Dynamic Analysis of Large Datasets with Animated and Correlated Views

VAST 2012 Mini Challenge # Award: Honorable Mention for Good Use of Coordinated Displays

Yong Cao
Virginia Tech

Reese Moore
Virginia Tech

Peng Mi
Virginia Tech

Alex Endert
Virginia Tech

Chris North
Virginia Tech

Randy Marchany
Virginia Tech

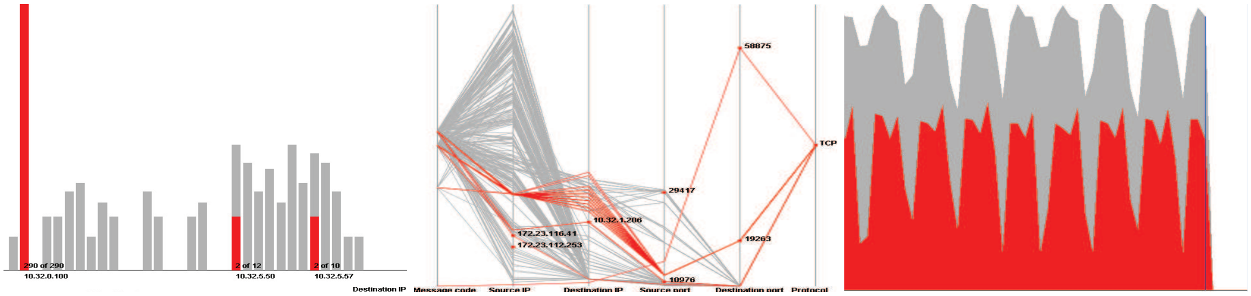

Figure 1: The correlated data views in *AVIST*: histogram view (left), parallel coordinate view (middle), and dynamic view (right).

Abstract

In this paper, we introduce a GPU-accelerated visual analytics tool, *AVIST*. By adopting the in-situ visualization architecture on the GPUs, AVIST supports real-time data analysis and visualization of massive scale datasets, such as VAST 2012 Challenge dataset. The design objective of the tool is to identify temporal patterns from large and complex data. To achieve this goal, we introduce three unique features: automatic animation, disjunctive data filters, and time-synced visualization of multiple datasets.

Index Terms: H.5.2 [Information Interfaces And Presentation]: User Interfaces—Interaction styles (e.g., commands, menus, forms, direct manipulation)

1 Introduction

“Big Data” and dynamic pattern mining have been the most dominated challenges in visual analytics. On one hand, substantial computational power and storage are required to handle massive scale datasets. On the other hand, real-time visualization and interaction are demanded to mine temporal patterns. The temporal information discovery from large datasets is our focus. We introduce a GPU-accelerated visual analytics tool, *Animated Visualization Toolkit (AVIST)*, to tackle such challenges. By taking advantages of the massive parallel computing power of the GPUs, we can analyze and visualize the large datasets, such as VAST 2012 Challenge datasets, in real-time for temporal pattern recognition. In addition, AVIST support multiple correlated data views, which includes histogram view, parallel coordinate view and dynamic view, as shown in Figure 1. The dynamics view shows the count changes of certain filtered events over a period of time. All these three views are interactively correlated to show highlighted and excluded information for intuitive visualization.

Of all the features included in the *AVIST*, we highlight the following three major advantages over the existing tools.

Automated Animation. In order to visually detect temporal patterns in a large dataset, it is critical to render the dynamic changes of the data. By playing the animation of the data changes in certain sampling rate, the hidden temporal patterns can be revealed. In *AVIST*, the animation control supports automatic forward playback, interactively dragging of the timeline bar, and interactive change of time sampling range. With the change the current time and time range, we can define a time window in which the information will be analyzed and visualization on three correlated views. By combining automated animation with three correlated data views, we can clearly visualize the temporal changes in the datasets, therefore discover interesting temporal patterns for further analysis.

