Visual Analytics for Network Security

Georgiy Shurkhovetskyy* Modern Sciences and Arts University
Ahmed Bahey† Nile University
Mohammad Ghoniem‡ Modern Sciences and Arts University

ABSTRACT

To visualize the VAST 2012 Mini Challenge 2 datasets, we use the InfoVis Toolkit (IVTK). Custom visualizations as well as extra interaction capabilities have been added to the toolkit. Custom-made Python scripts are used for data preprocessing purposes. In this work, we show how visualization tools may be combined to leverage network forensic analysis tasks.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Interaction Systems

1 INTRODUCTION

VAST 2012 Mini Challenge 2 aims to provide the security analysts of a regional office in "the Bank of Money" with visual analytics software promoting fast and informed decision making. The organization's network is crippled with rogue security software and suspicious computer activity that requires a forensic investigation. As a starting point, available data consist in firewall and IDS log files for two days, along with documentation describing the network topology, the priority of each IP address as well as the corporate usage policy of the information services.

In order to help security analysts achieve cyber situation awareness, we combine multiple interactive visualizations providing both global and detailed views of network activity. These visualizations are designed or customized to increase the capabilities of security analysts, and help them detect noteworthy events as well as temporal and topological patterns in the network activity throughout the observation period.

2 MATERIALS AND METHODS

2.1 Data Preprocessing

IVTK [2] can read many formats including XML and CSV files. Data are loaded into internal data structures such as tables, trees and graphs. For the purpose of this challenge, we use custom Python scripts to convert log files into CSV files with two objectives in mind. Firstly, the scripts are used to select fields required for each type of visualization and merging log file entries from the first day with entries from the second day. Secondly, for particular tasks, scripts are used to aggregate log entries at a half-hour time granularity. For example, we count outgoing connections from each unique source IP in every half hour interval throughout the two days. We also aggregate on tuple values such as the triplet (source IP, destination IP, destination port) among other possible combinations. The half-hour time granularity proved to be sufficient for observing activity trends in the network, and may be customized to meet specific analyst needs.

2.2 Visual Representations

Processed data are then loaded in IVTK, and displayed using various types of visualizations. Firstly, for an overview of network activity, outbound communications from the firewall log are represented as a heatmap, using a graph matrix visualization. The rows being half-hour time slices, while the columns are source IP addresses. The darker the color the more intense the activity. This helps the analyst see topological activity patterns (with regard to whole IP ranges) as vertical stripes and temporal patterns as horizontal stripes and get insight about the propagation of activity patterns over time as in figure 1. Several such patterns exist in the dataset. Similar heatmaps can be generated for source/destination IP and incoming/outgoing traffic on source/destination ports providing multiple points of view on network activity and allowing the analysts to see patterns and spot outliers.

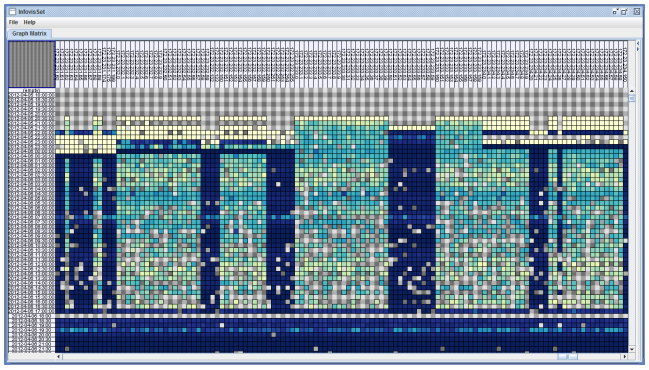

Figure 1: Heatmap showing IP addresses involved in IRC traffic

IVTK provides built-in color equalization and sorting capabilities. We have extended it with custom segment-by-segment sorting for IP addresses so that IP ranges would be displayed in a consolidated and orderly fashion.

