Visual Analytics for Situation Awareness of a Large-Scale Network

VAST 2012 Mini Challenge #1 Award: "Effective Video Presentation"

Chris Horn, Chris Ellsworth*

Applied Visions, Inc. – Secure Decisions division

ABSTRACT

Our entry for the VAST 2012 Mini Challenge #1 consisted of a homegrown visual analytic tool that displays the health and policy status of a large-scale network. By putting ourselves into the shoes of an analyst, we defined key dimensions of the data, and developed visualizations for each of them. Using our tool, we were able to identify several areas of concern within the provided data. We presented our findings in a video that depicts the visual analytic interaction sequence of two analysts using the tool to explore the data set. The accompanying audio is set as an informal narrative between the two authors to increase the level of audience engagement.

Keywords: Visual analytics, information visualization, VAST, multiple coordinated views, video presentation

Index Terms: H.1.2 [Models and Principles]: User/Machine Systems—Human information processing; H.5.2 [Information Technology and Systems]: Information Interfaces and Representation—Graphical User Interfaces; H.4.2 [Information Technology and Systems Applications]: Types of Systems—Decision support

1 INTRODUCTION

The 2012 VAST Challenge asked teams to create a situation awareness visualization of a fictitious bank's enterprise network consisting of 888,997 hosts. Specifically, there were two tasks: 1) create a visualization of the health and policy status of the entire enterprise at a point in time and 2) use our visualization tools to look at how the network's status changes over time.

This paper presents the design rationale behind the Secure Decisions team entry, aspects of the tool we built, and some technical details behind our process. Additionally, we discuss our approach to the production of our video submission.

2 DESIGN RATIONALE

We began our analysis by envisioning how network operations staff would use the tool to maintain awareness over their network. Even considering just one data dimension, "policy status" [1], we could envision an operator wishing to identify value patterns in geography, IP-space, and enterprise hierarchy. For example, an operator may ask "Do all hosts with a possible virus infection belong to the same [geospatial region, class B or C subnet, bank region]?".

Our first idea was to create a single visualization that would display multiple data dimensions to reveal such patterns. We experimented with several approaches, including grids, three-dimensional structures projected onto a world map, and tree maps. Through these attempts a design goal was to maintain a static visual position for each host or facility entity over time to allow operators to leverage their spatial memory.

The scale of the network, however, makes such approaches practically infeasible. For example, representing each host directly would only allow one or two pixels per host on a modern monitor. Even aggregating hosts to represent each of the 4,096 facilities would only allow 20x20 pixels per facility. At most, this could encode three data dimensions: x-y position could be approximately mapped to geographic coordinate, and pixel hue and brightness could each theoretically communicate one dimension. But the small area devoted to each entity in such visualizations makes them difficult to use in practice.

Thus, we decided to pursue an approach that employs multiple visualizations, each aimed at primarily one dimension of the data. The tradeoff in such a design is that operators must cycle through multiple "screens" to obtain an overall awareness of operations. The benefit is that each display is simpler and can support more detailed analysis (e.g., filtering and drill-down).

3 TOOL OVERVIEW

Our tool is organized around several top-level visualizations that each displays information focused on one aspect of operations. These aspects include: host downtime, policy status, maintenance activity, and average number of host connections.

One of the visualizations that provided the most insight into the network's health and status over time depicts host downtime. Shown in Figure 1, we used line charts to plot the number (left) and percentage (right) of hosts that are not reporting status over time, aggregated by bank region.[2] These charts are coordinated with two horizontally stacked bars (situated below the line charts) so that when a user brushes over a chart plot area, she can see whether the downtime in the region affects servers, workstations, or ATMs. Drilling down into a region to view facility-level detail is supported by double-clicking directly on a region's plotted

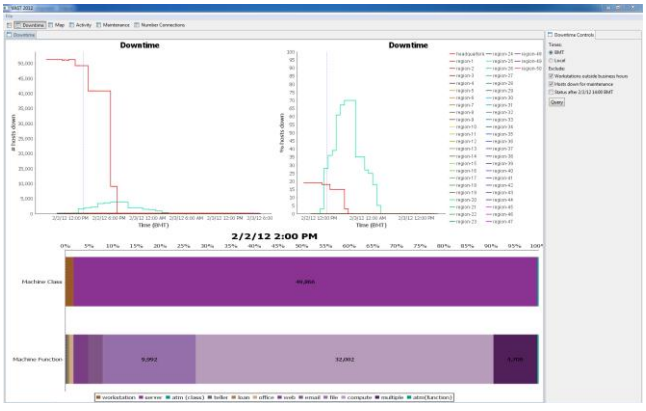

Figure 1. Headquaters and Region-25 unplanned downtime

* {chris.horn, chris.ellsworth}@securedecisions.com

[1] Defined as "How well the machine complies with security policy" and encoded as an integer between 1 and 5, inclusive.

[2] Hosts are housed in facilities that are located in a bank region.