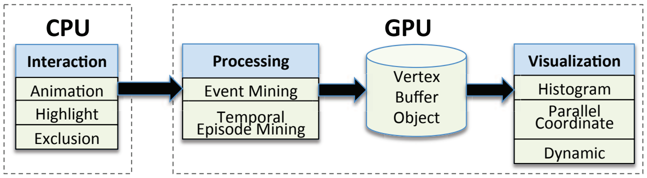

Figure 2: The *In-situ* visualization architecture of the *AVIST*.

Complex Disjunctive Data Filters. For large datasets, visual clutter is one of the most significant problem for knowledge discovery. For example, there can be most than millions of lines in parallel coordinate view for visualizing VAST 2012 Challenge dataset. In AVIST, we include highlight filters to make the user selected values stand out of the rest data, and exclusive filters to remove non-interesting data from the data views. These filters are in disjunctive normal form (DNF) and can be edited with the interactions through the interface in AVIST.

Time-Synced Visualization of Multiple Datasets. For many data visualization scenarios, such as VAST 2012 Mini Challenge #2, there are multiple datasets that are created independently but describe events in the same time period. In this case, it is very helpful to visualize these datasets side by side, and use the same time windows control (current time and time range) to synchronize the visualization of the datasets. AVIST supports the time-synced visualization and independent data filtering for multiple datasets. In VAST 2012 visualization contest, we can visualize and analyze both firewall log dataset and IDS dataset together in AVIST.

2 In-Situ Visualization Architecture

The unique contribution of our AVIST visual analytics tool lies behind the GPU-accelerated *In-Situ* visualization/processing architecture. For most of existing visual analytics tools, the computation of the data analysis/processing is taken at the CPUs, and the results are stored in the main memory before sending to GPUs for visualization. For large dataset visualization and processing, the data communication between CPU and GPU has become the performance bottleneck for real-time interaction. In AVIST, we use an in-situ visualization computing architecture on the GPUs, as shown in Figure 2. The data processing/analysis components of the tool is also executed on the GPU. Therefore, the processing results can be directly used in the data visualization components without additional data transfer from main memory to the GPU. We implement this feature with NVidia's CUDA computing framework and OpenGL's Vertex Buffer Object (VBO) support.

3 Results

We use the AVIST tool and identify three noteworthy events in the VAST 2012 Challenge data, which includes two datasets: Firewall log and Intrusion Detection System (IDS) log. The firewall log includes $23,711,341$ data records and the IDS log includes $35,948$ data records. These two datasets cover the network access information during a three-day period. For the dataset detail, please visit the challenge website.

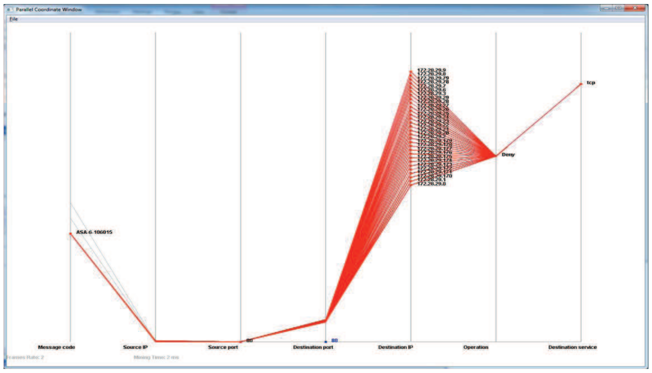

Figure 3: In parallel coordinated view, unknown destination IPs being scanned at port from 1100 to 5000. This is one snap shot at 4/5/2012 19:20:06, covers a period of 530 second. The tools searches 90285 records and uses the filter: *"message code = ASA-6-106015 and Source port = 80 and Destination service = tcp and Operation = deny and Destination IP =* $172.28.X.X$ *and Destination port != 80"*.

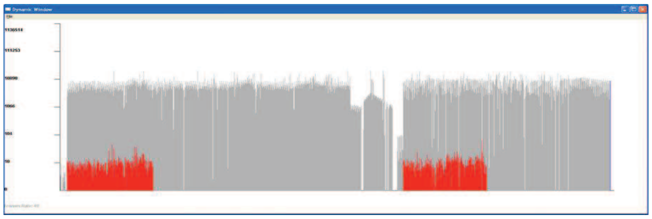

Figure 4: In Dynamic view, it shows the time periods of scans. X axis is time. Gray shows all the activities excluding "*Destination port = 80"*, Red shows the records that satisfied the filter in Figure 3. The red period shows the scanning period.

