A Survey of Visual Analytics Tools for Effective Decision-Making

R. Jordan Crouser, Erina Fukuda ‘18, and Subashini Sridhar ‘18J
Department of Statistical & Data Sciences, 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:

- We identify common **thematic groupings** among existing work, as well as interesting patterns of design around the utilization of various visual encodings.
- We also discuss **existing gaps** in the adaptation of visualization techniques for cybersecurity applications, and recommend avenues for future exploration.

Automated Analysis via Text Mining

Goal: obtain a high-level overview of the state of the practice through automated text mining on a large corpus of published work on visualization for cybersecurity:

- 161 papers published in IEEE Visualization for Cyber Security from 2004-2015
- Preprocessed to extract full text and metadata (authors, date of publication, etc.)

Approach:

- Simple bag-of-terms model [1] using single words, bigrams, and trigrams
- Eliminate all terms contained in the nltk [2] English stopwords library
- Normalize relative weight of terms using term frequency – inverse document frequency (TF-IDF) [3] Discard terms that appeared in more than 80% or fewer than 10% of the corpus, leaving 2,369 unique terms
- Compute the **cosine similarity** of each pair of documents (characterized by their respective TF-IDF vectors), and use this information to construct a complete pairwise distance matrix over all 161 publications

References

[1] C. E. Shannon. A mathematical theory of communication. The Bell System Technical Journal, 27:379–423, Oct. 1948.
[2] S. Bird, E. Loper and E. Klein. Natural Language Processing with Python. O'Reilly Media Inc. 2009.
[3] J. Ramos. Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning, 2003.
[4] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In Proceedings of the IEEE Symposium on Visual Languages, pages 336–343. IEEE, 1996.
[5] E. H.-h. Chi. A taxonomy of visualization techniques using the data state reference model. In Proceedings of the IEEE Symposium on Information Visualization, pages 69–75. IEEE, 2000.
[6] W. Zhuo and Y. Nadjin. Malwarevis: entity-based visualization of malware network traces. In Proceedings of the Ninth International Symposium on Visualization for Cybersecurity, pages 41–47. ACM, 2012.
[7] R. J. Crouser and R. Chang. An affordance-based framework for human computation and human-computer collaboration. In IEEE Transactions on Visualization and Computer Graphics, 18(12):2859–2868, 2012.

