VIAssist: Visual Analytics for Cyber Defense

John R. Goodall, *Member, IEEE* and Mark Sowul

***Abstract*—Analysis of voluminous computer network data has become a common practice for cyber defense, but few tools provide adequate support for cyber-infrastructure defenders' workflow, visual exploration, IP geo-location, scalability, collaboration, or reporting. The state-of-the-art in visual analysis tools for cyber defense is typically no more than spreadsheets and primitive charting. While familiar to users, this approach ignores the human perceptual ability to identify novel patterns and anomalies when data is presented graphically. This paper reports on a visual analytics systems, VIAssist, being developed for cyber-infrastructure protection that helps cyber defenders better understand the massive, multi-dimensional datasets to protect our nation's critical infrastructure.**

***Index Terms*—Visual analytics, visualization, site security monitoring, computer network security.**

I. 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, the number of cyber attacks against critical cyber-infrastructure has 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].

To combat this threat, we are developing technologies for cyber-infrastructure defenders to facilitate the discovery, analysis and understanding of cyber attacks. This visual analytics platform, VIAssist, shown in Figure 1, enhances situational awareness, facilitates collaboration and reporting, and enables the analysis and understanding of cyber events.

A Cognitive Task Analysis of Computer Network Defense (CND) analysts in commercial and military environments informed the system's design. Based on the results of this research, we know that cyber defenders 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 cyber defenders, and to report their hypotheses and findings.

New analysis tools must fit within defenders' current workflow and work with the tools and data they currently use. To this end, we have integrated VIAssist with the SiLK network flow analysis tools.[1] We have also provided a mechanism for plugging in additional commands.

Visual tools must bring together and link multiple information visualization views to present data from multiple perspectives. Tools should also take advantage of multiple displays that are common in today's workspaces. Data should be presented at different levels of detail to support multiple levels of visual analysis, from a high-level dashboard overview to linked visualizations to the low-level textual details of cyber-related data. VIAssist provides an intuitive, customizable dashboard to provide a big-picture overview of network flow data. Multiple visualizations are linked together to facilitate exploration and discovery. Different kinds of visualizations are provided to enable the analysis of events in network, temporal, and geographic contexts.

Even through network flow data is already somewhat aggregated, sizes can grow to be extremely large, so systems must address issues of scalability. VIAssist does so through automatic Smart Aggregation , which keeps both the size of data manageable and the presentation of data visually understandable. Drilling into the data from this aggregated state enables interactively increasing the level of detail.

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 users to easily create and reuse reporting templates that allow non-technical users to understand findings through the visualizations.

II. Related Work

NVisionIP is a visualization system aimed at increasing a CND analyst's situational awareness by visualizing flows at multiple levels of detail. [2] At the highest level of aggregation, NVisionIP displays a Class B network (65,534 IP addresses) as a scatter plot, with points representing an IP

Manuscript received March 26, 2009. This work was supported in part by the U.S. Department of Homeland Security under contract FA8750-08-C-0140 and the U.S. Department of Defense under contract F30602-03-C-0260.

J. R. Goodall is with the Secure Decisions division of Applied Visions, Inc. (phone: 518-632-4195; e-mail: johng@securedecisions.avi.com)

