4-25-2017

Retrospective on a Decade of Research in Visualization for Cybersecurity

R. Jordan Crouser
Smith College, jcrouser@smith.edu

Erina Fukuda
Smith College

Subashini Sridhar
Smith College

Retrospective on a Decade of Research in Visualization for Cybersecurity

R. Jordan Crouser*
Smith College

Erina Fukuda†
Smith College

Subashini Sridhar‡
Smith College

Abstract

Over the past decade, the visualization for cybersecurity (VizSec) research community has adapted many information visualization techniques to support the critical work of cyber analysts. While these efforts have yielded many specialized tools and platforms, the community lacks a unified approach to the design and implementation of these systems. In this work, we provide a retrospective analysis of the past decade of VizSec publications, with an eye toward developing a more cohesive understanding of the emerging patterns of design at work in our community. We identify common thematic groupings among existing work, as well as several interesting patterns of design around the utilization of various visual encodings. We also discuss existing gaps in the adaptation of information visualization techniques to cybersecurity applications, and recommend avenues for future exploration.

1 Introduction

Over the past decade, the visualization for cybersecurity (VizSec) research community has featured over 160 publications through workshops, symposia, and more recently at an annual conference sponsored by the Institute of Electrical and Electronics Engineers (IEEE). Through these works, the community has demonstrated the development and adaptation of numerous information visualization techniques to support the critical work of cyber analysts. These efforts have yielded many novel, sometimes highly-specialized tools and platforms, as well as more theoretical contributions such as task taxonomies. Despite this rich body of work, the relative youth of this community means that we lack a unified approach to the design and implementation of these systems.

In this work, we provide a retrospective survey and analysis of the past decade of VizSec publications, with an eye toward developing a more cohesive understanding of the emerging patterns of design at work in our community. We take a multi-pronged approach in order to provide a comprehensive reflection on the current state of VizSec research, utilizing methods ranging from text mining to the application of existing task analysis frameworks. We identify common thematic groupings among existing work, as well as several interesting patterns of design around the use of various visual encodings. We also discuss existing gaps in the adaptation of information visualization technologies for use in cybersecurity applications, and recommend avenues for the development of future systems.

2 Automated Analysis via Text Mining

In order to get a high-level overview of the state of the practice within the VizSec community, we wanted to begin with an approach that would introduce as little of our own assumptions and bias as possible. We therefore decided to conduct a preliminary analysis via automated text mining on the full body of work that has been published at various venues focusing on visualization for cybersecurity. We acknowledge that this is not a comprehensive survey of all work related to the utilization of visualization in support of cybersecurity. Indeed, this analysis is not intended to provide an authoritative digest of the topic. Rather, we hope that through a close examination of our own peer-reviewed publication practices, we might have the opportunity to reflect more deeply on our history and values as a community.

2.1 Dataset

The analysis reported in the remainder of this section was conducted on a collected a corpus of 161 papers published in IEEE visualization for cybersecurity between 2004 and 2015. We preprocessed each article by first extracting the raw text, and then gathering associated metadata.

2.2 Preprocessing and Computing Pairwise Distance

Because we wish to be able to group similar publications together, we first need to identify an intuitive distance metric by which we will define "similarity". For simplicity we'll begin by using a simple bag-of-terms model, in which we simplify each publication down the number and frequency of unique terms it contains [28]. We consider single words, as well as *bigrams* (phrases of length 2) and *trigrams* (phrases of length 3)[1] for this analysis. Using this model, each document can then be represented as as a vector of length k, where k is the number of unique terms across all publications in the corpus. We can then readily compare between these high-dimensional vectors in order to get a rough sense of how similar two publications are to one another.

Unfortunately, because some terms naturally appear more frequently than others, the above approach will tend to over-emphasize frequently-used terms with low information content. For example, terms such as "see Fig." or "as seen in" (which appear frequently across all scientific publications regardless of topic) are not particularly useful for our purposes. To mitigate this, it is useful to first eliminate commonly-used English words[2], and then to perform some normalization specific to our publication corpus.