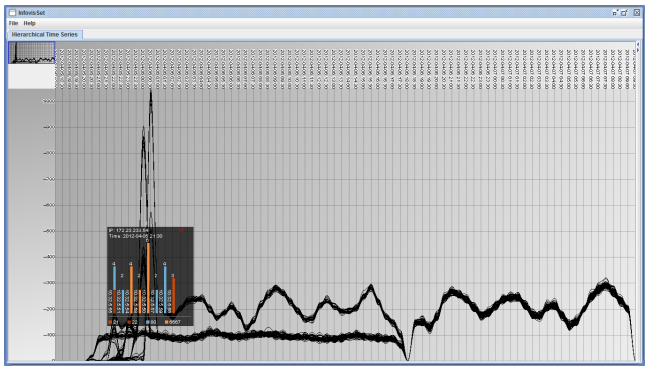

Figure 2: Timeseries visualization of IP addresses involved in FTP / SSH / IRC traffic

Using the same data as in the heatmap visualization, aggregated time series visualization display information about all network traffic. The use of aggregation as explained earlier is motivated by the fact that a computer is the most simple network element, whose malfunction reflects on its overall behavior. With aggregated Time Series, workstations with outstanding volumes of network traffic stick out clearly. Rich interaction and filtering capabilities let the

*e-mail: shurkhovetsky@gmail.com
†e-mail: ahmed.bahey@nileu.edu.eg
‡e-mail: mghoniem@msa.eun.eg

IEEE Conference on Visual Analytics Science and Technology 2012
October 14 - 19, Seattle, WA, USA
978-1-4673-4753-2/12/$31.00 ©2012 IEEE

user drill down into details and focus on salient events. The analyst can use regular expressions to filter the time series view using IP ranges as spotted in the heatmap view, in order to focus on the corresponding curves and hence reduce the clutter inherent to time series. Interesting temporal patterns may emerge as in figure 2. The analyst can also enable scale fitting to adapt the visualization to the filtered subset of data. A custom magic lens [1] displays a stacked histogram allowing the analyst to drill down into individual port activity broken down for each pair of source IP and destination IP during the selected half-hour slice. The bar next to each IP tells which ports are used and how many hits they receive. A colored port legend is displayed below the charts for the analyst's reference. Like heatmaps, we use time series to visualize various aspects of the firewall and IDS log files.

For a view of the different events and associated actions from the IDS log, a parallel-coordinates visualization [3] is used. This visualization helps detect different suspicious events in the network. For example, by color coding suspicious ports (such as FTP, SSH and IRC ports) and/or filtering on specific IDS classes of events. For example, in figure 3 one can see traffic involved in "attempted information leak" through ports 22/SSH and 161/SNMP. In IVTK, the parallel coordinates visualization comes with the ability to order coordinates, color and filter data in relation to data attributes. It also provides excentric labeling to avoid label overplotting.

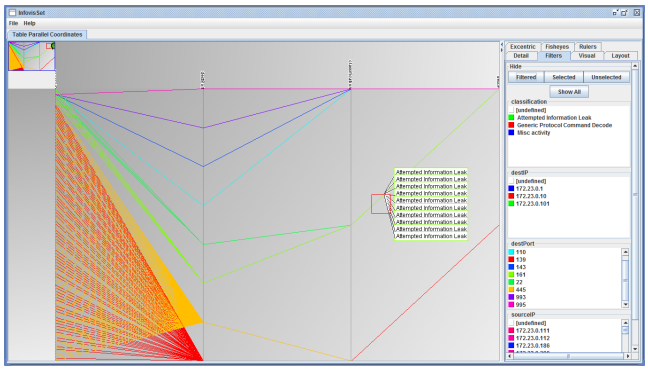

Figure 3: Parallel Coordinates visualization of IDS events

3 Analytical Process

Network activity analysis has to take into account internal limitations and official usage policy. A security analyst would start from an overview of network activity, examining time dependent patterns across all internal IP ranges. She would also look for inbound and outbound traffic and monitor sensitive ports such as "service ports", looking for the visual signature of common threats and anomalous workstation activity. After investigating the data certain violations of the policy or implementation loopholes may be detected. Concretely, the heatmap visualization is examined for patterns, since they can be easily observed on this view as vertical and horizontal stripes. Determining which IP addresses contribute to the pattern triggers further scrutiny in a more precise visualization, such as a time series visualization.

