

SemanticPrism: a Multi-Aspect View of Large High-Dimensional Data

VAST 2012 Mini Challenge 1 Award: Outstanding Integrated Analysis and Visualization

Victor Yingjie Chen¹, Ahmad M Razip², Sungahn Ko², Cheryl Zhenyu Qian³, David S.Ebert²

¹Computer Graphics Technology ²Electrical and Computer Engineering ³Interaction Design
Purdue University

ABSTRACT

We present a visual analytics system SemanticPrism, which aims to analyze large-scale high-dimensional datasets containing logs of a million computers. SemanticPrism visualizes the data from three different perspectives: geo-temporal, time series curve, and pixel visualization. With each perspective, we use semantic zooming to present more detailed information.

Index Terms: H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces;
H.1.2 [User/Machine Systems]: Visual Analytics;

1 INTRODUCTION

The provided data for VAST 2012 Mini-Challenge 1 has many dimensions: in addition to geographic location and time, it also has fields for activities, policies, and machine types. To make sense of the data, it is important to let an analyst see and compare all these different dimensions. Also, analyzing data for such a big and complex global organization, the analyst should not only be able to analyze the world as a whole, but also narrow down and investigate specific offices and the computers within those offices.

To meet these requirements, we developed the system SemanticPrism to visualize the given data from three aspects: geo-temporal visualizations of office health status, time series curves, and pixel visualizations of IP blocks. All these components are interlinked and provide two to four levels of semantic zooming to allow the user to drill down for more information.

2 THE SYSTEM

The system is developed using Adobe Flash, PHP, and MySQL. It is a web application that analysts can run using most modern web browsers. Also, the client-server structure of the web application is naturally suitable for such a problem by keeping the large-scale data in a central location.

2.1 Data transformation and aggregation

The first challenge is to transform the large amounts of data to make it effective for interactive analysis. Directly querying such a large dataset in its raw form is inefficient and can take hours to get a result. We thus created additional indices and tables to speed up the data query and enable responsive system performance. These new tables categorize branches, offices and computers, and aggregate computers with a combination of criteria (e.g. policy, office, and time). In a real-life implementation, such an aggregation and preprocessing could be performed while collecting data on the fly.

2.2 Geo-temporal visualizations

The default view of SemanticPrism is a geographic visualization

{chen489, mohammea, ko, qianz, ebertd}@purdue.edu

with a time slider (Figure 1). Offices are marked as dots of different shapes to encode the office types. To show the computers' health level, the office dots are colored according to the maximum policy violated by its computers at the chosen time.

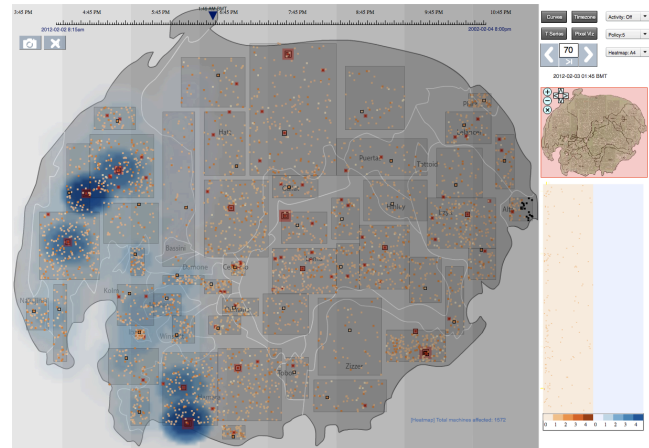


Figure 1: SemanticPrism map view – status at 2012-02-03 1:45am.

A dot with a darker shade of red represents higher policy violation of any computer in the office. With this visualization, if there is even one computer affected by a virus in any office (or other policy violation problems), the user is able to see it immediately. Dragging the time slider automatically updates the status of all offices to the newly selected time.

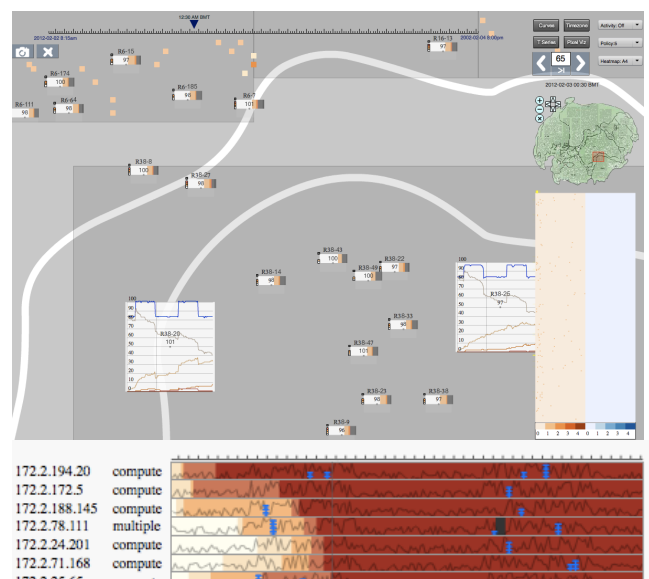


Figure 2: Four levels of semantic zooming to visualize an office

The SemanticPrism map (Figure 1) uses several layers to stack different information together. The heat map layer provides a

visualization of the geospatial distribution of computers with a certain policy or activity. The time zone layer shows the time zones and their local time. A layer of blinking dots highlights offices with a certain policy or activity flag for easy identification among the densely plotted office dots, with the size of the blinking dots reflecting the number of affected computers within that office. With this function, an analyst can easily identify abnormal activities and their growth extent, such as plugging in external devices to computers during nighttime or virus infections.