Rather the considering the pure frequency of each term, we utilize a measure known as Term Frequency-Inverse Document Frequency, or `tf-idf` [26]. This normalization amplifies the weight of terms that more uniquely distinguish certain documents, and minimizes the contribution of terms that occur commonly across the full corpus. We utilized the `TfidfVectorizer` module from the Python $scikit-learn$ toolkit for efficient computation of the `tf-idf` vector for each publication. We discarded any terms that appeared in more than 80% or fewer than 10% of the publications in the corpus, leaving 2,369 unique terms. We then compute the *cosine similarity* of each pair of documents (characterized by their

*e-mail: jcrouser@smith.edu

†e-mail:efukuda@smith.edu

‡e-mail:ssridhar@smith.edu

[1] We do not consider n-grams of length > 3, as information theoretic models of the English language suggest that n-grams of greater length do not generally yield more effective results [24].

[2] We eliminate all terms contained in `nltk`'s English stopwords library, available at: `http://www.nltk.org`

978-1-5090-6356-7/17/$31.00 ©2017 IEEE

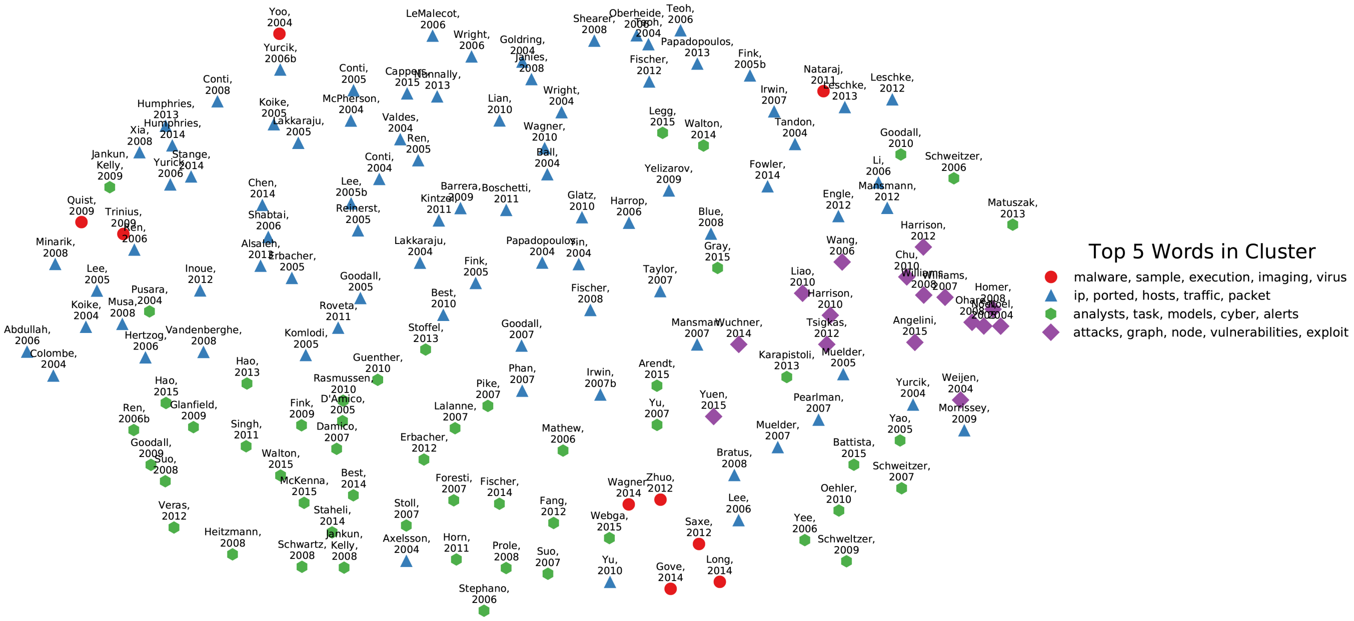


Figure 1: MDS projection of k-means clustering ($k = 4$) of 161 VizSec papers spanning the years 2004-2015. Distance between publications is calculated using TF-IDF vectors constructed from single words, bigrams, and trigrams.

respective `tf-idf` vectors), and use this to construct a complete pairwise distance matrix over all 161 publications.

