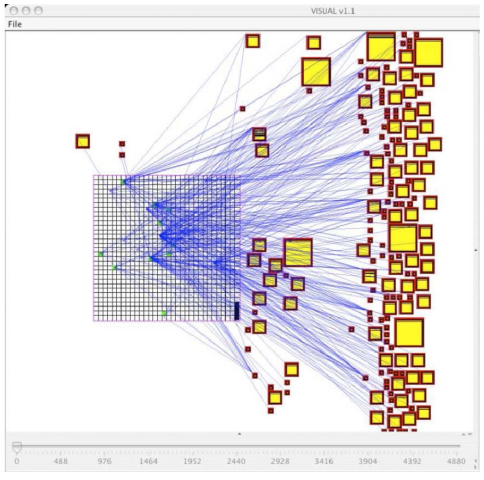
- URL: <http://ieeexplore.ieee.org/document/6007132/>   
- Title: A Survey of Visualisation Systems for Network Security  
- Authors: H. Shiravi, A. Shiravi, A.Ghorbani  
- Notes: Also includes me fiddling with the new in-line citation style. Has a super-useful table of where information can come from, could be good to link to later when looking at security events from IDS and how to format data.

##Review

Security Visualization is a very young term, and many common visualisation techniques are not designed for security related data (Shiravi, Shiravi, & Ghorbani, 2012). Manually traversing textual logs is not only frustrating and time consuming, but may result in important details being overlooked. This paper explores methods of showing administrators quantitative data in meaningful ways so as to better look for anomalies or patterns from sources such as intrusion detection systems, port scanning tools and firewalls.

The paper does a fantastic job showing various interesting ways of converting vast quantities of event types into graphs and other visual representations. This included splitting the events into different types: Network traces, security events, network activity context, user/asset context, network events and application logs. The image to the right shows 80 hours of network data on a network of 1020 hosts. The internal network is represented by the grid on the left, and external servers by squares on the right, wit square size denoting the level of activity. This is not extremely obvious upon first look.

One thing the paper does not do is really discuss the requirements of visualisation compared to merely parsing and reformatting text data. While the image above does a good job of showing that large amounts of data may be transferred to external servers from internal sources, we have no easy way of knowing if any of this data is malicious or coming from potentially unwelcome sources. If this data were formatted as an easily formatted excel table for example, we could apply filters very easily to look for patterns in data. This is a common theme in the paper, where visualisations show what at first seems to be useful information, but allows little to no exploration of potential anomalies once identified short of digging through raw data.

##Citation

Shiravi, H., Shiravi, A., & Ghorbani, A. a. (2012). 006 A survey of visualization systems for network security. *IEEE Transactions on Visualization and Computer Graphics*, *18*(8), 1313–1329. https://doi.org/10.1109/TVCG.2011.144

- URL: <http://jamia.oxfordjournals.org/content/jaminfo/18/5/544.full.pdf>   
- Title: Natural Language Processing: An Introduction  
- Authors: Prakash M Nadkarni, Lucila Ohno-Machado, Wendy W Chapman  
- Notes: Aimed at a medical audience but could still be relevant as an NLP intro

##Review

This introduction to Natural Language Processing acts as an overview of common machine-learning approaches currently being used and possible future directions of NLP, as well as some of the associations with IR – Information Retrieval. One of the first things defined is that of statistical NLP –NLP based on machine learning methods, learning via large annotated bodies of text which provided the standard they were looking to achieve. The paper quickly becomes complex, looking into data driven approaches to NLP and their drawbacks such as Hidden Markov Models (HMMs), which is a system where variables can switch between several states and generate possible outputs. The issue with HMM’s is that we can only see the output, not the process at it takes to comes to that output.

While this paper is primarily looking into NLP, it more looks into the methods of machine learning that could be applied to NLP. However, it still lists and explains a good number of NLP sub-problems which, while primarily aimed toward the medical field, are still relevant in computer science, such as sentence boundary detection and morphological decomposition – The act of separating words into smaller words.

The paper also has an interesting section focussing on the future of artificial intelligence and NLP, quoting heavily IBM’s Watson supercomputer and its attempt at beating humans in the game Jeopardy. With 16TB ram, Watson is designed to hold all of reference content in memory, as opposed to being disk-I/O-bound, which makes its seek time exceptionally slow. However, Watson can be easily misled with certain questions – Asking it “Which US city has two airports, one named after a World War II Battle, the other after a World War II Hero?” would be a multi-step process which Watson could not answer, as the reference content used for machine learning was structured as one sentence question and answers (“What/who is/are X?”).

##Citation

Nadkarni, P. M., Ohno-Machado, L., Chapman, W. W., Manning, C., Raghavan, P., Schuetze, H., … Miller, R. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association : JAMIA*, *18*(5), 544–51. https://doi.org/10.1136/amiajnl-2011-000464