Visual Analytics for Network Flow Analysis

John R. Goodall and Daniel R. Tesone
Secure Decisions division of Applied Visions, Inc.
{ johng, dant } @securedecisions.avi.com

Abstract

Large corporations and government agencies are continually bombarded by malicious network attacks through the cyber infrastructure. One common method to identify and assess the im pacts of these m alicious activities is through the m onitoring and analysis of network flow data. W hile already somewhat aggr e-gated, the data can quickly become overwhelming – a billion flow records a day for large organizations is not abnormal. W e have integrated our visual analytics toolkit w ith netw ork flow data to provide a seamless workflow for computer network defense analysts. This integration can facilitate the learning process of novice analysts and make expert analysts more productive.

1. Introduction

Computer networks are growing larger and more complex as commercial and government entities have increasingly come to depend on the cyber infrastructure. Against this backdrop of increased complexity and reliance on the network infrastructure, cyber attacks have also increased. The stakes have increased as well. The 2007 Russian cyber attack against Estonia hints at the future of cyber warfare: coordinated bots can attack and cripple the cyber infrastructure of a nation. [1] Such coordinated attacks are common against large corporations or even the U.S. government, and may soon have the devastating effects of the Estonian attacks.

The Department of Homeland Security's United States Computer Emergency Readiness Team (US-CERT), among other responsibilities, is charged with defending the U.S. government's civilian agencies. A similar organization, Joint Task Force-Global Network Operations (JTF-GNO) exists to defend the networks of the military. US-CERT currently collects over a billion network flows a day. A network flow is a uni-directional aggregation of individual packets across several dimensions, typically including source and destination IP address, source and destination port, and protocol. Computer Network Defense (CND) analysts at US-CERT and other organizations need to sift through this voluminous flow data to try to find indications of malicious activity and to analyze and assess the implications of these events. Finding the information they need within this mound of data using conventional, textual tools is time consuming, inefficient, and prevents the analysts from developing a mental model of the big picture of network activity and determining what is important. This understanding, Situational Awareness (SA), is the key component of CND. Endsley describes SA as "knowing what is going on around you," and within that knowledge of your surroundings knowing what is important [2]. One doesn't need to know everything, only those things necessary to make accurate, timely decisions.

To provide Situational Awareness to CND analysts, we are developing a visual analytics platform for providing analysts with a tight integration with their network flow data and tools. This integration allows novice users to immediately begin using the visual analytics platform for data analysis, a leap forward from the complicated textual, command-line tools that, while extremely powerful, require a deep knowledge of not just the domain, but of the tools' myriad options. Novices and expert CND analysts alike also need to learn the nuances of a particular environment that they are monitoring. This learning process of understanding normal behavior is required to understand what is abnormal and to be able to analyze, prioritize, and assess the impacts of events.

The visual analytics platform brings together and links multiple information visualization views into a multi-display system that enhances SA through multiple levels of visual analysis, from a high-level dashboard overview to linked visualizations to the low-level textual details of the data. This enables analysts to view the data from multiple perspectives and levels of details. Analysts can drill into the network flow data where they can interact with multiple visualizations and drill down into the details of the data.

978-0-7695-3568-5/09 $25.00 © 2009 IEEE
DOI 10.1109/CATCH.2009.47

Figure 1. VIAssist visual analytics platform.

The design of the system, VIAssist, was informed by a Cognitive Task Analysis of CND analysts in commercial and military environments. [3] The results of this research formed the use cases and informs the design of VIAssist; for example, motivating the collaborative and reporting functionality that differentiate the system from other network flow visualization systems.

The features and use of VIAssist is described in D'Amico, et al. [4]. This paper focuses on the integration of the visual analytics with a network flow analysis system, SiLK, the System for Internet-Level Knowledge. The SiLK network traffic analysis tools are used by US-CERT, JTF-GNO, and other organizations to collect, store, and analyze network traffic as flows.

2. Meeting the Needs of CND Analysts

Based on the results of the Cognitive Task Analysis, we know that CND analysts need to be able to understand the big picture, to answer questions they didn't know they had, to put events into their larger context, to collaborate and work with other CND analysts, and to report their hypothesis and findings. VIAssist, shown in Figure 1, provides an intuitive, customizable dashboard (shown at right in Figure 1) to provide a big picture view. Multiple visualizations are linked together (shown at left in Figure 1) to facilitate exploration and discovery. Different kinds of visualizations are provided to enable the analysis of events in network, temporal, and geographic contexts. Collaboration is supported in multiple ways: through shared lists of critical and potentially malicious IP addresses, annotations, workspaces, and expressions. Embedded communication and reporting tools enable analysts to easily create and reuse reporting templates that allow non-technical users to understand findings through the visualizations.