2.3 Results

We can now use this pairwise distance matrix to explore how the past decade of VizSec publications has unfolded. If we perform k-means clustering with $k = 4$, we see clear separation between thematic groups (see Fig. 1):

- In the **blue** cluster (85 papers), we find tools for cyber situational awareness such as VisFlowConnect [38, 41], NVisionIP [18, 19, 40], and NVisionCC [42] alongside work by Conti et al. [5, 6, 7] in using visualization low-level features to identify malicious activity.
- In the **green** cluster (51 papers), we find many higher-level frameworks to organize the space and process of designing visualization systems for cybersecurity applications, including work by Jankun-Kelly et al. [17], Staheli et. al [22, 30], and Suo et al. [31, 32].
- In the **purple** cluster (16 papers) we find systems which exploit hierarchical or graph-theoretic structure in order to identify vulnerabilities within a network, such as work by Harrison et al. [14, 15] and Williams et al. [36, 37].
- And finally, the **red** cluster (9 papers) consists of work in the area of malware analysis [11, 21, 23, 25, 27, 33, 34, 39, 43].

Upon closer inspection of the MDS projection in Fig. 1, we notice that the ***malware analysis*** cluster appears to be somewhat scattered. If we inspect the metadata associated with each paper (such as the date of publication), the reason becomes clear: the rapidly evolving landscape of malware analysis and the persistent threat of novel software exploits introduces significant differences in the terminology used in publications on this topic year-over-year.

This observation begs the question: does this analysis capture other interesting trends over time? When we further subdivide each cluster into bins for each publication year, an even richer story begins to unfold (see Fig. 2). In this view, we can see the gradual shift away from low-level ***forensic analysis***, which was a primary focus in the early years of utilizing visualization in support of cybersecurity. After a 6-year heyday averaging 10 papers per year under this umbrella, we see a slow taper as the community begins to turn its attention toward providing higher-level support in areas such as ***situational awareness*** and ***network defense***. We observe an increased interest in the use of models, both in the context of better understanding the analysts' tasks as well as developing a clearer picture of the connectivity of the network through the use of graph theory.

3 Application of Existing Taxonomies

We next explored whether there were any interesting patterns in the kinds of visualizations or visual metaphors the community was using across these categories. We first distilled the corpus down to a subset of 87 systems papers published at VizSec in the most recent 8 years. Because automated classification is a fallible process, the authors of this paper conducted a manual evaluation of the assigned

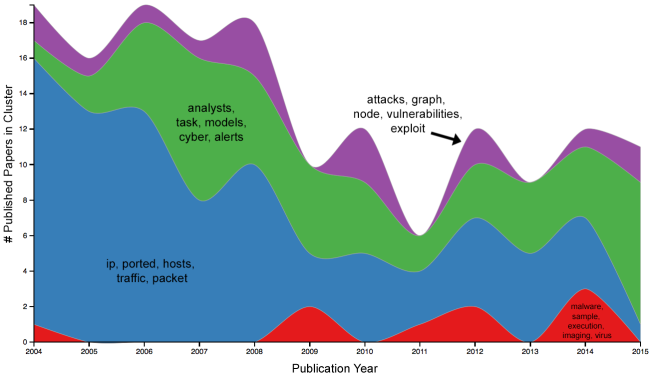
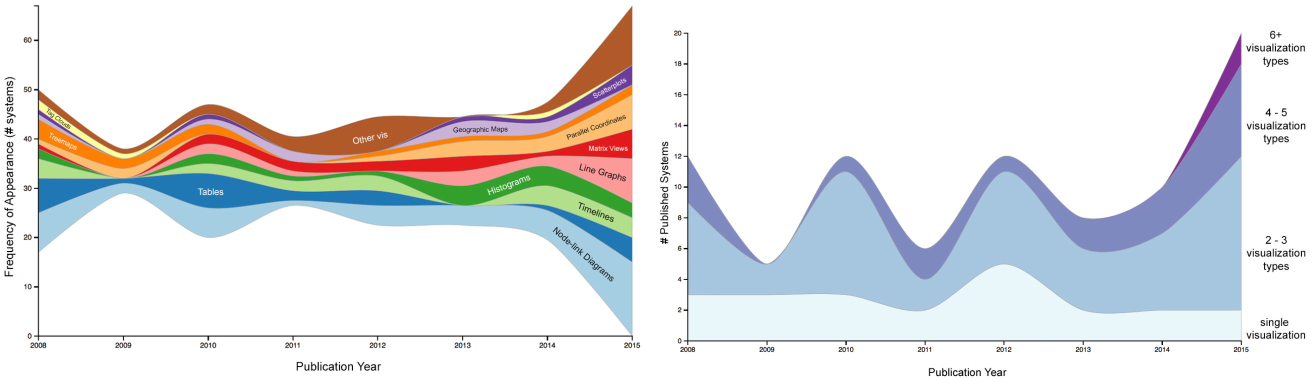