Identifying Meaningful Topics through Clustering

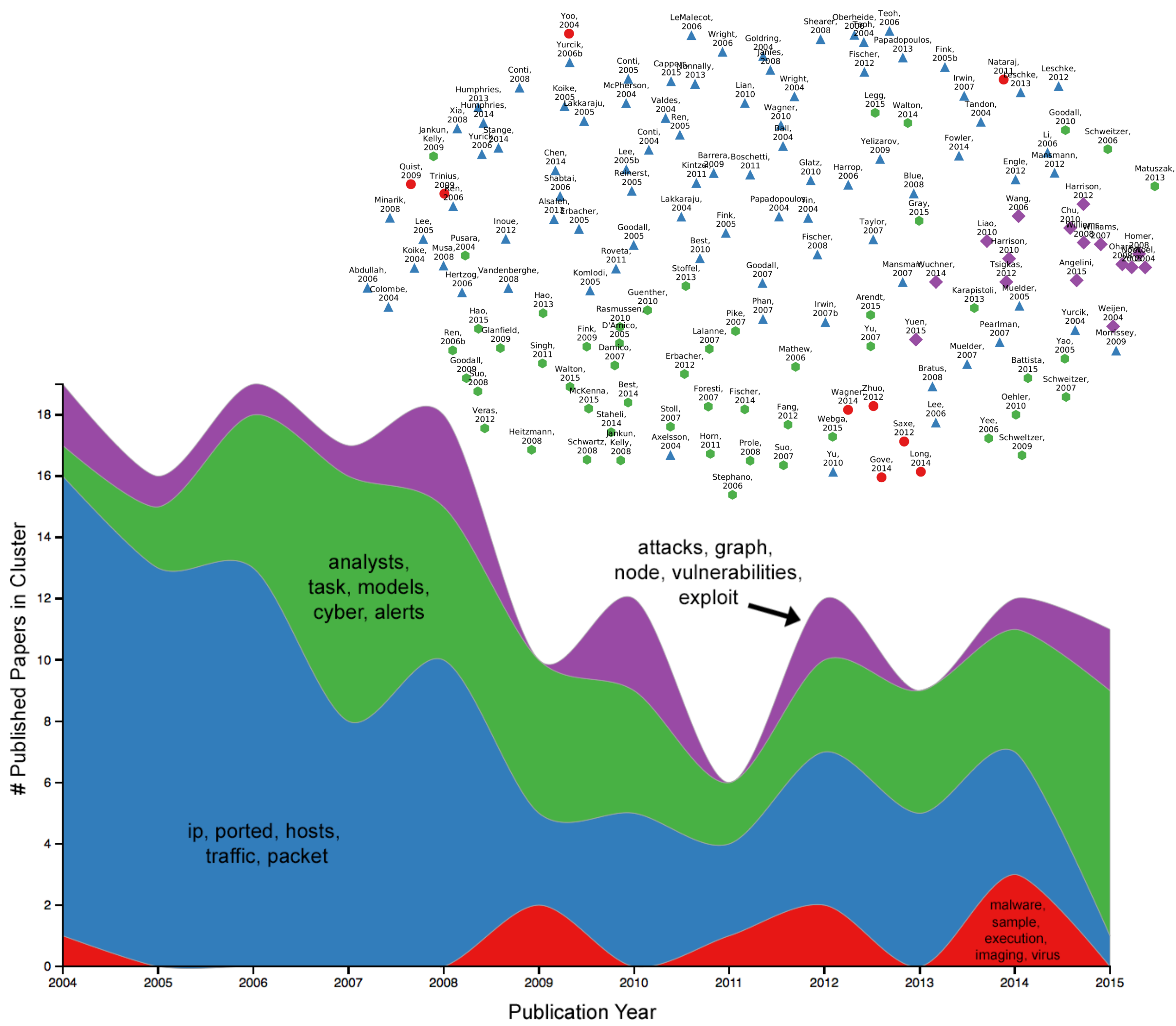


*Fig. 1: (top) MDS projection of **k**-means clustering (**k** = 4) of 161 VizSec papers spanning the years 2004-2015. Distance between publications is calculated using TF-IDF vectors constructed from single words, bigrams, and trigrams. (bottom) Distribution of the 4 automatically-generated clusters of VizSec publications. In order to illustrate the thematic groupings, we have included the 5 most frequently used words in each cluster.*

Performing k-means clustering on the extracted text, we observe 4 clear thematic groups (Fig. 1):

