CluVis: Dual-Domain Visual Exploration of Cluster/Network Metadata

(Special Session on Computer Security)

Christopher Waters* crw7@msstate.edu
Jonathan Howell jeh118@msstate.edu
T.J. Jankun-Kelly tjk@acm.org

Department of Computer Science and Engineering
James Worth Bagley College of Engineering
Mississippi State University
Mississippi State, MS 39762

ABSTRACT

CluVis, a prototype for visual monitoring and exploration of cluster and network metadata, is introducted. CluVis builds upon interactively added charts of cluster/network metadata (e.g., packets received, processor activity, etc.) created by the user. CluVis uses a dual-domain approach that depicts the active chart on each node of a computer cluster or communication network in one view while displaying the relationships between explored charts in another view. Thus, CluVis facilitates a hypothesis-driven exploration of computer system metadata for visual analysis.

Categories and Subject Descriptors

I.3.8 [**Computer Graphics**]: Applications; C.1.4 [**Processor Architectures**]: Parallel Architectures—*Distributed architectures*; C.2.3 [**Computer-Communication Networks**]: Network Operations—*Network monitoring*

Keywords

security visualization, information visualization, visual analysis, clusters, networks, metadata

1. INTRODUCTION

The monitoring of clusters requires an understanding of both the operational behavior of a single node and any correlation of behavior between the other nodes in the cluster. In this aspect, the analysis of cluster behavior is analogous to the analysis of communication networks except on a smaller, more local scale. When working with clusters, there are a limited number of ways to obtain information about its overall performance or current behavior. A majority of these methods produce such a large volume of heterogeneous data that it becomes impractical to analyze the data without some form of post processing. In order to analyze this information, we introduce CluVis, an ongoing effort to design a dual-domain cluster metadata visualization system.

CluVis uses a hypothesis-driven exploration process for analyzing cluster and network metadata; the user interactively builds charts comparing metadata statistics to observe the cluster behavior. For each node in the cluster or similar network, this metadata includes the number of network packets incoming and outgoing, CPU utilization and idleness, disk usage, memory usage, and other similar statistics. Relationships between these values can be explored to ensure proper usage and discover anomalies. In addition, there can be correlations between several nodes at once—for example, if multiple nodes in the system are receiving a large number of incoming packets, it is possible that a single node in the system is flooding the network interconnect. Thus, to understand the cluster activity, both the metadata relationships of a single node and between multiple nodes must be understood. CluVis facilitates this by presenting a topology view of the cluster network and a view displaying statistics for a single node (Figure 1). All relevant information is consolidated into a single picture, instead of being distributed over several screens. This visualization is effective because of the human visual system's ability to quickly perceive patterns and outliers in graphical data [3, 22]. CluVis builds upon methods in both graph- and multi-dimensional visualization.

2. RELATED VISUALIZATION WORK

CluVis is a system for depicting relationships between multiple cluster metadata sources (*data streams*) over a single cluster node and across all nodes simultaneously. Thus, it is both a *graph visualization* system in that it depicts cluster connectivity via a graph and a *multi-dimensonal visualization* system in that it depicts multiple aspects of the cluster metadata concurrently. This section discusses related graph and multi-dimesional visualization work, especially in the context of security visualization.

Computer security data often consists of multiple data streams such as network traffic logs, user activity records, and similar. A common approach to depict this data is to use *multi-dimensional visualization*, a collection of tech-