M. Sowul is with Applied Visions, Inc. (email: marks@avi.com)

[1] http://tools.netsa.cert.org/silk/

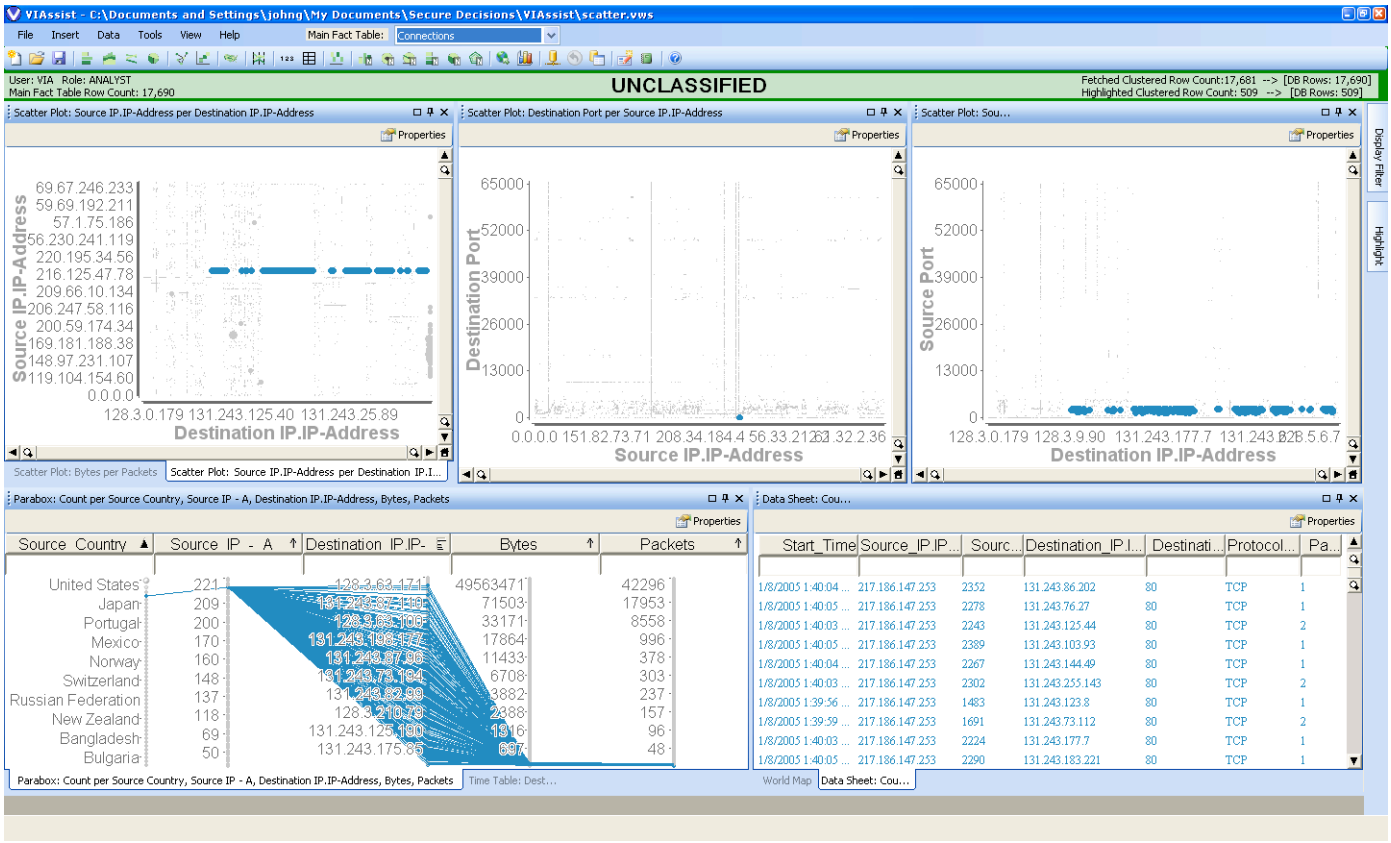

Figure 1. VIAssist visual analytics platform: coordinated multiple views highlight a single source IP address sending network packets to many internal destination addresses, as shown in the horizontal band in the scatter plot (top left), to a single source port (top center), from many source ports (top right). These connections are relatively small and short, as shown in the parallel coordinate view (bottom left), which shows the distribution of connections from a single source to many destinations. The details are linked to each of the other views (bottom right).

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 scatter plots similar to the one in NVisionIP, and other views representing lower levels of data can be concurrently 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. [3] While we have not yet linked the visual analytics to automatically create rules to find patterns in new data, we have moved towards providing automation. VIAssist allows users to automate and schedule the flow analysis collection; this will be described in more detail below.

FlowTag is a visualization tool to enable users to tag flow data to support analysis and collaboration. [4] Tagging allows analysts to label key elements during the analytic process to reduce the cognitive burden of analysis and to 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 computer network defense 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 it currently only supports 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 users within an organization, and data can be searched based on the annotations. Future plans include the ability to provide more robust tagging.

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. [5] It 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 user to revisit a query and change a hypothesis. 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 shares, 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 increases 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.

III. Visual Analytics for Cyber Defense

A. Support for Cyber Defenders' Workflow

1) Cognitive Task Analysis

We previously conducted a cognitive task analysis (CTA) to gain insight into the workflow, cognitive skills, and tools that cyber defenders rely on and the cognitive challenges and obstacles they face. We interviewed and observed 41 CND analysts from one commercial and six Department of Defense organizations responsible for network security. Participants varied in level of expertise from novice to expert and represented a variety of roles.