Semantic zoom allows the user to dynamically drill down and investigate the data in different levels of details (Figure 2). The user may use the navigator (Figure 1, right) to zoom and navigate the map. When zooming in, the space among office dots increase, thereby effectively providing more space to display more detailed information. Depending on this space, an office is visualized in one of four levels: Level 1 visualizes an office as a dot, colored to represent the highest level of policy violation of any computer in the office. Level 2 uses a horizontal color bar to show the percentage of computers in different policies, including being offline. With this visualization, the user may miss policies with small percentages of computers and deduce them as non-existent. To overcome this, icons are used to represent if computers with certain policies exist. Level 3 shows the growth curves of all policies in the office. The curves show the number of computers affected by each policy over time as well as their total. Level 4 shows the history status of each individual computer in the office (Figure 2, bottom). Each computer's policy status over time is visualized as different shades of red in a color bar. The curve in the middle of the color bar shows the number of connections. The computer's activity is visualized as stacked horizontal blue bars with the number of bars representing the activity flag number. The user uses this visualization to drill down to the lowest detail of a specific computer in a specific office.

The user can alternatively interact with the map to show more information. Clicking on a region box will show all offices in that region with level 2 details. The user can then zoom in the region view to show all offices in level 3 and 4 details. Clicking on an office dot will also pop up both the level 3 and 4 detail views.

2.3 Time series curves

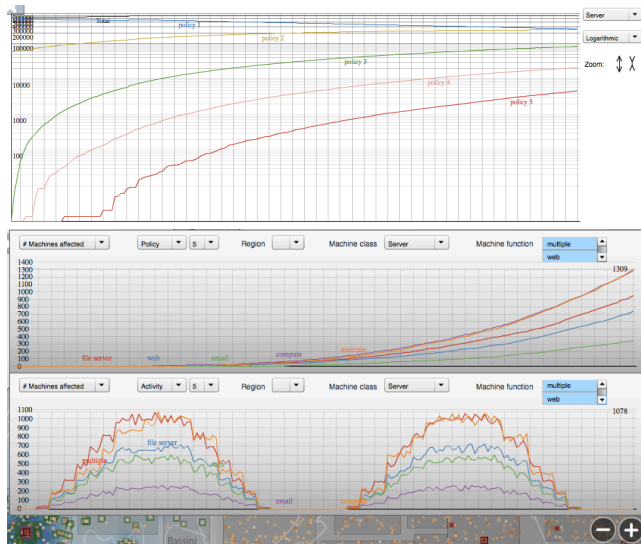


Figure 3: SemanticPrism Time Series Curves.

The SemanticPrism time series curves (Figure 3) provide an overview of the growth trends of policies, activities, and number of connections over the given time period. The default curve view

lets the user see the growth of policies and activities of one class of computers (Figure 3, top). The user can choose to use either linear or logarithmic scale to draw the curve. The logarithmic scale addresses the skewness of the curves towards large numbers and lets the user see the first moment a computer violates a policy, while the linear scale lets the user see the overall growth trend. The time series curve visualization also lets the user dynamically generate new curves and apply a combination of filters for computer class, computer functions, activities, and policies to visualize the number of computers affected or the number of connections. Having multiple panels allows the analyst to perform comparison among the generated curves of different filters.

2.4 Pixel visualizations



Figure 4: Pixel visualizations of IP blocks.

We incorporate a pixel visualization of IP blocks (Figure 4) to visualize the IP address space since it can provide hints to the network structure. Each panel shows the number of computers within an IP block that is affected by the selected activity and policy (Figure 4, top). In each of these five panels, the red (left) side is for policy and blue (right) side is for activity. Each pixel represents a group of computers in a particular IP address D-block. The X-axis encodes the IP's B-block (172.1 to 172.56), and the Y-axis encodes the C-block (0 to 255). The color of the pixel encodes the number of computers that carries the selected policy or activity flags in the D block.

The IP block pixel visualization has three levels of semantic zooming. The user can zoom in to see the time series curve of all C-blocks within one B-block (Figure 4, bottom). Zooming in further will show all individual computers grouped in the IP C-block using a similar format as shown in bottom of Figure 2.

3 CONCLUSION

With SemanticPrism, the interconnected three main components: geo-temporal visualization, time series curves, and pixel visualizations, help the analyst explore the data from different aspects. With semantic zooming, the analyst can explore the full spectrum of the data from getting an overview of the world as a whole, to locating problematic areas, to drilling down further to investigate individual computers.

4 ACKNOWLEDGEMENT

This work was supported in part by the U.S. DHS's VACCINE Center under Award Number 2009-ST-061-CI0001