*Corresponding Author

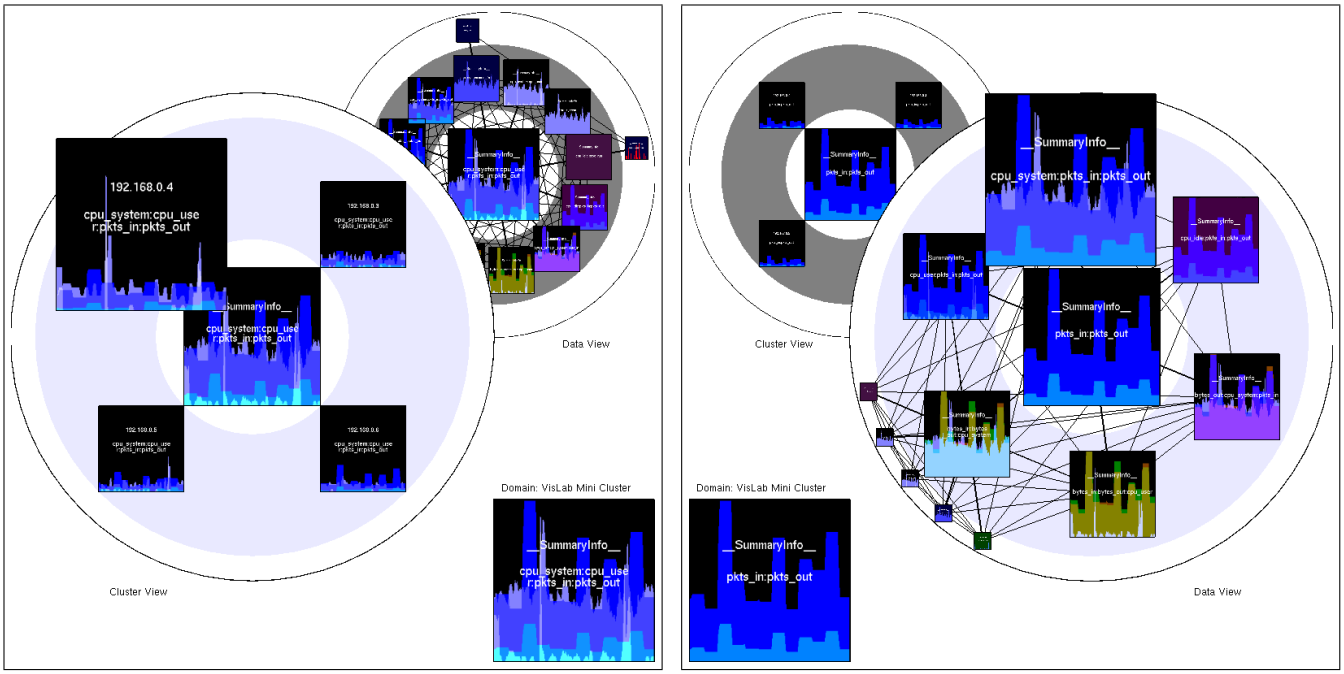

Figure 1: CluVis visualization of cluster metadata consisting of different data stream visualizations (charts). In the topology view, the active data streams are visualized on each cluster node. In the data view, all explored data stream combinations for the focused cluster node is displayed. In the topology view, links are based upon network connectivity; in the data view, edges are based upon similar metadata in a chart.

niques to display different variables of the security data simultaneously. Examples include the use of glyphs for visual intrusion detection data [4,10], multi-dimensional depictions of high-level [23,24] and packet-level [6] network traffic, and tools to directly monitor and audit computer [14, 20] and network [21] logs. All the aforementioned systems use customized visual aggregation methods to depict the multi-dimensional data; CluVis uses a different approach by only comparing two dimensions of the data at a time in a chart and allowing multiple charts to be explored. This approach is similar to the scatterplot matrix approach of Bertin [1], save that our charts are also connected via a graph.

Multi-dimensional visualization depicts the *contents* of the data; additional methods are needed to show *connections* between data. Connections in security visualization are common—computer networks are inherently a connected data structure. Connections are often depicted using *graph visualization* [8]. Examples of graph-based security visualization include visualization for network topology analysis [2,16,17] and applications to flaw/attack detection [5,15, 18,21]. Each uses graph visualization for the same purpose—to depict relationships. These relationships are between the nodes on the network, between network scans, or between packets sent to machines. But graph-only systems only tell part of the story; both the data and the relationships are needed for complete analysis. A few systems have combined multi-dimensional and graph visualization methods—e.g., VisAware's radial, multidimensional visualization [12] and TVN's comprehensive network traffic analyzer [7]. CluVis follows a similar approach, using two graphs in the visualization—one for network topology, the other for the relationships between depicted charts (Figure 1).