The results of this research formed the use cases and informed the design of VIAssist – for example, motivating the collaborative and reporting functionality that differentiates the system from other visualization systems. Dealing with massive data was one of the most significant challenges that the cyber defenders faced. This need motivated the Smart Aggregation functionality, discussed below. The results of the CTA are detailed in D'Amico, et al. [6]

2) Data Integration

Current network flow analysis tools used by cyber defenders lack key features or are limited in their scalability, which 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 complex and there is no straightforward path to getting started. Not only do novices need to learn 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 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. [7]

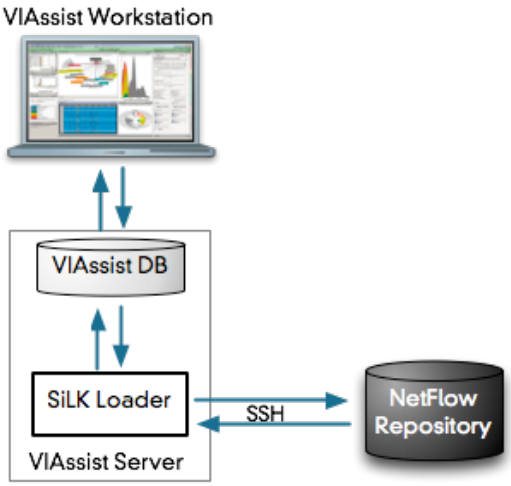

Figure 2. Network Flow integration.

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 cyber defender's workflow and 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 defenders' workflow rather than disrupting it, 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. This is notionally shown in Figure 2, which shows the flow of data between the VIAssist workstation, server and network flow repository. 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. The integration with network flow tools is described more fully in Goodall and Tesone. [8]

3) Data Tools

Cyber defenders also need to be able to use new tools, like VIAssist, in conjunction with their current tools. For example, cyber defenders often have a toolbox of websites and command-line tools and scripts 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 the user can operate on data attributes with command-line or web-based tools; for example, the user can highlight an IP address and perform an nslookup, or highlight a port and perform a port lookup.

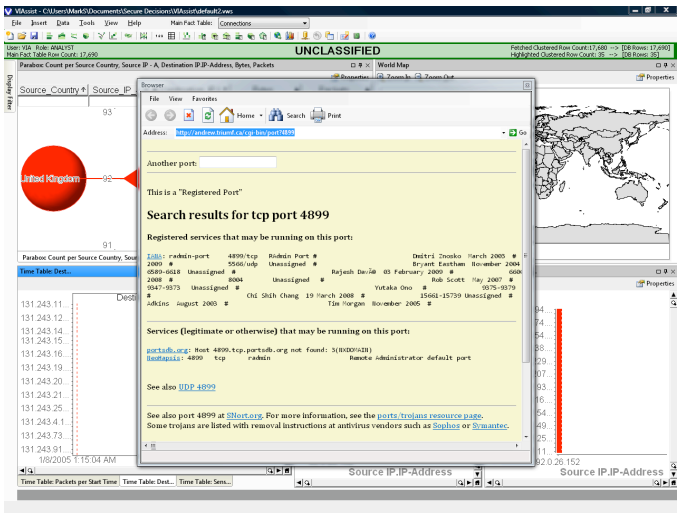

Figure 3. Using the port lookup data tool.

B. *Visual Information Seeking*

VIAssist can facilitate discovery by providing a platform for visual exploration. The process of discovering unexpected patterns is more fully described in D'Amico, et al. [9]

In support of the discovery and exploration process, VIAssist follows Shneiderman's Visual Information Seeking Mantra: overview first, zoom and filter, details-on-demand. [10] A customizable dashboard provides this overview. Users can semantically zoom into the data by increasing the fidelity of the data, and can filter using interface widgets such as sliders and checkboxes. Additionally, all of the visualization views in VIAssist are linked, so semantic zooming and filtering in one view is reflected in the others. Finally, there are methods for providing the details of selected visual items.

1) *Overview*

The overview of the dataset is provided by a customizable dashboard, consisting of any number of simple charts showing the Top N elements for any field in the data. For example, a bar chart can show the Top 10 source IP addresses and a pie chart can show the Top 5 protocols. The Top N rankings can be based on the record count or other numeric values, such as the number of packets or number of bytes.

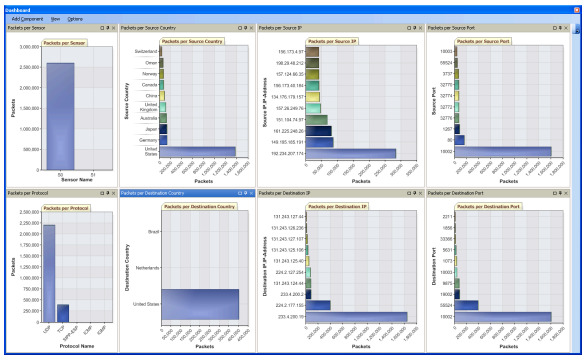

Figure 4. VIAssist dashboard window showing high-level overview of network activity.