Figure 2: Distribution of the 4 automatically-generated clusters of VizSec publications (***forensic analysis***, ***situational awareness***, ***network defense***, and ***malware analysis***) from 2004 to 2015.

978-1-5090-6356-7/17/$31.00 ©2017 IEEE



(a) Utilization of various classes of visual metaphor.

(b) Utilization of multiple visual metaphors in a single interface.

Figure 3: Temporal trends in the utilization of visualization types in VizSec publications from 2008 to 2015. Note not only the change in utilization of visual metaphors such as *parallel coordinates* and *treemaps* (left panel), but also the increasing tendency to utilize multiple complimentary metaphors in a single interface (right panel).

labels and manually corrected any obvious misclassifications, performing partial cross-validation on a random subsample to ensure internal validity. From this, we further restricted our in-depth analysis to the 78 papers categorized as ***forensic analysis* (30)**, ***situational awareness* (27)** or ***network defense* (23)**.

3.1 Visualization Type(s)

Drawing on taxonomies by Shneiderman [29], Chi [4], and Duke University [44], we identified 11 high-level types of visual mapping techniques commonly employed by the VizSec community: *node link diagrams* (46), *tables* (26), *timelines* (19), *matrix views* (17), *parallel coordinates* (17), *bar charts / histograms* (17), *line graphs* (17), *treemaps* (13), *geographic maps* (9), *scatterplots* (8), and *word clouds* (4). For each published system, we identified which of these high-level classes of visualization were utilized (see Fig. 5 for selected examples). When more than one visual type of visualization was used, we also noted whether the views were *coordinated* or *independent*. In the event that a view could not be readily classified into one of the above 11 groupings, it was labeled *Other Visualization Type* (29).

We again begin by first examining how the VizSec community has made use of various visual metaphors over time (see Fig. 3). In the left hand panel of Fig. 3, note the change in utilization of visual metaphors such as *parallel coordinates* (increasing beginning in 2012) and *treemaps* (slowly decreasing after 2010). Additionally, we observe in the right hand panel of Fig. 3 an **increasing tendency to utilize multiple complimentary metaphors** in a single interface.

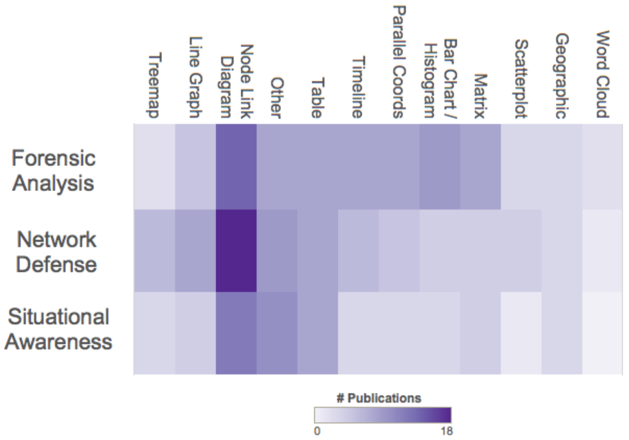


Figure 4: Utilization of visual metaphors across 3 analytic goals. Note the difference in use of *matrix views* versus *node link diagrams* in **forensic analysis** and **network defense**.