To the best of our knowledge, visualization research has not been specifically applied to small cluster visualization. As mentioned above, most of the above visualization systems focus on either the multi-dimsional data aspects or the connectivity aspects, not both. While many modern operating systems come with simple status monitors which display much of the metadata aspects we are interested it, these depictions do not consolidate the information cross-cluster nor depict the topology. Other monitoring tools, such as Ganglia [13] and RockSoft's Cluster Top [19] only allow users to compare similar cluster metrics. CluVis allows the comparison of metadata metrics over the entire network.

The closest extant system to CluVis is Fink et al.'s visual process monitor [5]. Fink et al.'s system uses a modified Linux kernel to correlate active processes with network traffic information for a better picture of ongoing activity. Fink et al.'s approach is both finer and larger grained than ours—CluVis does not focus on individual processes (it uses aggregate process information) and it is not designed for large or wide area network analysis (it focuses on smaller clusters). The two systems complement each other—CluVis can be used to determine the existence of unusual behavior across the cluster while Fink et al.'s can then be used to drill down to the specific processes causing the issue.

3. APPROACH: THE CLUVIS SYSTEM

There are three major components of the CluVis system: Its data collection mechanism, its dual-view visualization, and the interaction used for creating and exploring data charts. These are detailed in this section, with further examples presented in Section 4.

3.1 Data Collection

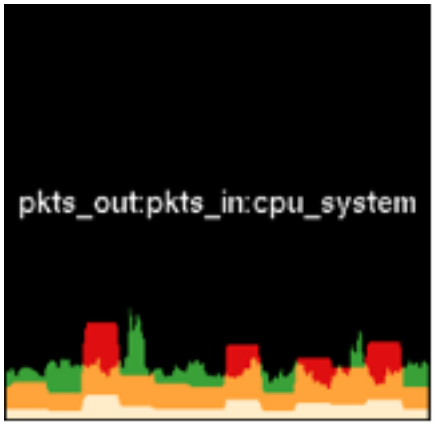

Figure 2: An example CluVis metdata chart. Three data series are visualized: packets in, packets out, and CPU system utilization.

Data is collected using Ganglia, a free and open source cluster metric collection and aggregation tool [13]; each metric consists of a particular data value measured over time (a data series). Our system builds upon the Ganglia data collection by organizing all of the metrics into files in a folder structure based on the cluster topology; this reports the overall statistics of the cluster. The data is stored at different temporal resolutions; in order to get a continuous series of data, CluVis keeps only the highest resolution data for a given time span. Each Ganglia metric file is assigned to one data series, and the data series are pooled and shared for the entire application.

Other data sources can be used in CluVis, so long as the data streams are identified in addition the the network topology. The second example in Section 4 demonstrates the use of CluVis to analyze the DARPA data set [11].

3.2 Visualization

The prevalent visual feature in CluVis is the usage of multiple user-created charts (Figure 2). Charts are one or more data series rendered onto a texture; a filled polyline chart is used for each data series displayed. Multiple data series rendered on top of each other are blended additively, which allows the user to quickly identify similar and dissimilar regions of the data. The cluster node id and the names of the used data series for a chart are rendered to the texture for identification. The texture is then cached and only regenerated when the data series change.

A chart represents one or more data streams from a given node. A chart is related to another chart if they share the same parent cluster node and at least one data series. Similarly, other cluster nodes could posses different charts showing the same collection of data metrics. These two forms of relationships are depicted via two graphs: One for the cluster topology and the second for the current cluster data (Figure 1). Two MoireGraphs, focus+context graph views (Figure 3), are used to depict this information; MoireGraphs were chosen because they allow visual content to be displayed at each graph node. At any one time, one of the graphs is in focus with the other slightly behind and offset from the foreground graph. The focused view is the only view that is considered for the interaction methods defined