The dashboard is intended to be simple in order to present high-level information that cyber defenders can quickly digest.

2) *Coordinated Multiple Views*

Our design follows the principle that no single visualization can accomplish a cyber defender's myriad tasks. Instead, rather than rely on any single visualization, as many systems do, we present coordinated multiple views. Each visualization view is coordinated and linked to the other views, as shown in Figure 5. Interactive linking and brushing (selecting a subset of the data with the mouse) allows visual data elements in one view to be highlighted in all of the other views. This technique makes it possible to see dependencies and correlations in the data by presenting multiple perspectives on the data. [11] This method provides more information than the sum of the component visualizations.

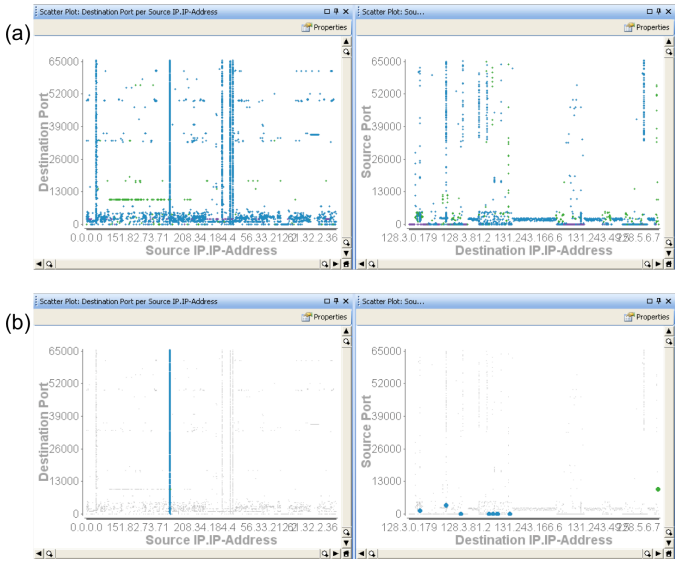

Figure 5. Above (a) shows two coordinated views; below (b) shows the same two views with a vertical band representing one source IP address brushed on the left and the corresponding data is highlighted on the right.

Views can be relocated and resized within the workspace. These workspaces can be saved and shared among cyber defenders. Multiple displays can be used to increase the number of linked views.

3) *Details-on-demand*

While users in most domains will require access to the details of their analytic exploration, this is particularly true in CND. Without access to the details, cyber defenders cannot make the final decisions about the importance or significance of an event. Details can be dynamically linked so that selection made in any visualization view is automatically shown in the tabular view.

C. *Geographic Visualization*

Understanding the geographic distribution of computer networking traffic is critical to analysis. In addition to exposing the associated country for all IP addresses to be used

in any visualization, we have also integrated a high-level geographic visualization. This visualization allows a cyber defender to determine the physical location of connection endpoints (IP addresses). The geo-location lookup data for this feature provides the country, city, and approximate latitude/longitude for all IP addresses. The geographic visualization can show IP distribution and location in several ways: *region shading*, which colors each country according to its relative contribution to all of the network traffic in the visible data set; *endpoint nodes*, which plots the endpoint of each connection as a variable sized node (according to its data attributes) at its approximate latitude and longitude; and *connection links*, which plots each connection as a variable sized link (according to its data attributes) between its source and destination endpoints.

For each of these visualizations, the user can choose to color its visual markers in one of two ways, either as a monochromatic scale, indicating each marker's relative value, or by mapping specific values or sets of values to individual colors. For endpoint nodes, each node's size can also be scaled to its relative weight. Each of these values can be a simple count of the number of records (i.e. the number of connections), or it can be weighted by numeric values, such as the number of bytes or packets.

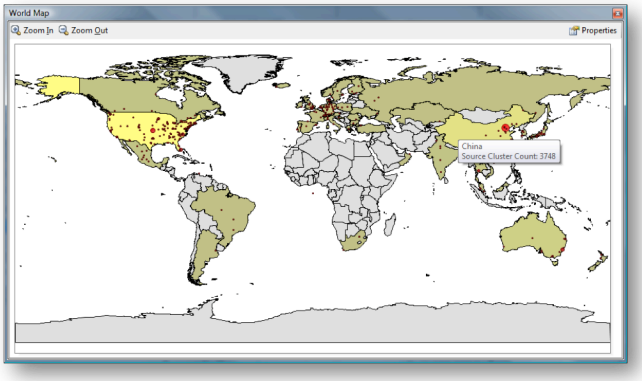

Figure 6. Geographic visualization of source addresses as shaded regions for an overview of geographic distribution, and individual points for distribution within a region.