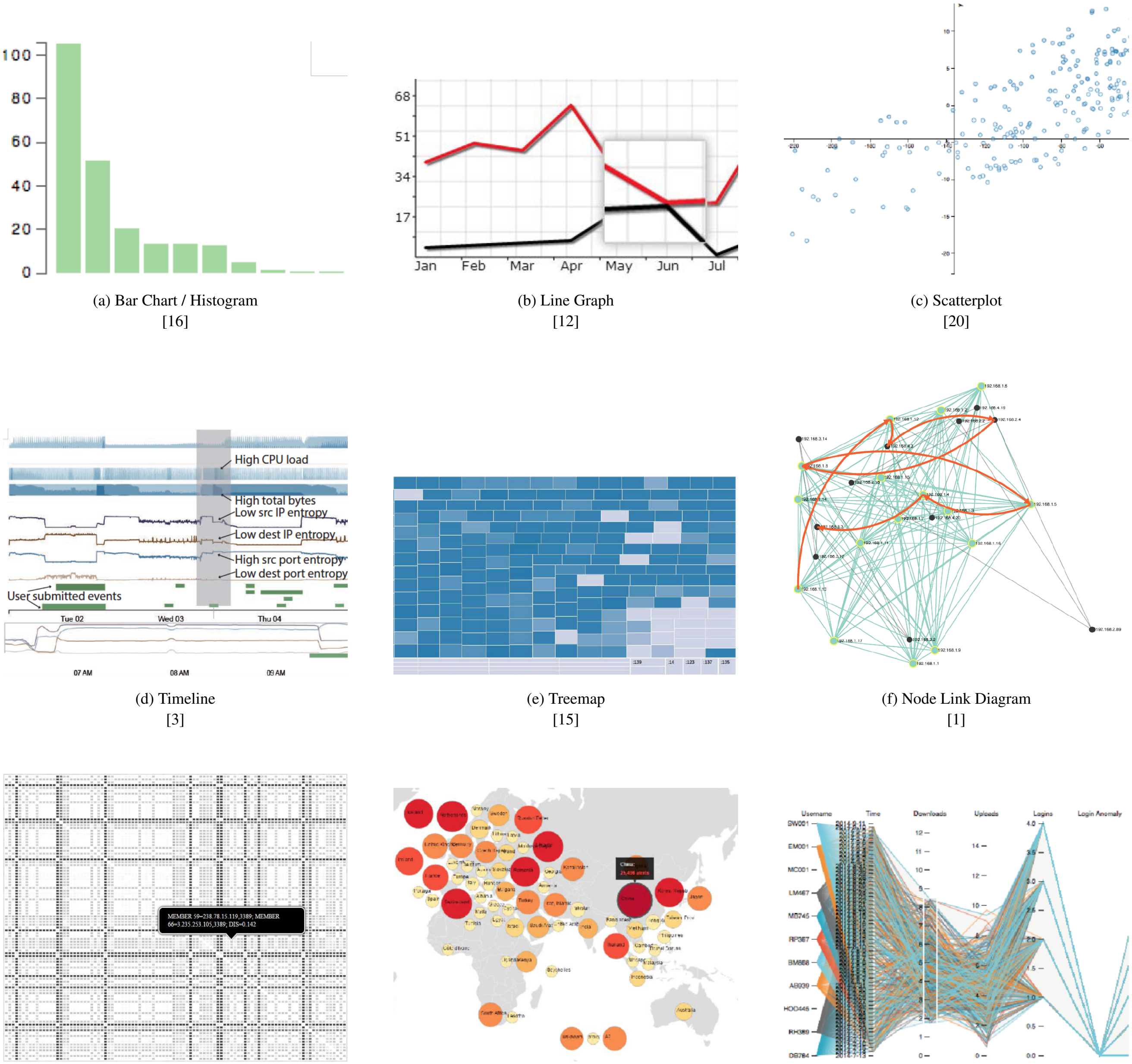
If we then examine the distribution of these various visual metaphors across the ***forensic analysis***, ***situational awareness*** or ***network defense*** classes, some interesting patterns of design begin to emerge (see Fig. 4). For example, we notice that the use of *geographic views* and *tables* is relatively consistent across all categories. Intuitively, geographic maps are also employed relatively infrequently due to the fact that logical topology is often more informative than physical topology. We also observe a dramatic difference in the utilization of *matrix views* versus *node link diagrams* between the ***forensic analysis*** and ***network defense*** classes. This suggests that these views may provide different affordances [8], providing opportunities for further exploration.