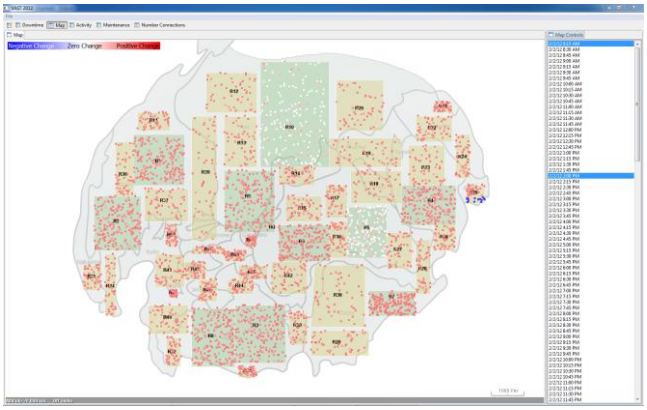

Figure 2. Map depicting change in policy status between two times

series.

Using this view, we were able to detect a downtime event that affected many facilities in Region-25 (cyan), plus the return to operation of an entire datacenter, Datacenter-5 (red).

To help mitigate our less-than-ideal choice of plotting too many series on one chart, we designed a simple algorithm to assign colors that aims to maximize the color distance between an arbitrary number of series. This algorithm works by taking the total number of series and dividing them into groups of five. Each group is assigned a hue (determined by evenly partitioning the hue wheel) that is then used to create five shades, which are assigned to each series in the group.

A second visualization that was particularly useful was a map that could show the average policy status of hosts in each facility, or the change in average policy status of hosts in each facility between two times. Primed with the knowledge that a large downtime event affected many facilities in Region-25, we used the map shown in Figure 2 to better understand the dynamics of the outage.

In Figure 2, a user has chosen to show the change in average policy status by selecting two times in the right pane. Clicking through a sequence of times revealed a southwest to northeast pattern of Region-25 facility outages and then recoveries. The speed and localization of these outages suggested that a weather event swept, knocking out power.

4 Technical Details

4.1 Data Import & Augmentation

We began our investigation by importing the provided comma-separated value data into a Microsoft SQL Server[3] 2008 R2 database and indexing each column to improve access time. The database contained two tables: one containing an entry for each facility and the second with 158+ million host status reports.

Additionally, we constructed derived columns to improve the speed and ease with which we could run more advanced queries. The first derived column we added was facility "timezone", based on longitude. Using that, we then added "localtime" to the host status reports. Having the local time allowed us to easily determine if updates occurred within or outside of business hours.

Because we wanted to know which hosts were "down" at any given time (i.e. not reporting status), we created an additional table that was populated with an entry for each host that did not report a status at any given time. This table contained optional columns representing "excuses" for being down. For example, a workstation down between 6:00 PM and 7:00 AM is not necessarily a problem because employees are encouraged to shut their workstations off outside of business hours. Similarly, a host that has raised an "activity flag" of 2 ("going down for maintenance") would be expected to show up in the downtime table until the next time that it shows up in the status report table.

Similar tables were created to drive the maintenance and number of connections visualizations, although they were aggregated to highlight trends across bank regions.

4.2 Tools

Our tool selections were made to reduce time and acquisition costs. Our goal was to progress as rapidly as possible from a question to a visually-driven answer. In addition to SQL Server, we made use of Eclipse RCP[4] and SWT[5] for our user interface, JFreeChart[6] for plots, NASA's World Wind[7] for geographic mapping, Excel[8] for quickly exploring query results, and whiteboards for ideation.

5 Presentation

We presented our findings as a video screen capture that depicts the visual analytic interaction sequence of using our tool to analyze the data set. Because such videos are often as engaging as that previous sentence, we set the video as an informal narrative between the two authors. To further enhance entertainment and engagement, we each assumed a character role: one as an enthusiastically observant interrogator, and a second as a just-the-facts analyst. The video was organized to simulate the exploration of the data set by these two characters. The general formula for the dialogue is one of call and response: the enthusiastic interrogator calls attention to a salient feature of a visualization, and is answered by the more somber analyst, who explains the relevant underlying data and interpretation of the visualization.

Audio was captured and edited using Audacity[9] and imported into Camtasia Studio[10] where it was synchronized with the video.

6 Conclusion

Using off-the-shelf libraries and some database-driven processing of the raw data, we were able to create a visual analytic tool that we subsequently employed to discover several anomalies and areas of concern in the bank's enterprise network.

7 Acknowledgements

We would like to thank Dr. Anita D'Amico, Eric Gilmore, Brianne O'Brien, and Ken Prole for their constructive criticism and suggestions during the course of our analysis and development of our tool.

[3] http://www.microsoft.com/sqlserver/en/us/default.aspx

[4] http://www.eclipse.org/home/categories/rcp.php

[5] http://www.eclipse.org/swt/

[6] http://www.jfree.org/jfreechart/

[7] http://worldwind.arc.nasa.gov/java/

[8] http://office.microsoft.com/en-us/excel/

[9] http://audacity.sourceforge.net/

[10] http://www.techsmith.com/camtasia.html