In Figure 6, for example, endpoint nodes are drawn as red points and sized according to the number of packets, and countries are shaded by source activity in yellow. Brighter shades of yellow indicate more activity. In this example, the US is the source of most of the traffic, and an endpoint in the middle of the US is the source of a noticeably high number of connections. China is another heavy source of connections, and an endpoint in the northeast of China is the single largest source of connections. Darker countries have less source traffic. Countries in gray are not present in the data being displayed.

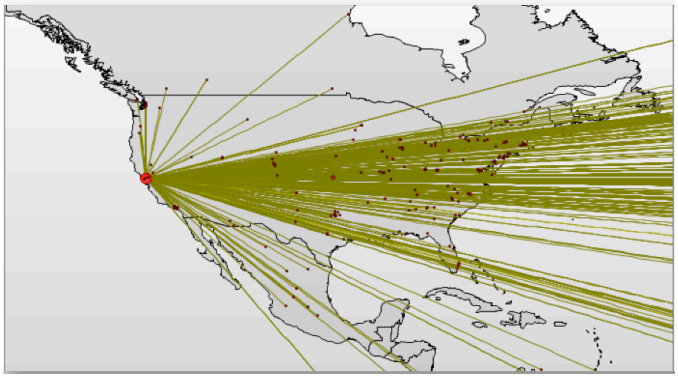

Figure 7. Connections plotted as lines, with total activity for each point plotted by size. This dataset shows most of the connections have one endpoint in San Francisco.

Region shading is determined through the relative contribution to the network traffic, either by its total contribution to traffic (source + destination), or by its contribution only as a traffic source or destination. Latitude and longitude values can be used to place individual endpoint nodes on the map. Once again, the user can choose whether to plot all traffic, or only source or destination points. Connection links show each network flow as a line from its source to destination. Users can choose to display lines directly from source points to destination points, or aggregate these to show connections only between countries.

D. Scalability

Dealing with large data sets is a pressing problem in dealing with network flow data, which can quickly scale to huge volumes that overload most systems. Traditional data management and interactive retrieval approaches have often focused on solving the data overload problem at the expense of users' Situation Awareness. For VIAssist, we developed a new data management strategy, coined Smart Aggregation, as a powerful approach to overcome the challenges of both massive data sets and maintaining Situational Awareness. By combining automatic data aggregation with user-defined controls on what, how, and when data should be aggregated, the system can handle massive amounts of data while maintaining a user's Situational Awareness. This approach ensures that a system is always usable in terms of both system resources and human perceptual resources. Details of the Smart Aggregation approach and a comparison with other data management methods are further described in Tesone and Goodall [12].

E. Collaboration

VIAssist offers several methods of collaboration, including annotations and structural tagging. Data elements can be annotated and shared among users. These annotations are freeform and the elements of an annotated item are tagged in the central database, and can be colored, highlighted and filtered. Annotations allow users to communicate hypotheses, findings and status across shifts and locations.

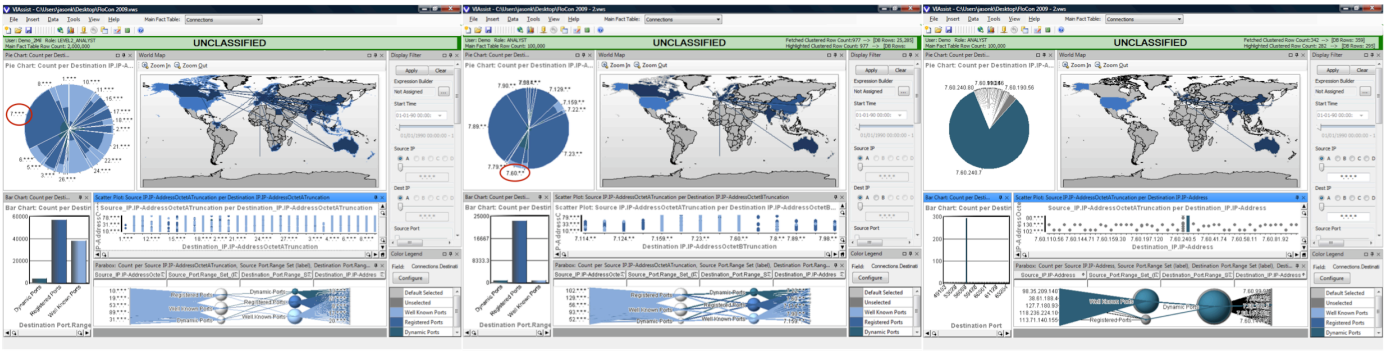

Figure 9. 2,000,000 records are passed through the Smart Aggregation, which (a) automatically aggregates the destination IP addresses by the first octet, from which (b) the user can drill in to the first two octets, and finally (c) drill in