4 Future Opportunities: Streaming Data

In their 2014 paper of the same name [2], Best et al. identified "7 Key Challenges for Visualization in Cyber Network Defense" (paraphrased): *increasing data volume, heterogeneity of data sources, isolated / non-linked data sources, poor data quality, cadence / temporal scale, progression of threat escalation*, and *balancing risk vs. reward*. Through the increasing trend toward coordinated views we are beginning to address issues of heterogeneous data, and to a lesser extent issues of isolated data by facilitating cognitive linking. Issues of balancing risk / reward and poor data quality are fundamental to the process of developing human-machine analytical systems, though we are making strong progress in expanding our cannon of design [22] and evaluation [30] tools.

The remaining 3 challenges (*increasing data volume, cadence / temporal scale*, and *progression of threat escalation*) all touch on the unique issues raised in applying visualization techniques to **streaming data**. As noted by many, blind application of established strategies for visualizing static data often falls apart when applied to streaming data, even when the systems were designed for similar tasks in similar domains [2, 9, 10]. Such challenges are not restricted to the cybersecurity domain; in the wake of ever-evolving data landscapes and human intelligence that proves difficult to scale, they are pressing issues facing the visualization community as a whole.

In 2016, Pacific Northwest National Laboratory in collaboration with the Laboratory for Analytic Sciences and various academic



(a) Bar Chart / Histogram
[16]

(b) Line Graph
[12]

(c) Scatterplot
[20]

(d) Timeline
[3]

(e) Treemap
[15]

(f) Node Link Diagram
[1]

(g) Matrix View
[13]

(h) Geographic Map
[22]

(i) Parallel Coordinates
[35]

Figure 5: 11 high-level classes of visualization types.

collaborators organized a Workshop on Streaming Visual Analytics. The goal of this workshop was to develop a guiding vision for streaming visual analytics, and to identify important research directions needed to achieve this vision. Many of the more than 50 workshop attendees have ties to the VizSec community, either as researchers or as analysts working in the trenches. We believe that by continuing to engage with the larger VA community on these issues, we will develop novel approaches that enable users to understand complex emerging events and make appropriate assessments from streaming data.

5 Conclusion

This paper provides a retrospective analysis of the past decade of VizSec, with the goal of engaging the community in developing a more cohesive understanding of emerging patterns of design. Utilizing text-mining approaches, we have identified common thematic groupings among existing work. From this base, we have manually analyzed a collection of recent systems publications in order to extract interesting trends around the use of visual encodings in various applications. We highlighted existing gaps in the adaptation of visualization techniques to cybersecurity applications, and hope this will provide avenues for the development of future systems.

978-1-5090-6356-7/17/$31.00 ©2017 IEEE

References

[1] M. Angelini, N. Prigent, and G. Santucci. Percival: proactive and reactive attack and response assessment for cyber incidents using visual analytics. In *Visualization for Cyber Security*, pages 1–8. IEEE, 2015.

[2] D. M. Best, A. Endert, and D. Kidwell. 7 key challenges for visualization in cyber network defense. In *Visualization for Cyber Security*, pages 33–40. ACM, 2014.

[3] S. Chen, C. Guo, X. Yuan, F. Merkle, H. Schaefer, and T. Ertl. Oceans: Online collaborative explorative analysis on network security. In *Visualization for Cyber Security*, pages 1–8. ACM, 2014.