Following this approach, we discovered traffic through forbidden ports due to asymmetric firewall rules e.g. outbound IRC traffic is allowed on port 6667 while inbound traffic is blocked. We also spotted traffic through ports that fulfil identical/similar functions to blocked ports. We resorted to Internet search in order to find out more about ports and associated services.

For example, one noteworthy temporal pattern concerned the IP range (172.23.123.x – 136.x) which had a high volume traffic with recurring temporal pattern through both days. Switching to the aggregated time series visualization and filtering on that IP range, the magic lens reveals that this traffic is directed to the same range of external websites (10.32.5.x) and on the same port (6667). Port 6667 is used for IRC (Internet Relay Chat) messaging, while the "Bank of Money" policy restricts the use of computers to business purposes. Thus this traffic may in fact be considered suspicious and needs further investigation from the network administrators.

Another noteworthy event was detected using the parallel coordinates visualization of data from the IDS log file. Inspecting this visualization, one can easily notice that communication on port 21/FTP is denied by the organizations policy, but the firewall logged successful port 22/SSH traffic. Knowing that port 22 can also be used for file transfer and remote command execution over a secure connection, this traffic can also be considered suspicious and should be investigated more by the network administrators, especially when tied in with IRC traffic from the same computers.

The magic lens can be used to characterize multiple types of security threats. For instance, when the data is aggregated on destination IP and the lens displays the source IP addresses that connect to it, if the chart uses a large variety of colors as in figure 4, this indicates that many connections are attempted through many different ports. The analyst will have to decide whether a port scan was detected in this instance or whether this activity pattern is normal or legit. Likewise a distributed denial of service (DDoS) attack can be visually characterized by a suspiciously high number of connections to one destination IP inside the network from multiple sources which follow similar temporal activity patterns. All such information are displayed by our tools and it is up to the security analyst to distinguish between normal and malicious behavior.

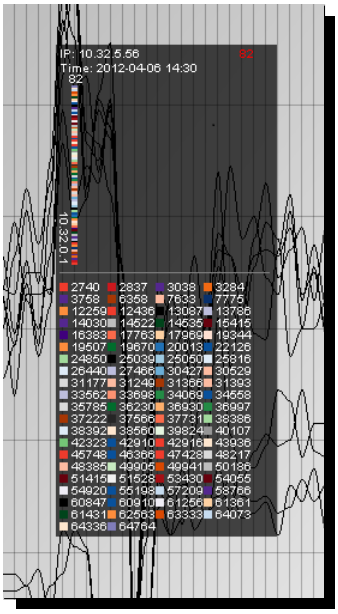

Figure 4: Portscan on the firewall

4 Conclusion

The use of visualization tools in network security helps increase situational awareness, and assists the administrators with their investigative tasks. In this work, we customize IVTK, a rich opensource visualization toolkit, to meet the VAST 2012 challenge. Interactive visual analytics tools allow security analysts to examine network activity and detect abnormal events with ease. Linking expert knowledge with insights gained from the visualization tools lets the analysts make well informed decisions about salient events. Further improvements include stronger integration between views and the use of a wider set of visualizations. While the methods presented proved useful for post-mortem analyses, we believe that with slight changes they may also be successful at monitoring live data.

References

[1] E. A. Bier, M. C. Stone, K. Pier, K. Fishkin, T. Baudel, M. Conway, W. Buxton, and T. DeRose. Toolglass and magic lenses: the see-through interface. In *Conference companion on Human factors in computing systems*, CHI '94, pages 445–446, NY, USA, 1994. ACM.

[2] J.-D. Fekete. The infovis toolkit. In *Proceedings of the IEEE Symposium on Information Visualization*, INFOVIS '04, pages 167–174, Washington, DC, USA, 2004. IEEE Computer Society.

[3] A. Inselberg and B. Dimsdale. Parallel coordinates: A tool for visualizing multi-dimensional geometry. In *IEEE Visualization*, pages 361–378, 1990.