**Cyber Security Data Visualization**

**Group:**

Christopher Storey and Victor Wu

**Contact Info:**

[cstorey1@umbc.edu](mailto:cstorey1@umbc.edu)   
[tr59337@umbc.edu](mailto:tr59337@umbc.edu)

**Abstract:**

The visualization of large volumes of network traffic is a challenging research problem. The purpose of this project is to display network data in ways that minimize clutter while maintaining the source and destination relationships. To address this issue, we plan on creating a CloudLines graph to display our data.

**Motivation:**

The visualization of large volumes of network traffic is a challenging research problem. We are doing this project because we are interested in cyber security and because the application of effective visualization techniques to the field of cyber security is a hot research topic. Novel approaches to providing overview clues within a large streaming corpus are needed to help users grasp the big-picture and to spot trends and anomalies. This is important because practitioners often find it challenging to display large volumes of data in ways that allow users to extract meaning and insight.

**Research questions:**

We will be addressing source-destination relationships of network traffic and how to display the data in a way that minimizes clutter while maintaining the relationship between nodes.

**Preliminary hypotheses:**

We believe that our visualization tool will require us to translate categorical and temporal attributes of network flow data into visual patterns to enable users to sift through gigabytes of traffic. This translation will use color and scale to add dimensionality and clarity, resulting in minimization of graphic density and occlusion.

**Hardware and Software:**

To implement our visualization, we will be creating a web application using javascript and D3 libraries to display data.

**Tasks and Metrics:**

We will be using a CloudLines graph to represent the relationship between source and destination nodes in network traffic.

**Literature Review**

The application of visualization techniques to the field of cyber security is an active area of research given the complexity, large volumes of data, and speed with which security professions have to sift through the hundreds of signals present in typical network traffic. For a security practitioner to be successful, one has to readily identify actionable patterns within the data. Over the years, many techniques have been proposed, but effective idioms to reliably visualize network security data continue to elude the community. This paper attempts to review the effectiveness of several visualization examples from the literature that illustrate good and bad approaches to this problem.

## **Overview**

(Munzner, 2015) proposes eight rules of thumb for effective use of visualization techniques to convey meaning to views of various forms of data. These rules of thumb come from the author’s personal experience, as well as from empirical human factors studies of idioms in a variety of use cases. The eight rules of thumb are: 1) No Unjustified 3D; 2) No Unjustified 2D; 3) Eyes Beat Memory; 4) Resolution over Immersion; 5) Overview First, Zoom and Filter, Detail on Demand; 6) Responsiveness is Required; 7) Get It Right in Black and White; and 8) Function First, Form Next. Often there are trade offs in the use of each form of representation that have been identified by researchers. This review will examine each example in terms of its effectiveness and the tradeoffs relative to the use case.

## **Comparative Review**

(Abdullah, Lee, Conti, & Copeland, 17-19 June 2002) discusses the use of stacked histograms of aggregate port activity to depict network flow in a way that aids administrators in recognizing attacks in real time. The authors use drill-down techniques to selectively provide finer details without obscuring routine display of data. This paper builds upon earlier work done by Becker et al (1995) and Cox et al (1995 using glyphs to represent nodes and lines to represent links. Furthermore it expands upon Erbacher (2002) in the use of glyphs and nodes where line type and circle attributes show information about the system and link. Lastly, it expand Flodar’s (1998) and SecureScope’s (2003) approach to visualize the layout of a network, with emphasis on identifying statistics of the network and trends over time. The authors implement a prototype that effectively used stacked histograms for security purposes. The techniques proposed handled detecting anomalies quickly and effectively solved issues with scaling by using for port numbers, IP addresses and time.

(Conti, 2007) is a survey book that examines key aspects of the security visualization problem. It provides an easy introduction for practitioners that are new to the field. The book contains great introductory discussion of relevant open source visualization application such as AfterGlow, EtherApe and VisFlowConnect. Additionally, it provides a list of classical visualization research papers and books that could be used for further study. The author builds upon prior work to explain in detail the challenges inherent in visualizing large network datasets. In Chapter 7, Firewall Log Visualization, he explores the use of color coding and hierarchical link graphs to depict Firewall and intrusion detection data. This discussion is effective in illustrating the pitfalls of occlusion. In Chapter 8, Intrusion Detection Log Visualization, he discusses the many idioms that could be applied to large network datasets and uses Treemaps and an illustrative example. His discussion of drill down in this chapter provides a great example of how to effectively guide the user to key information. In Chapter 9, Attacking and Defending Visualization Systems, he surveys the many systems that have attempted to address the difficulty of coping with large datasets with the overall hope that the reader would be prompted to explore novel techniques based upon viewing the failures of others. In his exploration, he explores the difficulty of dealing with labeling, occlusion, and scaling. All of these overarching problems areas distort or clutter the display and make visualization of large amounts of network data ineffective in most visualizations today.

