional offices. The reason why issues last till 3am each day might because the maintenance will be able to apply on a daily time.

It is obvious to observer from Figure 5 that there are a few policy status bars showing big positive differences compared to other regions. The headquarter's activity flag bar also shows a big plus in height which worth investigating. This region includes one headquarter and 5 data center. We speculate about one of the data center occurs anomaly. A further study in this section will tell in detail.

With the same KML file generator, there are more we can do by simply modifying the aggregation solution. We have already mentioned that grouping by branches will be both accurate and efficient when study one specific region or area. We could also calculate more in detail. For instance, if we only count how many record's activityflag value is '2' in each region. This scenario will show which region's offline machine number is more than other's. The relevant animation might represent a outstanding trend if one area suffered by a geographic-based dynamical anomaly.