CND analysts also need to be able to use new tools, like VIAssist, in conjunction with their current tools. For example, analysts often have a toolbox of command-line tools and scripts and web sites that they use to gather additional information about the attributes of a network flow record. VIAssist provides an extensible context-sensitive framework for integrating existing commands so that data attributes can be operated on by command-line or web-based tools; for example, the user can highlight an IP address and perform an nslookup or whois command to find out additional information for that IP address.

The current network flow analysis tools used by analysts lack key features or are limited in their scalability that reduces their utility. Command-line network flow tools can be powerful, especially when the output can be piped, or chained together, into other tools. However, these tools lack an overview of the data beyond simple statistics. Perhaps more importantly, they are difficult to learn and use. The syntax is often complex and there is no straightforward path to getting started. Novices need to learn not just this complicated syntax, but they often do not have a full understanding of the domain.

Spreadsheets can be useful, especially as a means of organizing data for simple charting, but are not a scalable solution, lack advanced visualizations, and do not have collaboration features required in the CND domain. Analysts from US-CERT identified some of

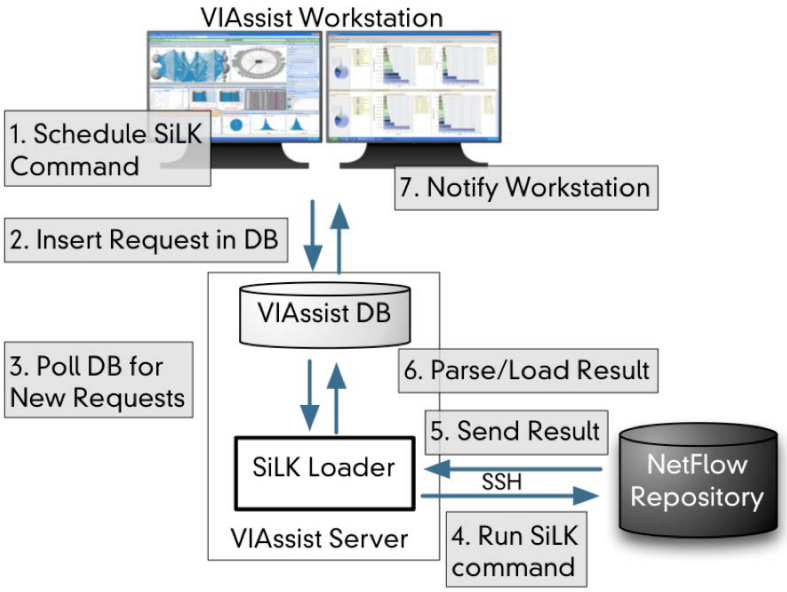

Figure 2. Network flow integration workflow.

the drawbacks of this approach, which included a limited plotting engine, a hard limit on the number of records that could be imported, and the need to format and import the data. [5] This last point is central to this paper, easing the process of moving data from a network flow repository to a visual analytics environment where novices and experts can explore and analyze the data.

Other network flow visualization tools require data to be inserted manually into a database or exported from a binary flow format into a text file that can then be brought into the visualization tool. This interrupts the CND analysts' workflow. Additionally, it requires that users be proficient in the network flow tools, which are typically command-line tools with a multitude of options that a new user can find daunting. Thus, to provide a means for fitting in with rather than disrupting a CND analysts' workflow and to provide a method for novice users to immediately begin visual data analysis and foster learning, we tightly integrated the flow data repository with the visual analytics system. Additionally, it is possible to not only query the network flow repository on demand, but to import existing commands and to schedule the commands to be run periodically (e.g. nightly). This kind of automation is a functionality often ignored in visual analytics systems.

3. Network Flow Integration

To demonstrate the utility of integrating network flow analysis tools into a visual analytics environment, we integrated VIAssist with the SiLK flow collection and analysis tools. SiLK is a collection of open-source traffic collection, processing, storage, and analysis tools. The SiLK tools are powerful and used by many expert CND analysts. These network flow analysis tools can be useful in preprocessing data, augmenting flow data with metadata, aggregating data, and filtering the data based on a complex series of conditionals.