(Milos Krstajic, 2011) proposes the novel use of CloudLines to visualize time-series data. The key benefit identified by this approach is the ability to display large amounts of data in a limited space. CloudLine plots are constructed of individual circles, representing event, that are then placed temporally along a timescale. An obvious application of this technique is to visualize network or cyber security data. The author’s attempt to overcome many of the trade offs that could potentially sideline this approach as an effective means to display large amounts of data. For example, in early versions of this technique, they recognized the difficulty that some users have distinguishing between events that are depicted very close together. Munzer calls this potential pitfall occlusion – where some objects cannot be seen because they are hidden behind others. As a result of this challenge, a lens-based interaction was added to allow direct access to these events. Likewise, opacity was added to represent the relative density of plots along the linear line. This density translates into areas of interest that might be of interest to the reader.

All of the above design choices make CloudLine plots effective representations of time-series data. Figure 1 was taken from (Milos Krstajic, 2011). The most noticeable elements of the visualization are the contrasts that are inherent between the temporal lines representing each political candidate. There does not appear to be any essential information missing, nor elements, which carry no information. Colors are used effectively to distinguish between the political candidates, while opacity is used to highlight areas of intensity. Because all of the elements used within this chart have been tailored to a particular purpose, this visualization does not encourage any misperceptions.

The CloudLine plots as depicted in this article could possibly be improved if they were integrated hierarchically with other idioms. For example, drill down functionality could be added provide a layered view of the data. This would augment the existing lens-based interaction, while improving the user’s ability to overcome occlusion. The implementation of this effect must guard against creating a cognitive load on the viewer when switching from one view to the other.

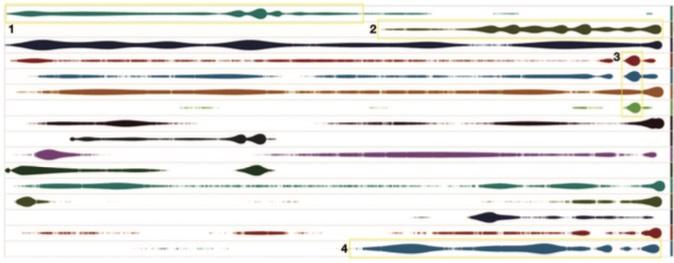


Figure 1: CloudLines visualization showing the relative number of appearances in the news media by sixteen different politicians during the month of February 2011.

(Florian Mansmann, 2007) extend concepts from logistics and resource planning to the visual analysis of network traffic. This article explores the use of various types of Treemaps – space-filling layouts of nested rectangles. The authors conclude that a geographically based Treemap called a HistoMap should be used to depict the top layers of the network and provide an overview for the user. Additionally this top-level view provides scalability, squareness, and stability benefits. For lower levels of the network hierarchy, the authors believe that Strip Treemaps are most appropriate due to their maintaining the input order of nodes (Florian Mansmann, 2007, p. 1111).

(Munzner, 2015) discusses the pros and cons of Treemap extensively. On the one hand, the idiom of treemaps can effectively show hierarchy relationships. However, they are most effective for hierarchies that are shallow rather than deep (Munzner, 2015, p. 214). The application of Treemaps to the representation of internet traffic at the TCP/IP level is therefore a mistake, given the many layers of abstraction from Automated Systems at the top layers, down to the individual IP addresses near host systems. The most noticeable elements of the visualization depicted in Figure 2 are the boundary lines that are allocated to each node. These relationships are the most important. There is a lot of information missing in this view because of the scale. It is extremely difficult for the view to see detail down to the granular level of IP addresses in this view. Most of the diagram provides no information. Color is used effectively to show intensity or the number of connections from low (blue) to high (red). It is possible that a viewer could misperceive the importance of a particular node or activity based upon the scale of the hierarchy. In the domain of internet security, even the small nodes are important.

The Treemap plots as depicted in the article could possibly be improved if they were combined with some additional algorithms that would assist in lessening the cognitive load. According to Munzner, it might be helpful, given the size of the depicted dataset, to create a useful overview for top-down exploration. This would allow the user to find context within the data and starting points for further analysis.