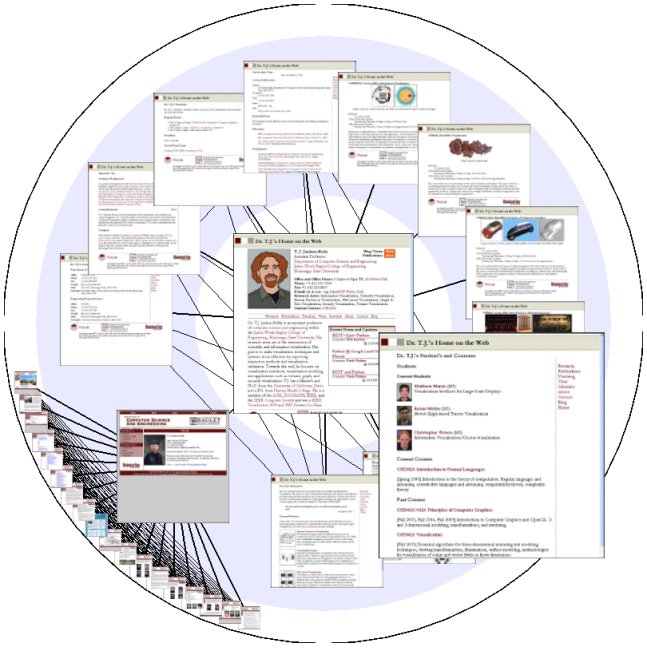

Figure 3: The MoireGraph visualization. The focus node is given the majority of the screen space; each subsequent level is given half the previous space. More screen space can be given to a user-specified level or node to facilitate node comparison (as shown) [9].

by the MoireGraph. The user switches between the two views by clicking on the unfocused view with the mouse; this triggers a smooth animated transition between the two focus states.

Each view focuses on different elements of the multi-dimensional cluster data. The topology view (Figure 1, left), positioned on the left side of the visualization, represents the topology of the cluster, with each graph node being a cluster computer. The graph nodes drawn in the topology view display the same metrics as the focus of the data view. The focus node on the topology view specifies the current cluster node of the visualization. The data view (Figure 1, right), positioned on the right side of the visualization, displays one graph node for each combination of data streams that have been explored. The graph nodes in the data view display the data from the cluster node specified by the focus of the topology view. Data view graph nodes that display metrics that are in a common category (CPU, Memory, Communication, Load, Disk) are connected. This means that nodes that display more than one category of metric are connected to two other nodes of both categories. Connecting the nodes in this way allows the user to identify correlations.

3.3 Interaction

There are three primary interactions a user may have with the CluVis system. First, they can change either the active view or the focus of the active view. Changing the view has been described previously; changing the active focus had different effects depending on the active view. If the topology view is active, changing its focus will change which cluster node is used for the data series displayed in the data view. If

the data view is active, changing its focus will change which collection of data series is charted for each cluster node on the topology view to match the new focus.

The second major interaction is chart construction. Charts may be added or removed from the graph. When a new chart is added, a list of all available data streams is presented to the user from which the user may pick and choose to add to the chart. In addition, each data stream can be assigned specific colors (interface not shown); the color will change for every currently displayed charts.

The final major interaction is the selection of the time span for the display. As previously mentioned, the cluster metadata may possess information from different time scales at different resolutions. By choosing which range of time to depict, the finest resolution data covering the entire range is displayed for the given charts. Each time a new time range is chosen, the entire dual-graph interface is updated.

There are other minor actions to facilitate exploration of the cluster metadata. MoireGraphs include interactions to enlarge non-focus nodes or entire graph levels for comparison purposes, to provide tool tips detailing additional information about the chart, and similar interactions [9]. In addition, the non-focused view may have its graph nodes rotated about its center in order to un-occlude charts that are behind the active view. This rotation also assists in comparison of that data.