In order to ensure that processor-intensive data queries would not impact the transactional data repository, we separated the visualization repository from the network flow repository. This ensures that the data analysis will not interfere with the data collection. This is also the workflow used at US-CERT and JTF-GNO.

3.1 Workflow

The workflow of scheduling and running a SiLK command, shown in Figure 2, is described below:

1. A user first schedules a new request within VIAssist to query the network flow repository. SiLK stores data in a compact, binary network flow format. The user creates the request in an intuitive dialog discussed further in the next section.

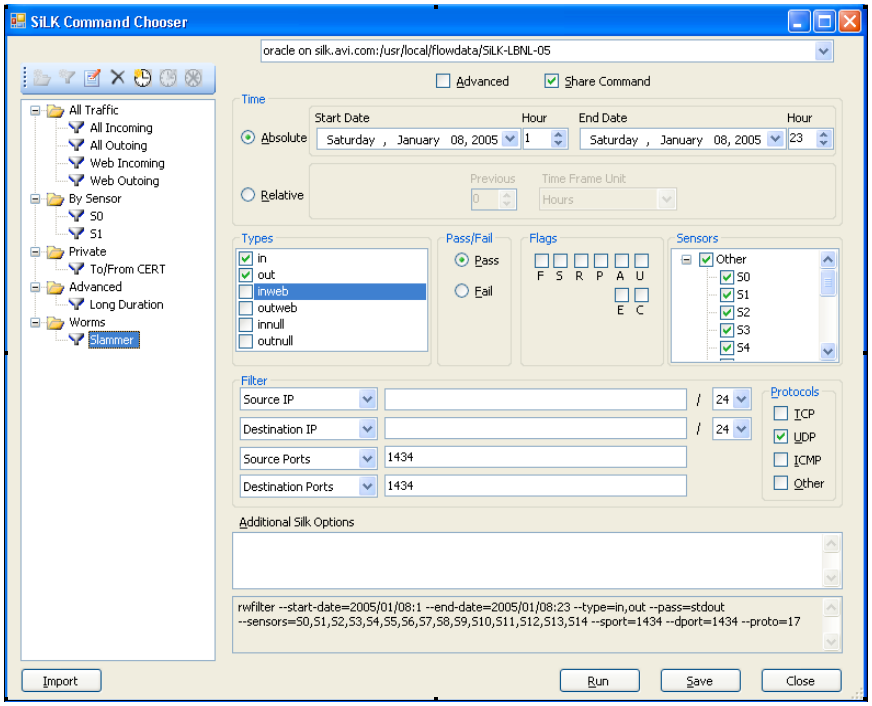

Figure 3. Network flow command chooser.

2. When the command is created, or an existing command is selected by the user, a new request is inserted into the VIAssist database.
3. A separate service, the SiLK Loader, which runs as a daemon on the VIAssist database server, continually polls the database for new requests.
4. When a new request is found, the SiLK Loader translates that request into SiLK commands that are run from the command-line as an external process. The command is sent to the flow repository over an encrypted secure shell (SSH) connection using public key authentication.
5. The compressed, binary SiLK data is piped back over the SSH connection where it is translated into text and parsed into a fixed-width, delimited format a database batch loading tool can upload.
6. After cleaning up the data and adding additional metadata (e.g. the user who issued the request) to the flow data, the data is loaded into the database using a batch loading tool.
7. When the loading is finished and any required post-processing has completed (e.g. adding geo-location information to IP addresses), the SiLK Loader sends a notification to the VIAssist application, which then notifies the user that the new data has been loaded and is ready to be visualized.

This entire process is transparent to the user. The user only needs to be able to import their current SiLK commands or point-and-click on some common filtering options in order to pull data from a network flow repository. The workflow is customized for a SiLK network flow repository, but additional types of flow types could easily be swapped in. Likewise, we are currently using an Oracle database, but VIAssist was designed to be database agnostic, and the data load mechanism would simply need to change the batch load command.

