Galaxy: Link Space Visualization and Analysis of Network Traffic

Elisha Peterson, Ryan Mukherjee, Saurabh Vyas, Duane Cornish*

Organization or School

ABSTRACT

Galaxy is a heterogeneous event visualization tool that couples a replayable timeline of events with a unique link graph view that leverages the concept of "link space" to show significant detail about events in addition to connectivity. Galaxy allows the analyst to import multiple data sources as "channels" in a single workspace. Using the replay feature and a variety of filtering mechanisms, the analyst can explore the data to discover patterns of activity or behaviors that play out over time. Additional tools help the analyst piece together a story for the desired audience. For the VAST Challenge 2013 MC3 data, we added a query capability to Galaxy to pull data from a MySQL database with netflow, firewall, and host data, and also imported relevant network information into the tool. We conclude this note with an example of one sequence of suspicious activity.

Keywords: Galaxy, event visualization, network visualization, link graph, netflow, firewall, link space.

Index Terms:

1 INTRODUCTION

Galaxy is an event visualization tool developed by JHU-APL to help analysts find meaning in large, heterogeneous collections of events. It shows not only who talked to whom, but also when they talked and what they talked about. For the VAST Challenge data sets, Galaxy helps the analyst distinguish between different kinds of network and firewall traffic, as determined by port number, and also determine the sequence of traffic over time.

2 GALAXY

Galaxy is a visualization tool designed for event analysis. Its principle view is a link graph view that is synchronized with a timeline of events. Supporting views include a trends view, a filter view, a text query/table results view, and an IP-centric view.

Analysts use the tool by querying or ingesting data sources, as *channels* in a workspace. The timeline view presents each channel separately, as in Figure 1. This helps to correlate events from different sources. The timeline is also used to filter the link graph view for a specific time window.

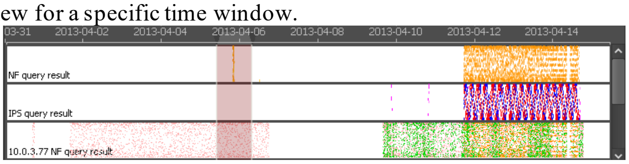

Figure 1: Galaxy's timeline view displaying multiple data channels. Colors indicate different traffic types.

A fundamental assumption in Galaxy is that events have different "flavors," and the tool colors these flavors differently. In the case of netflow data, the flavors are server ports, which usually map to particular protocols or services being used. These colors are used throughout the tool, so that events of each flavor can be distinguished in every view.

2.1 Link Space

The most unique feature in Galaxy is the use of *link space* in the graph view. The basic idea of link space is to support mapping information about a link onto a two-dimensional space between nodes (see Figure 2). In theory, this allows any two-dimensional visualization to exist on a link, and can therefore increase the expressivity (number of dimensions of data) of the view. In practice, the use of these two dimensions must be balanced against the readability of the result.

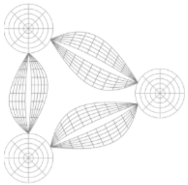

Figure 2: *Node Space* and *Link Space* in a 3-node network, representing the "canvas" on which node and link information can be displayed.

A parallel concept is *node space*, a two-dimensional space at the nodes, which is leveraged in many existing visualization techniques (see [1], [2]).

By default, Galaxy uses link space to present information about the relative volumes of different flavors of events, as well as the time at which the events occurred as shown in Figure 3, which visualizes 1200 ftp netflow records. One or two arcs exist between each pair of edges, indicating directionality; bi-directional events represent exchanges of information. Colors distinguish two common FTP ports. The gradient along the edges represents the relative position of the events with respect to the selected time period (not shown). The gradient is most helpful in identifying highly regular communications, but also provides information about the volume and/or order of events. Finally, the glyphs at the nodes display relative percentages of traffic at an individual node, and the node colors are used to distinguish internal vs. external traffic.

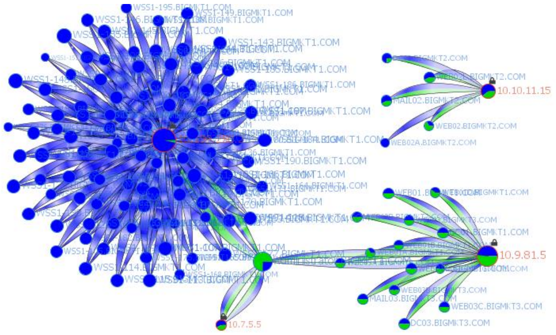

Figure 3: Depiction of 1200 events divided two "flavors" of traffic (both FTP) in Galaxy using link space.

* elisha.peterson@jhuapl.edu, ryan.mukherjee@jhuap.edu, saurabh.vyas@jhuapl.edu, duane.cornish@jhuapl.edu

In total, the view depicts up to six dimensions of each event: source, destination, time, flavor, volume, directionality. It may also add additional information about the source and destination. It does this with minimal clutter to the view, allowing for a much richer setting for discovering anomalies and telling stories.