[4] E. H.-h. Chi. A taxonomy of visualization techniques using the data state reference model. In *Information Visualization*, pages 69–75. IEEE, 2000.

[5] G. Conti and K. Abdullah. Passive visual fingerprinting of network attack tools. In *Visualization and Data Mining for Computer Security*, pages 45–54. ACM, 2004.

[6] G. Conti, E. Dean, M. Sinda, and B. Sangster. Visual reverse engineering of binary and data files. In *Visualization for Computer Security*, pages 1–17. Springer, 2008.

[7] G. Conti, J. Grizzard, M. Ahamad, and H. Owen. Visual exploration of malicious network objects using semantic zoom, interactive encoding and dynamic queries. In *Visualization for Computer Security*, pages 83–90. IEEE, 2005.

[8] R. J. Crouser and R. Chang. An affordance-based framework for human computation and human-computer collaboration. *Transactions on Visualization and Computer Graphics*, 18(12):2859–2868, 2012.

[9] E. Della Valle, S. Ceri, F. van Harmelen, and D. Fensel. It's a streaming world! reasoning upon rapidly changing information. *Intelligent Systems*, 24(6):83–89, 2009.

[10] R. F. Erbacher. Visualization design for immediate high-level situational assessment. In *Visualization for Cyber Security*, pages 17–24. ACM, 2012.

[11] R. Gove, J. Saxe, S. Gold, A. Long, and G. Bergamo. Seem: A scalable visualization for comparing multiple large sets of attributes for malware analysis. In *Visualization for Cyber Security*, pages 72–79. ACM, 2014.

[12] L. Hao, C. G. Healey, and S. E. Hutchinson. Flexible web visualization for alert-based network security analytics. In *Visualization for Cyber Security*, pages 33–40. ACM, 2013.

[13] L. Hao, C. G. Healey, and S. E. Hutchinson. Ensemble visualization for cyber situation awareness of network security data. In *Visualization for Cyber Security*, pages 1–8. IEEE, 2015.

[14] L. Harrison, X. Hu, X. Ying, A. Lu, W. Wang, and X. Wu. Interactive detection of network anomalies via coordinated multiple views. In *Visualization for Cyber Security*, pages 91–101. ACM, 2010.

[15] L. Harrison, R. Spahn, M. Iannacone, E. Downing, and J. R. Goodall. Nv: Nessus vulnerability visualization for the web. In *Visualization for Cyber Security*, pages 25–32. ACM, 2012.

[16] C. Humphries, N. Prigent, C. Bidan, and F. Majorczyk. Corgi: Combination, organization and reconstruction through graphical interactions. In *Visualization for Cyber Security*, pages 57–64. ACM, 2014.

[17] T. Jankun-Kelly, J. Franck, D. Wilson, J. Carver, D. Dampier, and J. E. Swan Ii. Show me how you see: Lessons from studying computer forensics experts for visualization. In *Visualization for Computer Security*, pages 80–86. Springer, 2008.

[18] K. Lakkaraju, R. Bearavolu, A. Slagell, W. Yurcik, and S. North. Closing-the-loop in nvisionip: Integrating discovery and search in security visualizations. pages 75–82. IEEE, 2005.

[19] K. Lakkaraju, W. Yurcik, and A. J. Lee. Nvisionip: netflow visualizations of system state for security situational awareness. In *Visualization and Data Mining for Computer Security*, pages 65–72. ACM, 2004.

[20] P. A. Legg. Visualizing the insider threat: Challenges and tools for identifying malicious user activity. In *Visualization for Cyber Security*, pages 1–7. IEEE, 2015.

[21] A. Long, J. Saxe, and R. Gove. Detecting malware samples with similar image sets. In *Visualization for Cyber Security*, pages 88–95. ACM, 2014.

[22] S. McKenna, D. Staheli, and M. Meyer. Unlocking user-centered design methods for building cyber security visualizations. In *Visualization for Cyber Security*, pages 1–8. IEEE, 2015.