To highlight the simplicity of this approach, consider the alternatives. One would have to log into the network flow repository, use the command-line filtering tool specific to that network flow type to get the requested data. A simple request looks like this:

```
rwfilter --data-rootdir=/flowdata --site-config-
file=/usr/local/silk.conf --start-date=2008/10/08 --
end-date=2008/10/09 --type=in,inweb --pass=stdout
--sensors=S0,S1,S2,S3 --protocol=6
```

This command gets 24 hours of incoming TCP data from four sensors and outputs it in binary format to standard output, where another command will parse the data into human-readable text. Then the data will need to be copied to the database server, manually transformed into a format that the database can interpret, and the database tools must then be used to load the data. This process is both tedious and requires familiarity not just with domain-specific network flow

analysis tools, but also with general database tools. Novice users will not know where to begin.

3.2 User Interface

The user interface for creating or importing SiLK commands is shown in Figure 3. This interface is specifically designed to be simple, presenting the most important options in a way that a novice can easily understand, without cluttering the display with all possible options. The command-line filtering syntax is arcane and it is easy to make a mistake. Although the details of how to structure the command-line tools are not required to begin using the visual analytics environment, the presence of the syntax at the bottom of the interface can help new users learn. As they configure different options through the interface, the corresponding command is automatically built and displayed.

Many expert CND analysts already have a library of commonly used commands, which can be imported into the system. Thus, expert users can easily import their existing commands, or directly edit the command in a free-form text box. All commands can be shared or stored privately for the analyst. Sharing the commands enables expert users to import or create a library of commonly used commands that novice and intermediate users can use either as-is or as a starting point for their own commands.

Commands can be run immediately or scheduled to run on an ongoing basis. This automation allows users to create a command to fetch data from the flow repository every night that can then be examined the following day. Coupled with the embedded reporting templates, which allow users to save and reuse PowerPoint-like templates, reports of the previous day or night's activities can easily be generated and shared with other analysts and management.

5. Related Work

NVisionIP is a visualization system aimed at increasing an analyst's situational awareness by visualizing flows at multiple levels of detail. [7] At the highest level of aggregation, NVisionIP displays a Class B network (65,534 IP addresses) as a scatterplot, with points representing an IP address that has had a flow. NVisionIP also provides capability to drill down into the data through a small-multiple view and a histogram of host details. VIAssist can be configured to show scatterplots similar to the one in NVisionIP and other views representing lower levels of data can be linked with it, allowing for simultaneous understanding of the flow data.

NVisionIP was extended to "close the loop" by allowing users to create rules from the visualization that can then automatically alert on new data. [8] While we have not yet linked the visualization to automatically create the rules required to find patterns in new data, we have moved towards providing automation. VIAssist allows users to automate and schedule the flow analysis collection, a process that will be described in more detail below.

FlowTag is a visualization system to enable users to tag flow data to support analysis and collaboration. [9] Tagging allows analysts to label key elements during the analytic process to reduce the cognitive burden of analysis and maintain context. Tagging can also be used for sharing and collaboration. Tagging has become popular recently with social networking and social bookmarking sites; adapting the concept to CND is a logical step. FlowTag brings the popular concept of tagging to the problems of analyzing and sharing network security data. Any data item can be annotated within VIAssist, but currently only support more structured tagging of IP addresses as either critical to an organization (internal addresses) or noteworthy external addresses. Annotated and critical/noteworthy IPs can be shared among multiple analysts within an organization, and data can be searched based on the annotations. Future plans include the ability to provide more robust tagging support.

Isis supports the analysis of network flows through two visualization methods, progressive multiples of timelines and event plots, to support the iterative investigation of intrusions. [10] Isis combines visual affordances with structured query language (SQL) to minimize user error and maximize flexibility. Isis keeps a history of a user's investigation, easily allowing a query to be revisited and a hypothesis to be changed. Two of the strengths of Isis are to present the user with a very high-level overview and then allow them to drill into the data, a trait that VIAssist also has, and to provide the user with a visual history of their investigation. This workflow encourages users to try different hypotheses without fear of losing their investigatory thread.

All of these systems share one thing in common. They require users to input their flows manually into a relational database or they read directly from a flow or text file. Both of these approaches have their advantages and disadvantages. A database permits more scalability, but requires knowledge of SQL and the import mechanisms specific to the database vendor. Tools that read directly from a flow file are more straightforward, but require parsers for all of the many different flow formats. Reading from a text file, rather than a native binary format specific to one flow tool or another, alleviates this problem, but these text files can

be extremely large. VIAssist takes the best of each of these approaches: taking advantage of the scalability of relational databases and the small file size of the SiLK binary data format. The flow integration is specific to the SiLK tools, but a new parser that interfaces with other flow formats could easily be integrated.