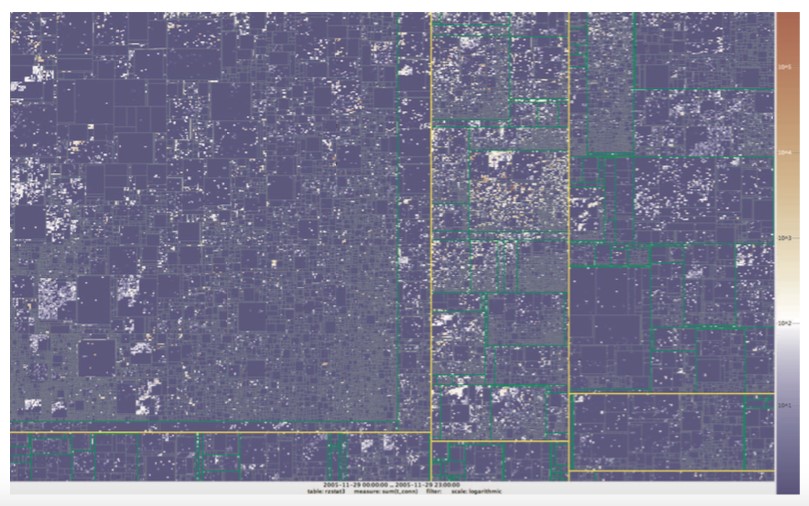


Figure 2: A Treemap visualization of outgoing traffic connections through a university gateway on November 29th, 2015, showing 197427 IP prefixes.

(Marode & Chavan) is a survey paper that examines how to translate large amounts of network traffic logs (in text based data) into visual patterns (animation) with the goal of enabling network administrators or security officers to sift through network traffic more effectively. The authors explore the prior art and point to insight that show promise or that could be expanded upon. Of particular note, the use of traffic header data for anomaly detection, the use of data mining techniques with parallel coordinates to visualize big network traffic data, and the use of a neural network to reduce dimensionality within the space of network and logging information into a self-organizing map are all implementation areas that should be explored further. Additionally, the author’s presentation of 24 network visualization tools that are under development by various organizations shows the depth of ongoing research in this area. The author’s specifically call out the use of parallel coordinates, bundled views, and 3-dimensional visualization as promising network visualization techniques.

(Meeks, 2015) provides a quick overview and introduction to the power of the D3.js JavaScript library in implementing the many visualization idioms discussed in the literature. The author provides detailed coding examples to demonstrate how to implement most major charting and visualization types within D3.js. In particular to study, Chapter 6, Network Visualization, focuses on representing networks as nodes, edges and links. Chapter 11 of this book, Big data visualization, focuses on techniques to create data visualization with large amounts of data.

## **Conclusions**

By grouping rules of thumb about the current state of knowledge on visualization techniques, Munzner provides a service. Using the taxonomy she provides, it is now easier to examine visualizations from a more holistic viewpoint and is an aim of this course. The literature discusses a broad listing of idioms that could be applied to cyber security data. Exploring the trades that can be made between different design elements are useful excursions and contain suggestions for future research. This team will attempt to expand upon these idea in the course design project.

# **Bibliography**

Abdullah, K., Lee, C., Conti, G., & Copeland, J. A. (17-19 June 2002). Visualizing Network Data for Intrusion Detection. *Proceedings of the 2002 IEEE Workshop on Information Assurance and Security*, 30-38.

<<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.106.5388&rep=rep1...>>

Becker, R. A., Eick, S. G., & Wilks, A. R. (March 1995). Visualizing Network Data. *IEEE Transactions on Visualization and Computer Graphics*, 16-21.

<<http://igva2013.wikispaces.asu.edu/file/view/visualizingNetworkData.pdf>>

Conti, G. (2007). *Security Data Visualization: Graphical Techniques for Network Analysis.* San Francisco, CA: No Starch Press, Inc.

Florian Mansmann, D. A. (2007). Visual Analysis of Network Traffic for Resource Planning, Interactive Monitoring, and Interpretation of Security Threats. *IEEE Transactions on Visualization and Computer Graphics* *, 13* (6), 1105-1112.

<<http://www.ncbi.nlm.nih.gov/pubmed/17968053>>

Krstajic, M., Bertini, E., & Keim, D. A. (n.d.). CloudLines: Compact Display of Event Episodes in Multiple Time-Series. 2432-2439.

<<https://kops.uni-konstanz.de/bitstream/handle/123456789/17483/Keim.pdf?sequence=1>>

Mansmann, F., Keim, D. A., North, S. C., Rexroad, B., & Sheleheda, D. (2007). Visual Analysis of Network Traffic for Resource Planning, INteractive Monitoring, and Interpretation of Security Threats. *IEEE transactions on visualization and computer graphics 13*.

<<https://kops.uni-konstanz.de/bitstream/handle/123456789/5472/InfoVis2007NetworkSecurity.pdf?sequence=1>>

Marode, D. K., & Chavan, R. K. (n.d.). Survey of Network Traffic Visualization Techniques. *International Journal Computer Technology & Applications, Vol 5 (3)*, 876-883.

<<http://www.ijcta.com/documents/volumes/vol5issue3/ijcta2014050311.pdf>>

Meeks, E. (2015). *D3.js In Action.* Shelter Island, NY: Manning Publications Co.

Munzner, T. (2015). *Visualization Analysis & Design.* Boca Raton, FL: CRC Press, Taylor & Francis Group.