HIVEBEAT - A Highly Interactive Visualization Environment for Broad-Scale Exploratory Analysis and Tracing

VAST 2012 Mini Challenge 1 Award: Honorable Mention for Comprehensive Visualization Suite

Robert Krüger* Harald Bosch* Steffen Koch* Christoph Müller* Guido Reina* Dennis Thom*

Thomas Ertl*

Institute for Visualization and Interactive Systems (VIS), Universität Stuttgart, Germany
Visualization Research Center Universität Stuttgart (VISUS), Germany

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—GUI; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Search process

1 INTRODUCTION

The VAST Challenge 2012 deals with large data network analysis. The challenge scenario is centered around the computer network of the fictitious Bank of Money (BoM). BoM operates 888,977 computers which are geographically distributed on a fictitious landmass.

The provided data has a hierarchical structure grouping machines in business units, facilities, machine classes and machine types. For each machine, up to 192 status logs in a two day time period were provided. These logs report on three status attributes: policy, activity and number of connections. The main task of Mini-Challenge 1 was to highlight up to five anomalies in the massive data set. We address this challenge by presenting HIVEBEAT, a highly interactive visualization environment for broad-scale exploratory analysis and tracing. It offers eight interactive views which visualize different aspects of the data, support brushing and linking and are complemented with time-dependent and sequence-based filtering. Some of these components build on ideas from earlier works [2, 3, 4]. With HIVEBEAT, we were able to identify five anomalies and justify them with plausible hypotheses.

2 ANALYTICAL PROCESS

In large infrastructure environments, anomalies of machine and network behavior can occur in arbitrary dimensions, scales and variations. Using multi-faceted filter tools and visualization perspectives, our system architecture enables the analyst to cover everything from unusual activity signatures of individual machines up to cross-domain policy violations and global network patterns that indicate large-scale cybersecurity threads. For example, short-timed anomalies located in a specific business unit and only affecting a single type of machine, would be hidden in a huge amount of other
data and noise. Our goal was to provide mechanisms to separate these anomalies from the rest of the data. However, achieving this by exploration of all possible combinations would have been a time-intensive task. We therefore support both overview and detail visualizations of the data (Fig. 1). The analyst may start by getting an overview of the BoM in the map view (Fig. 1 c), selecting a region attracting his interest and take a look at some aspects in the line plots and histogram (Fig. 1 h)), filter again and focus on some machines showing abnormal behavior in the low-level matrix view (Fig. 1 g). HIVEBEAT offers different options to filter the data. To verify hypotheses the user can make use of the data filter (Fig. 1 b) to search for time-dependent patterns in the attribute values. For example one could search for machines for which the policy state deteriorates after a device has been added within a specified time period. Moreover, it is also possible to explore the data by directly selecting or highlighting a subset in a specific view. In both cases the system automatically updates all other visualizations to that selection or highlighting, allowing the user to switch to the one which supports best further analysis at any point. To explain some of them in more detail we will now focus on the process of finding one of the five identified anomalies.

3 DETECTING AN ABNORMAL CONNECTION BEHAVIOUR

One business rule of BoM encourages the staff to turn off workstations at night. However, this rule is only followed for about 60 percent of all workstations and a violation of it is therefore not suspicious by itself. Relevant issues might become visible nevertheless in context with other attributes. In order to discover relevant transitions between single states (e.g. policy) and combined states (e.g. policy/connection) of a large number of machines, we employ an aggregated higher-order state graph visualization (Fig. 2), following an idea of B las et al. [1]. To examine the employees’ workstation behavior, we subsequently select different workstation types and observe their combined policy/connection state transitions. This reveals an anomaly in connection numbers exhibited only by teller machines in certain policy states. To investigate this phenomenon over time, we observe these machines using a parallel coordinates visualization in combination with a time slider (Fig. 1 d, a). A group of 9 teller machines with a disproportionately high number of connections stands out during the first night. In the following night at the same time, already 893 machines, including the ones of the first night, show the same behavior. By highlighting affected machines we detect that they are all associated with the same business unit and by looking at the map (Fig. 1 c), we localize them in Region-10. To further inspect this behavior we employ a low-level matrix view, which shows details about every selected machine and time step. It maps policy to fill-color, activity to border-color, and number of connections to the size of each glyph (Fig. 1 g). Since it is not possible to present all machines in the display, the user has to scroll through the data. This makes it difficult to find similar machines once a suspicious one has been found. As a solution to this problem, the system supports unsupervised learning. The user can cluster the machines with respect to any of their attributes. Furthermore, the distance measure can be changed depending on the current question. Applying k-means clustering with a root mean square distance over the sequences, all machines showing the above-mentioned abnormal connection patterns are grouped together (Fig. 3). Thus it can be seen that the pattern occurs exactly from 2:00 am to 5:00 am (local time) in all affected machines. As the inspection on the detailed level does not reveal any specific event that explains the anomaly, we suggest Region-10 staff to investigate whether the increase in connections was caused by BoM infrastructure or not. In the latter case, the machines’ suspicious, concerted behavior might have been caused by a bot net. And, even worse, this would not have been detected as policy deviations, which might require updating virus scanners or policy definition.

4 CONCLUSION

We presented HIVEBEAT, an interactive system to visually analyze the BoM data set. By applying exploratory analysis starting with an overall visualization of the data and examining those further in lower-level views we have been able to find five hidden anomalies in the large data set. Future work could focus on providing a more generic way to analyze sequential data within our tool.

ACKNOWLEDGEMENTS

We would like to thank the IEEE VAST Challenge 2012 team. This work was partially funded by the German Federal Ministry for Education and Research (BMBF), the German Science Foundation (DFG), the European Commission as part of the VASA project, the SPP 1335, the FP7-project PESCaDO (FP7-248594) and the doctorate program Digital Media in cooperation with the Stuttgart Media University (HdM).

REFERENCES