Users can also tag IP addresses that are known bad actors and known critical hosts. For the former, users can tag these 'Hot IPs' either personally or organization wide, which are then stored centrally and available for all users to use for color assignments or filtering and highlighting. For example, if an organization is being repeatedly targeted by a group of attackers within the same IP address space, these addresses could be globally flagged, resulting in those addresses being highlighted in the various visualizations used by all users. For the latter, users can tag 'Critical IPs' that are critical to the organizational mission, to allow the most important machines to be prioritized higher during analysis than, for example, workstations.

F. Reporting

Our conversations with CND analysts revealed reporting to be an important part of their workflow, both for routine briefings and for escalation of suspicious or interesting findings. To this end, we developed an integrated report designer component to allow users to construct these reports easily, without even leaving the application. These reports can contain screenshots of the workspace and its component visualizations, as well as text and simple drawing annotations.

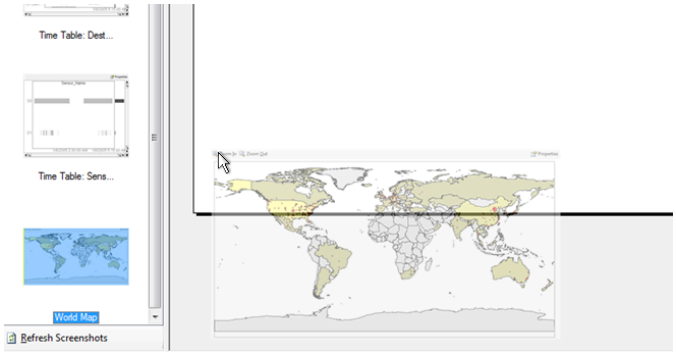

Figure 8. Drag and drop reporting interface.

Modeled after PowerPoint, the report designer uses a familiar drag and drop interface, shown in Figure 8, allowing users to drag visualization screenshots and annotations onto the canvas, as well as alter a component's layout and size. Moreover, these operations offer live previews – for example, when resizing or moving an image, a semi-transparent "ghost" image previews its new size and position. This feature can be expensive, so it is automatically disabled if it begins to take a long time to draw the screen. It is also disabled when the application is running remotely under Remote Desktop or Terminal Services.

When creating slides for routine briefings, it is common for cyber defenders to create the same basic report each time: the same visualizations will be present (but with different data), some of the same text will be used each time, and certain "boilerplate" content such as titles and logos will be the same. To this end, we provide support for "templates." When a user opens a template, the report designer will populate each visualization screenshot with the current state of the same visualization. Text that the user chose not preserve in the template will remain only as placeholder text boxes that the user can populate with whatever is appropriate (current date, for example).

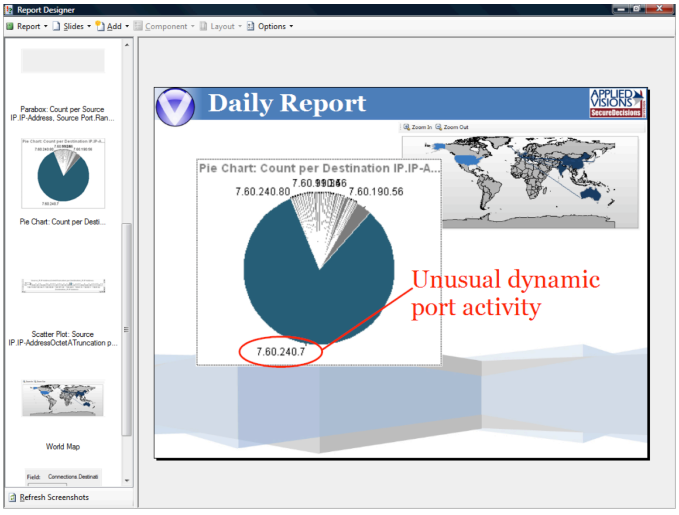

Figure 10. Report template, automatically populated with two of the visualizations and manually annotated to highlight the data.

Once the user is satisfied with the report, he or she can export it directly to a standard PowerPoint or PDF file from

within the application, to distribute to colleagues or use in a briefing or presentation.