Finding 1: Port Scanning. In the firewall log dataset, external websites Source IPs $10.32.x.x$ were scanning unknown Destination IPs $172.28.X.X$, which are unknown because they are not part of the bank network diagram. These scans start at 18:20 on each day, and there are four consecutive destination port scans for 2.5 hours, at source port 80, and destination port 1100-5000. Immediately after these scans, the IDS stabilizes. The finding is illustrated in Figure 3 and Figure 4.

Finding 2: Suspicious IRC traffic on Source Port. There is unexpected IRC traffic in the bank. External website IPs $10.32.5.X$ used IRC port 6667 as source port to access destination IP 10.32.0.1. These accesses are all denied, as shown in Figure 5. These accesses started at 4/5/2012 20:27:16.

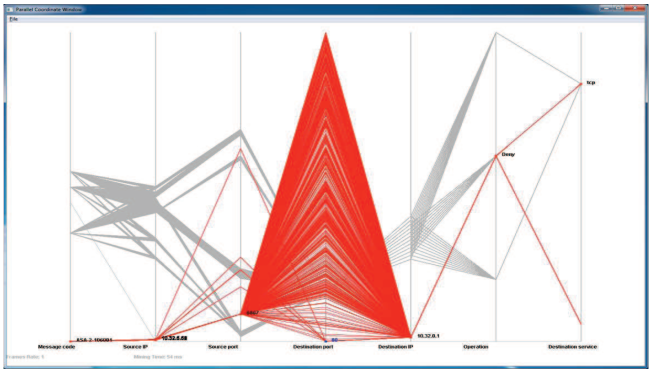

Figure 5: IRC traffic at source port 6667. This is the snap shot of Parallel Coordinate Window at 4/6/2012 15:00:06 during 1000 second period. The highlight filter is *"Source port = 6667"*. These accesses use the random destination ports.

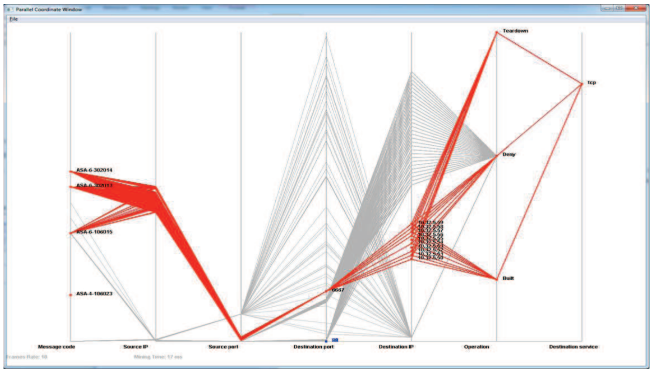

Figure 6: In parallel coordinated view, a set of bank workstations (IP $172.23.X.X$) were accessing external websites $10.32.5.X$ using destination IRC port 6667. This snapshot is taken at 4/6/2012 08:19:36 during 2,000 second period. Totally $334,392$ records are filtered using the highlight query of *"Destination Port = 6667"*.

Finding 3: Suspicious IRC traffic on Destination Port. Workstation Source IPs $172.23.x.x$ in internal network accessing external websites Destination IPs $10.32.5.X$ using destination IRC port 6667, starting at 20:21:06 on April 5th, and continuing for the rest of the period. The traffic pattern is very close to the previous IRC traffic pattern, as shown in Figure 5. These communications scan the source ports from 1100 and up. Source IP 172.23.0.108 starts after the break in data. The result is shown in Figure 6.

The video demonstration of the *AVIST* tool and the detailed explanation of the three noteworthy events can be found at *http://www.cs.vt.edu/∼yongcao/videos/VT-Cao-MC2.mov*.