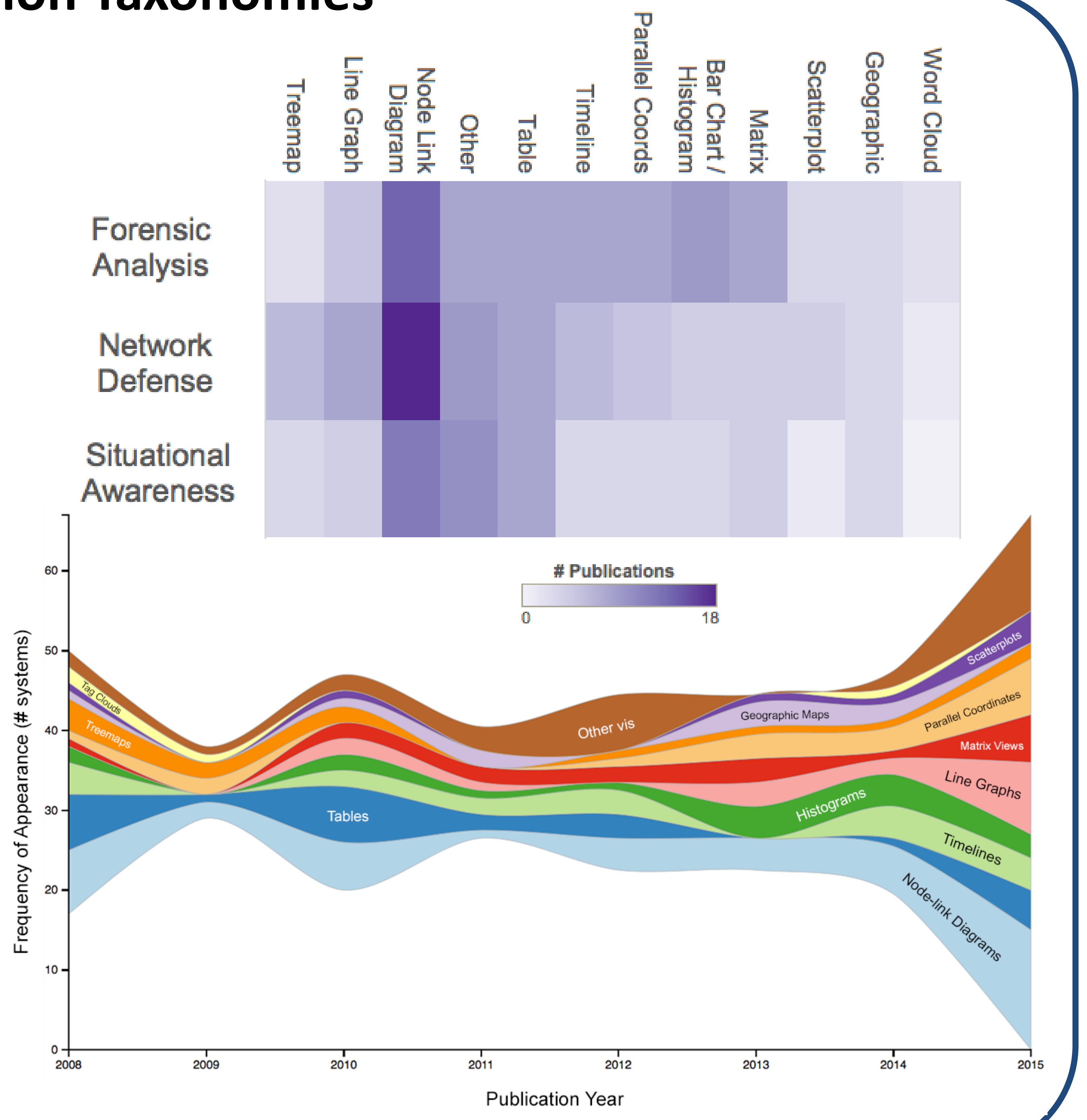
- In the blue cluster (85 papers), we find tools for cyber situational awareness such as VisFlowConnect, NVisionIP, and NVisionCC alongside work by Conti et al. in using visualization low-level features to identify evidence of malicious activity
- In the green cluster (51 papers), we find many higher-level frameworks which organize the spaceand process of designing visualization systems for cyber security applications, including work by Jankun-Kelly et al., Staheli et. al, and Suo et al.
- In the purple cluster (16 papers) we find many systems and frameworks which exploit hierarchical or graph-theoretic structure in order to identify network vulnerabilities, such as work by Harrison et al. and Williams et al.
- The red cluster (9 papers) consists of work in the area of malware analysis.

Applying Existing Visualization Taxonomies

Drawing on taxonomies by Shneiderman [4], Chi [5], and Duke University [6], we identified 11 high-level visual mapping 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).

Examining the distribution of these various visual metaphors across the fore`nsic analysis, situational awareness or network defense classes, some interesting patterns of design begin to emerge (see Fig. 2 top). For example, 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 [7], providing opportunities for further exploration.

In the bottom pane of Fig. 2, we highlight temporal trends in the utilization of the 11 high-level visualization types in VizSec publications from 2008 to 2015. Note the change in utilization of visual metaphors such as parallel coordinates (increasing beginning in 2012) and treemaps (slowly decreasing after 2010).


Treemap
Line Graph
Node Link Diagram
Other
Table
Timeline
Parallel Coords
Bar Chart / Histogram
Matrix
Scatterplot
Geographic
Word Cloud
Forensic Analysis
Network Defense
Situational Awareness
# Publications
Frequency of Appearance (# systems)
Tag Clouds
Treemaps
Tables
Other vis
Geographic Maps
Scatterplots
Parallel Coordinates
Matrix Views
Line Graphs
Histograms
Timelines
Node-link Diagrams
Publication Year