4. EXAMPLES AND DISCUSSION

4.1 Cluster Analysis

CluVis is designed to facilitate iterative exploration of cluster metadata. The visualization summarizes the vital statistics of the cluster at a glance, and allows users to combine different statistics in order to assist hypothesis testing (e.g., *Is an attack causing CPU utilization to spike? Let's check the correlation of network traffic and processor usage.*). The data charts are rendered in such a way that the user can find trends among different metrics, even if the metrics are of completely different categories. The dual graph layout allows the user to analyze multiple machines at the same time, which can help show if certain behavior is commonplace across the cluster. The chart construction interface allows the user to customize the visualization to their needs. These, combined with the basis of the graph connectivity, allow the user to explore the cluster's activity.

Figure 1 shows the visualization of Ganglia data collected from a four-node cluster over a few days. In the data view of the cluster (right), we can see that there is an abrupt increase in system CPU usage (dark blue in the enlarged graph node). Focusing on a chart with system CPU usage in it and switching to the topology view (left) shows us that the increase only occured on one machine (192.168.0.4, enlarged graph node). This straightforward example shows how effectively relationships can be extracted from the cluster data.

4.2 LAN Analysis

To assess the use of CluVis for visualizing network traffic metadata, the DARPA Instrusion Detection Evaluation Data Set was examined [11]. The 1998 training data TCP packet dumps were pre-processed into data streams of packet frequency and size for each IP address in the data set. Due to the data's large size, the topology is based on the bytes of the IP addresses; addresses of child nodes n levels from the root begin with the same n bytes; non-leaf nodes contain the union of its child addresses.

To demonstrate the iterative CluVis exploration process, we start by looking at the summary of incoming traffic (Figure 4, left). Due to the large number of IP addresses used in the data, only the "inside" IP addresses and a few "outside" IP addresses from the off-line evaluation are being used. From the topology visualization, it clear to see that IP addresses beginning with 172 received the highest amount of incoming traffic (enlarged node). Following the topology, we find that all of the traffic was directed at a single computer: 172.16.114.50 (Marx).

With Marx as the topology focus, we switch to the data view to understand the traffic through that machine (Figure 4, right). Looking at the incoming traffic data for Marx, there are five major "spikes" of traffic that occur (center node). By adjusting the chart time slider and comparing the times to known attacks in the data, we find that the first, smaller spike is a smurf attack on the machine. The second and third spikes are from two larger smurf attacks that occured in the fifth week, the second of which originating from "all attackers". The charts show that this attack created more inbound traffic for Marx than all other attack during the seven weeks. The final two spikes, occuring one week later, also exhibit the effects of two more smurf attacks.

This example demonstrates the hypothesis-driven exploration process that a user can follow while exploring with CluVis. Both views can be used to discover relationships between different sources and the data metrics across those sources. For instance, the early conclusion that Marx had the most incoming traffic was facilitated by the available visual comparison between sources in the topology view.

5. SUMMARY

CluVis is a cluster/network metadata visualization system for monitoring and security analysis. A data and topology view are presented in order to explore data at a specific machine level and at a cluster/network-wide level respectively. This dual-domain facility and the ability to iteratively generate visualizations as new hypotheses are formed contribute to CluVis' utility.

5.1 Future Work

Development of CluVis is ongoing; there are several opportunities for future development. Foremost, the scalability and functionality of CluVis should be tested empirically. Only via user studies can we gain a measurable understanding the effectiveness of CluVis, especially as the number of cluster machines and visualized data streams increases. The MoireGraph system has been used effectively for graph of approximately 300 nodes, so we suspect that similar performance could be gained for CluVis. For clusters and data views of larger size, grouping mechanisms could be used.

Other future work includes the exploration of other forms of visualizations besides simple charts. Other multi-dimensional visualizations such as scatter plots or parallel coordinates may be worthwhile to depict other forms of patterns or correlations amongst the cluster metadata. Another visual enchancement would be to add a threshold line to the data charts to help augment the visual comparison of charts.