3 Approach

The VAST Challenge MC3 data sets included two weeks of netflow and firewall traffic data for a hypothetical "Big Market" corporation, as well as Big Brother log data for individual hosts. "Enrichment data" such as the mapping between specific IP's and hostnames or roles within the enterprise was also provided.

Given the size of the data (~100 million events), we ingested the data sets into a MySQL database, and indexed the resulting tables on IP (or hostname) and port.

Initial analysis for basic trends was done using a tool called DRAFT (Data Reading and Fingerprinting Tool), which automatically generated summaries of the data broken down by key features such as IP, port, and/or time. This high-level analysis indicated several significant anomalies such as unusually noisy IP addresses or ports. These key indicators were noted as potential starting points for further analysis.

The primary analysis tool was Galaxy, which was customized in a few ways to support the MC3 data. First, we built a custom SQL query capability to load data by IP/host, port, and/or time window. We also built a Galaxy plugin to query and display the host-based information provided by the Big Brother logs. We also imported the enrichment data into Galaxy (a built-in feature). Events were set to be colored by common ports (for netflow), and for firewall rule.

Our analysis was accomplished through iterative exploration of the data, both beginning with specific ports likely to be interesting (such as those commonly used for IP's), and IP's likely to be interesting, whether due to the network map or the output of DRAFT. During analysis, additional queries were made as necessary to load additional channels into Galaxy.

4 Sample Analysis

One suspicious address uncovered during analysis was the external IP 10.0.3.77. Figures 4-5 show differences in activity centered at 10.0.3.77 and the internal machine wss2-23.bigmkt2.com during the two weeks of data. In the first week, 10.0.3.77 communicates over port 25/smtp (pink) with the mail server. The workstation wss2-23.bigmkt2.com has http, ntp, and some ssdp/upnp traffic, none of which is particularly suspicious by itself. The yellow/orange arcs show unrelated but definitely suspicious scanning activity elsewhere in the network (scans appear as uni-directional since they typically include content in only one direction).

In the second week, 10.0.3.77 starts to communicate with Big Market's web servers, and we see additional highly regular SSH activity (orange) with several of the user machines including wss2-23.bigmkt2.com. Throughout the same time period, the workstation communicates with some of the same external IP's as in the first week, but the communications are now highly regular.

The anomalies indicated here are starting points for further analysis. A potential explanation is a spamming or phishing campaign that succeeds with some users, leading to unauthorized access to the user machines, but more work would be required to verify this or to supply an alternate explanation.

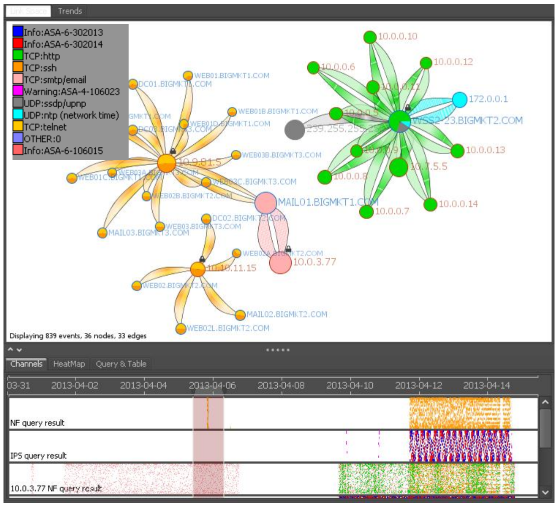

Figure 4: In the first week of data, activity at 10.0.3.77 consists of numerous regular smtp/email communications.

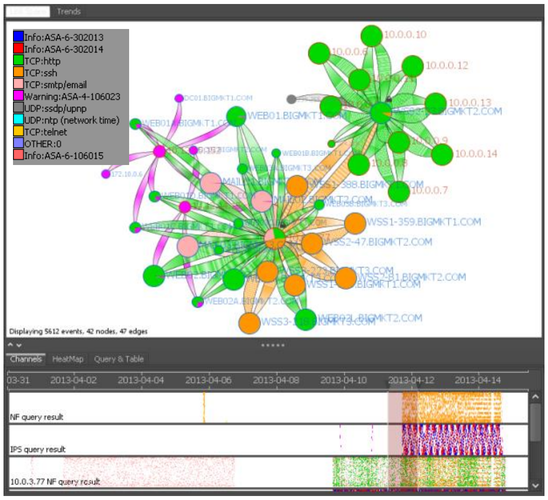

Figure 5: In the second week of data, activity at 10.0.3.77 includes additional http and ssh traffic.

5 Conclusion

Galaxy builds upon traditional link graph visualizations by taking advantage of *link space* to display information about flavor, volume, and time that is not typically displayed. Coupled with a timeline, this provides the analyst numerous opportunities to discover anomalies, piece together a story of network activity, and present that story to an audience.

References

[1] J. Heer, "Exploring Enron: Visual data mining of e-mail", 2005. Retrieved June 20, 2012, from http://jheer.org/enron/v1/.

[2] J. Pearlman and R. Rheingans, "Visualizing network security events using compound glyphs from a service-oriented perspective", *VizSEC*, 2007.