6. Future Work

We are currently extending VIAssist to include more generic tagging of data, the ability to undo user interactions, and are adding additional visualization displays specific to CND in order to further foster learning. While VIAssist currently allows users to tag IP addresses as either "critical," for important local hosts, or "hot," for hosts suspected of malicious behavior, we plan to extend this tagging functionality to be more generic, allowing analysts to create and share tags, which can then be used to visualize, filter, and highlight data.

We are also looking into more robust flow integration and automation methods. One area that has potential is to automatically translate the current view of the data (i.e. the data shown and the filters applied) into a SiLK command. This would allow novices even greater flexibility in beginning to work with VIAssist, as they would be able to interact with the data visually and translate that to future automated queries to the flow repository.

7. Conclusions

We have architected a workflow that fits within CND analysts' current workflow, making the time-consuming and often difficult task of getting data from a network flow repository into a visual analytics platform transparent to the user. This data gathering process can be shared among analysts and automated to bring network flow data into the visual analytics platform on a regular basis. The visual analytics platform for network flow analysis described in this paper can foster learning and has the potential to make novices and experts both more productive, more quickly by allowing them to focus on data analysis rather than data parsing and by bringing the data into an environment that is conducive to helping them learn the nuances of behavior within a particular environment.

8. Acknowledgements

This work was sponsored by the Department of Homeland Security (DHS) Science and Technology (S&T) Directorate under contract number FA8750-08-C-0140 and the Department of Defense under contract F30602-03-C-0260.

9. References

[1] M. Landler and J. Markoff, "Digital Fears Emerge After Data Siege in Estonia", *The New York Times* , 2007, http://www.nytimes.com/2007/05/29/technology/29estonia.html, Retrieved 10/10/2008.

[2] M.R. Endsley, "Theoretical Underpinnings of Situational Awareness: A Critical Review", *Situation Awareness Analysis and Measurement: Analysis and Measurement* , M.R. Endsley and D.J. Garland (eds.), Lawrence Erlbaum Associates, Mawah, NJ, 2000.

[3] A. D'Amico, K. Whitley, D. Tesone, B. O'Brien, and E. Roth, "Achieving Cyber Defense Situational Awareness: A Cognitive Task Analysis of Information Assurance Analysts", *Proc. of Human Factors and Ergonomics Society 49th Annual Meeting*, HFES Press, 2005, 229-233.

[4] A.D. D'Amico, J.R. Goodall, D.R. Tesone, and J.K. Kopylec, "Visual Discovery in Computer Network Defense", *IEEE Computer Graphics and Applications* 27(5), IEEE Press, 2007, 20-27.

[5] L. Rock and J. Brown, "Flow Visualization Using MS-Excel Visualization for the Common Man", *Proc. of FloCon*, 2008, http://www.cert.org/flocon/2008/presentations/, Retrieved 10/10/2008.

[6] J.R. Goodall, W.G. Lutters, and A. Komlodi, "I Know My Network: Collaboration and Expertise in Intrusion Detection", *Proc. of ACM Conf. on Computer-Supported Cooperative Work (CSCW)*, ACM Press, 2004, 342-345.

[7] K. Lakkaraju, W. Yurcik, and A.J. Lee. "NVisionIP: NetFlow Visualizations of System State for Security Situational Awareness", *Proc. of ACM Workshop on Visualization and Data Mining for Compu ter Security (VizSEC/DMSEC)*, ACM Press, 2004, 65–72.

[8] K. Lakkaraju, R. Bearavolu, A. Slagell, and W. Yurcik, "Closing-the-Loop: Discovery and Search in Security Visualizations", *Proc. of IEEE Workshop on Information Assurance and Security (IAW)*, IEEE Press, 2005, 58–63.

[9] C.P Lee, and J.A. Copeland, "Flowtag: A Collaborative Attack-Analysis, Reporting, and Sharing Tool for Security Researchers", *Proc. of ACM workshop on Visualization for computer security (VizSEC)*, ACM Press, 2006, 103–108.

[10] D. Phan, J. Gerth, M. Lee, A. Paepcke, and T. Winograd, "Visual Analysis of Network Flow Data with Timelines and Event Plots", *VizSEC 2007: Proc. of Workshop on Visualization for Computer Security* , J.R. Goodall, G. Conti, and K.L. Ma (eds.), Springer, Berlin, Germany 2007, 85-99.