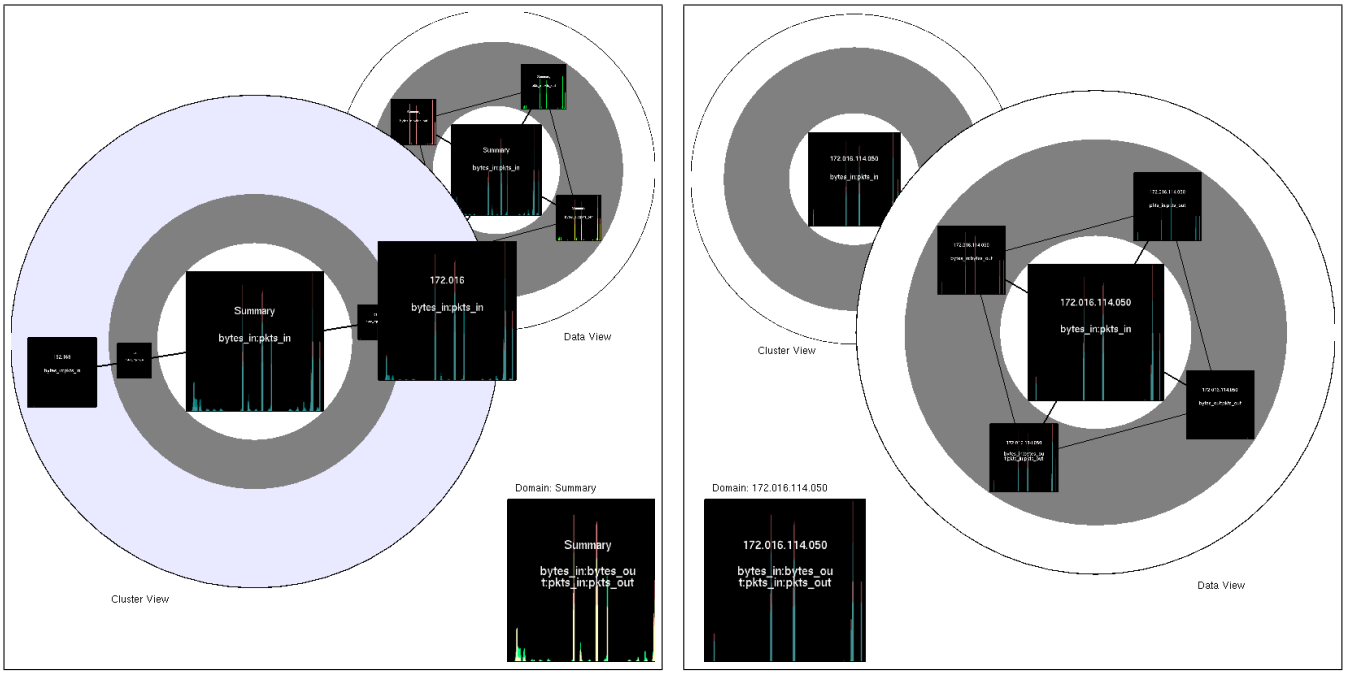

Figure 4: 1998 DARPA Intrusion Detection Evaluation Training Data Set Exploration

Acknowledgments

This work was funded in part by the Mississippi State University Office of Research and the Center for Computer Security Research via a grant from the National Security Agency. Joseph Langley and Benjamin Craig assisted in the development of the original CluVis prototype.

6. REFERENCES

[1] J. Bertin. *Graphics and Graphic Information-Processing*. de Gruyter, 1981.

[2] B. Cheswick, H. Burch, and S. Branigan. Mapping and visualizing the internet. In *Proc. of the 2000 USENIX Annual Techincal Conf.*, 2000.

[3] A. D'Amico and M. Kocka. Information assurance visualizations for specific stages of situational awareness and intended uses: Lessons learned. In *Proc. of VizSec 2005*, pages 107–112, 2005.

[4] R. F. Erbacher, K. L. Walker, and D. A. Fincke. Intrusion and misuse detection in large-scale systems. *IEEE Computer Graphics and Applications*, 22(1):38–48, January/February 2002.

[5] G. A. Fink, P. Muessig, and C. North. Visual correlation of host processes and network traffic. In *Proc. of VizSec 2005*, pages 11–19, 2005.