[23] L. Nataraj, S. Karthikeyan, G. Jacob, and B. Manjunath. Malware images: visualization and automatic classification. In *Visualization for Cyber Security*, page 4. ACM, 2011.

[24] S. T. Piantadosi, H. Tily, and E. Gibson. Word lengths are optimized for efficient communication. *National Academy of Sciences*, 108(9):3526–3529, 2011.

[25] D. A. Quist and L. M. Liebrock. Visualizing compiled executables for malware analysis. In *Visualization for Cyber Security*, pages 27–32. IEEE, 2009.

[26] J. Ramos. Using tf-idf to determine word relevance in document queries. In *1st instructional conference on machine learning*, 2003.

[27] J. Saxe, D. Mentis, and C. Greamo. Visualization of shared system call sequence relationships in large malware corpora. In *Visualization for Cyber Security*, pages 33–40. ACM, 2012.

[28] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, Oct. 1948.

[29] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages*, pages 336–343. IEEE, 1996.

[30] D. Staheli, T. Yu, R. J. Crouser, S. Damodaran, K. Nam, D. O'Gwynn, S. McKenna, and L. Harrison. Visualization evaluation for cyber security: Trends and future directions. In *Visualization for Cyber Security*, pages 49–56. ACM, 2014.

[31] X. Suo, Y. Zhu, and G. S. Owen. Measuring the complexity of computer security visualization designs. In *Visualization for Computer Security*, pages 53–66. Springer, 2007.

[32] X. Suo, Y. Zhu, and S. Owen. A task centered framework for computer security data visualization. In *Visualization for Computer Security*, pages 87–94. Springer, 2008.

[33] P. Trinius, T. Holz, J. Göbel, and F. C. Freiling. Visual analysis of malware behavior using treemaps and thread graphs. In *Visualization for Cyber Security*, pages 33–38. IEEE, 2009.

[34] M. Wagner, W. Aigner, A. Rind, H. Dornhackl, K. Kadletz, R. Luh, and P. Tavolato. Problem characterization and abstraction for visual analytics in behavior-based malware pattern analysis. In *Visualization for Cyber Security*, pages 9–16. ACM, 2014.

[35] S. Walton, E. Maguire, and M. Chen. Multiple queries with conditional attributes (qcats) for anomaly detection and visualization. In *Visualization for Cyber Security*, pages 17–24. ACM, 2014.

[36] L. Williams, R. Lippmann, and K. Ingols. An interactive attack graph cascade and reachability display. In *Visualization for Computer Security*, pages 221–236. Springer, 2007.

[37] L. Williams, R. Lippmann, and K. Ingols. Garnet: A graphical attack graph and reachability network evaluation tool. In *Visualization for Computer Security*, pages 44–59. Springer, 2008.

[38] X. Yin, W. Yurcik, M. Treaster, Y. Li, and K. Lakkaraju. Visflowconnect: netflow visualizations of link relationships for security situational awareness. In *Visualization and Data Mining for Computer Security*, pages 26–34. ACM, 2004.

[39] I. Yoo. Visualizing windows executable viruses using self-organizing maps. In *Visualization and Data Mining for Computer Security*, pages 82–89. ACM, 2004.

[40] W. Yurcik. Tool update: Nvisionip improvements (difference view, sparklines, and shapes). In *Visualization for Computer Security*, pages 65–66. ACM, 2006.

[41] W. Yurcik. Tool update: visflowconnect-ip with advanced filtering from usability testing. In *Visualization for Computer Security*, pages 63–64. ACM, 2006.

[42] W. Yurcik, X. Meng, and N. Kiyanclar. Nvisioncc: a visualization framework for high performance cluster security. In *Visualization and Data Mining for Computer Security*, pages 133–137. ACM, 2004.

[43] W. Zhuo and Y. Nadjin. Malwarevis: entity-based visualization of malware network traces. In *Visualization for Cyber Security*, pages 41–47. ACM, 2012.

[44] A. Zoss. Introduction to data visualization: Visualization types, Dec 2015.

978-1-5090-6356-7/17/$31.00 ©2017 IEEE