IV. Implementation

VIAssist is built on Microsoft's .NET Framework, and combines several different third party visualizations from different vendors into one coherent application. This was a challenge: some are unmanaged COM components and some are native .NET controls, each one has its own data model and its own configuration options, and each has its own share of quirks and bugs. Nevertheless, we developed a framework that allows coordination of these views.

VIAssist determines what views are available to the application by reading in a metadata file that contains information describing each view, such as the vendor name and visualization name, as well as the name of the class that wraps the visualization and the assembly (compiled code) containing this class. When the user chooses to insert a view into the application, this information, coupled with the .NET Framework's reflection capabilities, is used to instantiate the specified class at runtime. In this way, the available visualizations are not hard-coded into the application but merely require the presence of metadata and the compiled code in order for VIAssist to make use of them.

VIAssist's data retrieval is also handled by metadata, which describes the fields available in the database and how to retrieve them (table names, column names, data type, etc.). Each view specifies what fields it requires, and this information is fed into a SQL generator that creates the database query necessary to retrieve the data. The Smart Aggregator determines how much data will be returned and, if necessary, replaces these fields with aggregated versions in order to reduce the amount of data retrieved from the database.

V. Scenario

This scenario demonstrates the efficacy of VIAssist in finding anomalous network traffic. Figure 11 depicts a basic VIAssist workspace.

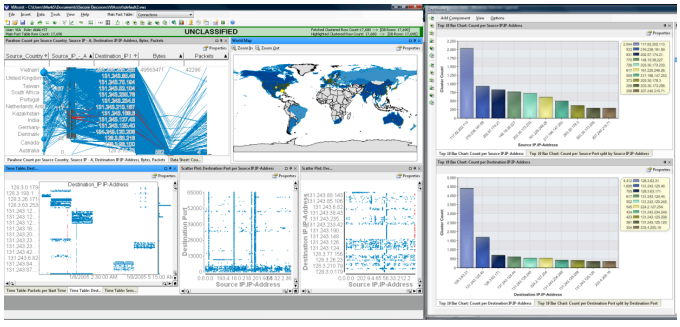

Figure 11. VIAssist workspace: the left side contains the main analysis window, while the right side contains the dashboard overview.

There are several views open in the main window. Clockwise from top-left, there is a parallel coordinates view showing the relationships between several data attributes, a geographic view with countries shaded by source and source points plotted in yellow, a scatter plot depicting source IP vs. destination IP, a scatter plot depicting source IP vs. destination port, and finally a temporal scatter plot showing activity at each destination IP for any given time.

The dashboard shows two bar charts representing the top 10 source and destination IPs. Beginning with the most active source IP, clicking its bar in the dashboard directs the other views to highlight only data with that source IP (Figure 12). Looking at the scatter plot of source IP vs. destination port (bottom center), a vertical row is highlighted. Such a vertical row in this scatter plot implies that a single source IP was connecting to one or more machines on every port; this almost certainly implies that this source IP is doing a port scan.

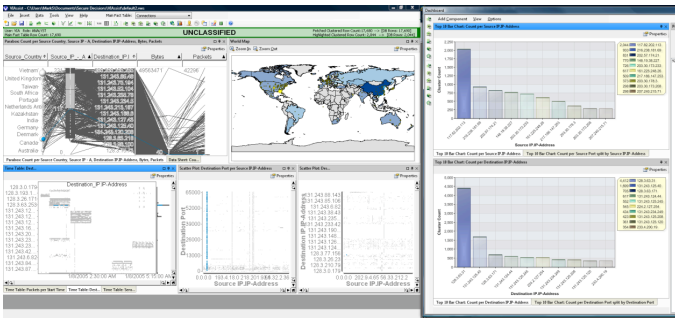

Figure 12. Selecting the most active source IP via the dashboard.

Drilling in to this source IP will filter out the rest of the data, as shown in Figure 13, which shows that this source is located in China (visible in the parallel coordinates and geographic views) and connected to only one destination IP: 128.3.63.31 (visible in the parallel coordinates view and source IP vs. destination IP scatter plot). The dashboard shows that this destination IP is also the most active overall.

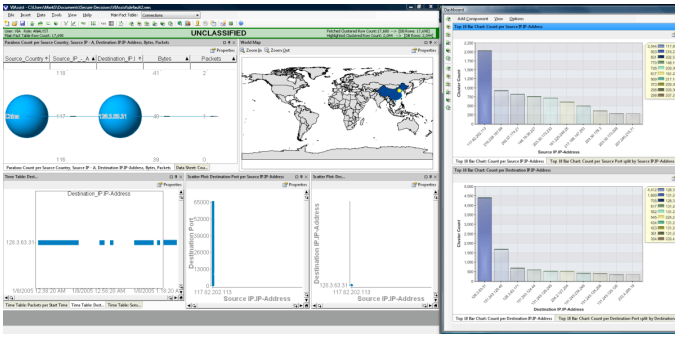

Figure 13. Display after filtering to most active source IP.