[6] L. Girardin. An eye on network intruder-administrator shootouts. In *Proc. of the Work. on Intrusion Detection and Network Monitoring (ID'99)*, 1999.

[7] J. R. Goodall, W. G. Lutters, P. Rheingans, and A. Komlodi. Preserving the big picture: Visual network traffic analysis with TN. In *Proc. of VizSec 2005*, pages 47–54, 2005.

[8] I. Herman, G. Melançon, and M. S. Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Trans. on Visualization and Computer Graphics*, 6(1):24–43, January-March 2000.

[9] T. J. Jankun-Kelly and K.-L. Ma. Moiregraphs: Radial focus+context visualization and interaction for graphs with visual nodes. In *Proc. of the 2003 IEEE Symp. on Information Visualization*, pages 59–66, 2003.

[10] A. Komlodi, P. Rheingans, U. Ayachit, J. R. Goodall, and A. Joshi. A user-centered look at glyph-based security visualization. In *Proc. of VizSec 2005*, pages 21–28, 2005.

[11] R. Lippmann, R. K. Cunningham, D. J. Fried, I. Graf, K. R. Kendall, S. E. Webster, and M. A. Zissman. Results of the DARPA 1998 offline intrusion detection evaluation. In *Recent Advances in Intrusion Detection*, 1999.

[12] Y. Livnat, J. Agutter, S. Moon, and S. Foresti. Visual correlation for situational awareness. In *Proc. of the IEEE Symp. on Information Visualization 2005*, pages 93–102, 2005.

[13] M. L. Massie, B. N. Chun, and D. E. Culler. The Ganglia distributed monitoring system: design, implementation, and experience. *Parallel Computing*, 30(5-6):817–840, 2004.

[14] J. McPherson, K.-L. Ma, P. Krystosk, T. Bartoletti, and M. Christensen. PortVis: A tool for Port-Based Detection of Security Events. In *Proc. of VizSec/DMSEC 2004*, 2004.

[15] C. Muelder, K.-L. Ma, and T. Bartoletti. A visualization methodology for characterization of network scans. In *Proc. of VizSec 2005*, pages 29–38, 2005.

[16] T. Munzner. Exploring large graphs in 3d hyperbolic space. *IEEE Computer Graphics and Applications*, 18(4):18–23, July/August 1998.

[17] T. Munzner, E. Hoffman, K. Claffy, and B. Fenner. Visualizing the global topology of the mbone. In *Proc. of the 1996 IEEE Symp. on Information Visualization*, pages 85–92, 1996.

[18] S. Noel, M. Jacobs, P. Kalapa, and S. Jajodia. Multiple coordinated views for network attack graphs. In *Proceedings of VizSec 2005*, pages 99–106, 2005.

[19] P. M. Papadopoulos, M. J. Katz, and G. Bruno. NPACI Rocks: Tools and Techniques for Easily Deploying Manageable Linux Clusters. *Concurrency and Computation: Practice and Experience*, 15(7-8):707–725, 2003.

[20] T. Takada and H. Koike. Tudumi: Information visualization system for monitoring and auditing computer logs. In *Proc. of the 6th Intl. Conf. on Information Visualization*, 2002.

[21] S. T. Teoh, T. J. Jankun-Kelly, K.-L. Ma, and S. F. Wu. Visual data analysis for detecting flaws and intruders in computer network systems. *IEEE Computer Graphics and Applications*, 24(5):27–35, Sep/Oct 2004.

[22] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, 2000.

[23] X. Yin, W. Yurcik, M. Treaster, Y. Li, and K. Lakkaraju. Visflowconnect: netflow visualizations of link relationships for security situational awareness. In *VizSEC/DMSEC '04: Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*, pages 26–34, 2004.

[24] W. Yurcik, K. Lakkaraju, J. Barlow, and J. Rosendale. A prototype tool for visual data mining of network traffic for intrusion detection. In *Proc. of the DMSEC'03*, 2003.