The temporal scatter plot indicates that this port scan took place over a period of 40 minutes (12:38 AM to 1:18 AM). Such a slow scan could indicate that the source was trying not to arouse suspicion. The parallel coordinates view also shows that every connection consisted of a single 40-byte packet, so it is unlikely that the source managed to infiltrate the system or exfiltrate any data.

It will be interesting to see what other connections there

were to this destination IP address. Selecting this destination IP in the dashboard highlights every other vertical line in the source IP vs. destination port scatter plot, showing that this destination has been the target of port scans by not one but several source IPs, including each of the three most active source IPs. This is shown in Figure 14.

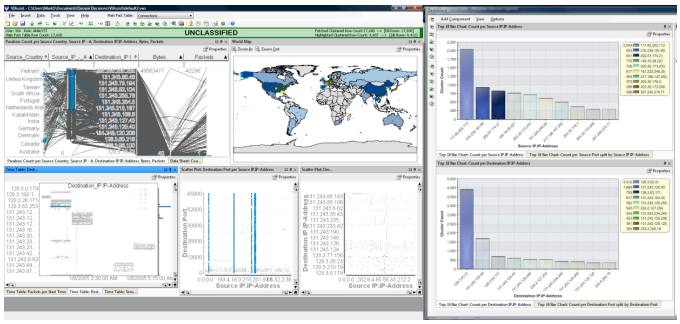

Figure 14. Selecting the most active destination IP via the dashboard.

VIAssist has quickly exposed suspicious behavior within this dataset. The next step would be to create a report to summarize these findings.

VI. Future Work

We are currently developing a visualization framework that will allow us to develop custom visualizations and to integrate existing visualizations into the VIAssist platform.

We are also exploring extending VIAssist to include more generic tagging of data. 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 users to create and share tags, which can then be used to visualize, filter, and highlight data.

We are also exploring adding functionality to easily import new data sources. Currently, a manual modification of a meta-data file that defines the fields in the data source is required. By automating this, cyber defenders – and users from other domains beyond CND – could quickly import and visualize new data sources.

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.

This product includes GeoLite data created by MaxMind, available from http://www.maxmind.com/.

References

[1] J. Kirk, "Estonia recovers from massive denial-of-service attack," in *Network World*, 2007.

[2] K. Lakkaraju, W. Yurcik, and A. J. Lee, "NVisionIP: netflow visualizations of system state for security situational awareness," in *VizSEC/DMSEC: Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*, New York, NY, USA, 2004, pp. 65-72.

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

[4] C. P. Lee and J. A. Copeland, "Flowtag: a collaborative attack-analysis, reporting, and sharing tool for security researchers," in *VizSEC: Proceedings of the 3rd international workshop on Visualization for computer security*, New York, NY, USA, 2006, pp. 103-108.

[5] D. Phan, J. Gerth, M. Lee, A. Paepcke, and T. Winograd, "Visual Analysis of Network Flow Data with Timelines and Event Plots," in *VizSEC 2007: Proceedings of the Workshop on Visualization for Computer Security*, J. R. Goodall, G. Conti, and K. L. Ma, Eds.: Springer, 2008, pp. 85-99.

[6] 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," in *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting*, 2005, pp. 229-233.

[7] L. Rock and J. Brown, "Flow Visualization Using MS-Excel Visualization for the Common Man," in *FloCon*, 2008.

[8] J. R. Goodall and D. R. Tesone, "Visual Analytics for Network Flow Analysis," in *Proceedings of the Cybersecurity Applications & Technology Conference For Homeland Security (CATCH)* Washington, D.C.: IEEE Press, 2009, pp. 199-204.

[9] 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,* vol. 27, pp. 20-27, 2007.

[10] B. Shneiderman, "The eyes have it: A task by data type taxonomy of information visualizations," in *Proceedings of the IEEE Symposium on Visual Languages*, 1996, pp. 336-343.

[11] D. A. Keim, "Information Visualization and Visual Data Mining," *IEEE Transactions on Visualization and Computer Graphics,* vol. 8, pp. 1-8, 2002.

[12] D. R. Tesone and J. R. Goodall, "Balancing Interactive Data Management of Massive Data with Situational Awareness through Smart Aggregation," in *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*: IEEE Press, 2008, pp. 67-74.