In Book of Abstracts of the International Symposium on Digital Humanities, pages 39-43, Linnaeus University, Växjö, Sweden, 2016.

Network Visualization for Digital Humanities: Two Case Studies of Visual Analyses for Text Analytics

Ilir Jusufi Department of Media Technology Linnaeus University, Sweden
ilir.jusufi@lnu.se

Andreas Kerren Department of Computer Science Linnaeus University, Sweden
andreas.kerren@lnu.se

INTRODUCTION

Much of the data created nowadays in fields such as Digital Humanities (DH) is of relational nature, such as social or semantic networks. Researchers often decide to depict networks as node-link diagrams to make a better sense of the complex nature of data (cf. Figure 1). Understanding the topology of such a network can be very important. For instance, if we show our friends as network nodes and their friendship as edges between the nodes, it becomes easy to identify groups of friends from different social settings (work friends, high school friends, etc.).

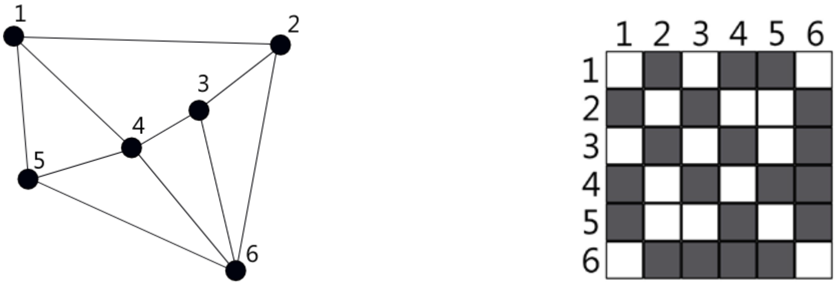

Figure 1. The same (undirected) network depicted by using two different visual representations: as node-link metaphor on the left hand side, and as a matrix metaphor on the right.

Networks usually have additional attributes attached to their elements (Kerren et al., 2014). For instance, we can model a number of documents in a repository as nodes and use edges to describe co-authorship. Additionally, we might want to explore other aspects of such a corpus, like the keywords for each document, its genre, and various other data associated. Here, it is often desirable to get an overview about the network structure and how different data values relate to this structure. For the more general case of visual text analytics, we refer to the surveys of (Kucher and Kerren, 2015; Kerren et al., 2014). In this paper, we present two case studies for visualizations in DH with a focus on publication networks. But first, we will introduce our data sets used in these studies.

Data

Both approaches presented here use the so-called *Jigsaw data set* (D1) containing metadata for every IEEE InfoVis and VAST conference paper (Stasko et al., 2014). Here, nodes represent the papers. The edges between nodes represent co-authorship, i.e., if two papers share an author, then their node representations are connected with an edge. This creates a co-authorship network. Each paper has additional metadata attached to it, such as concept terms for describing the content of the paper or publication and conference data.

An additional data set (D2) has been used for our first case study. This data set is very similar to the first one, but instead of concept terms, we have computed and attached the most used keywords for each document published in conferences within HCI and Information Visualization venues.

CASE STUDY 1

There are several ways to visually encode additional attributes to the nodes in a network (Kerren et al., 2014). For instance, we can position specific nodes close to appropriate regions in the display that hold some specific semantics. One example is shown in Figure 2. In this screenshot we see the *JauntyNets tool* visualizing the D2 data set (Jusufi et al., 2013). The most used keywords, i.e., the attributes, are shown as rectangles in the circular layout. Each paper with a certain number of occurrences of the given keyword has an invisible link to it. This invisible link acts as a spring and pulls the papers towards itself. The higher the number of occurrences of the keyword, the stronger the pull. In consequence, nodes move to those attributes with the highest pull and may build clusters. Users can additionally group certain attributes according to a desired semantics. In the figure, for instance, we have created three groups of attributes (called *interaction*, *mobile*, *graph*).

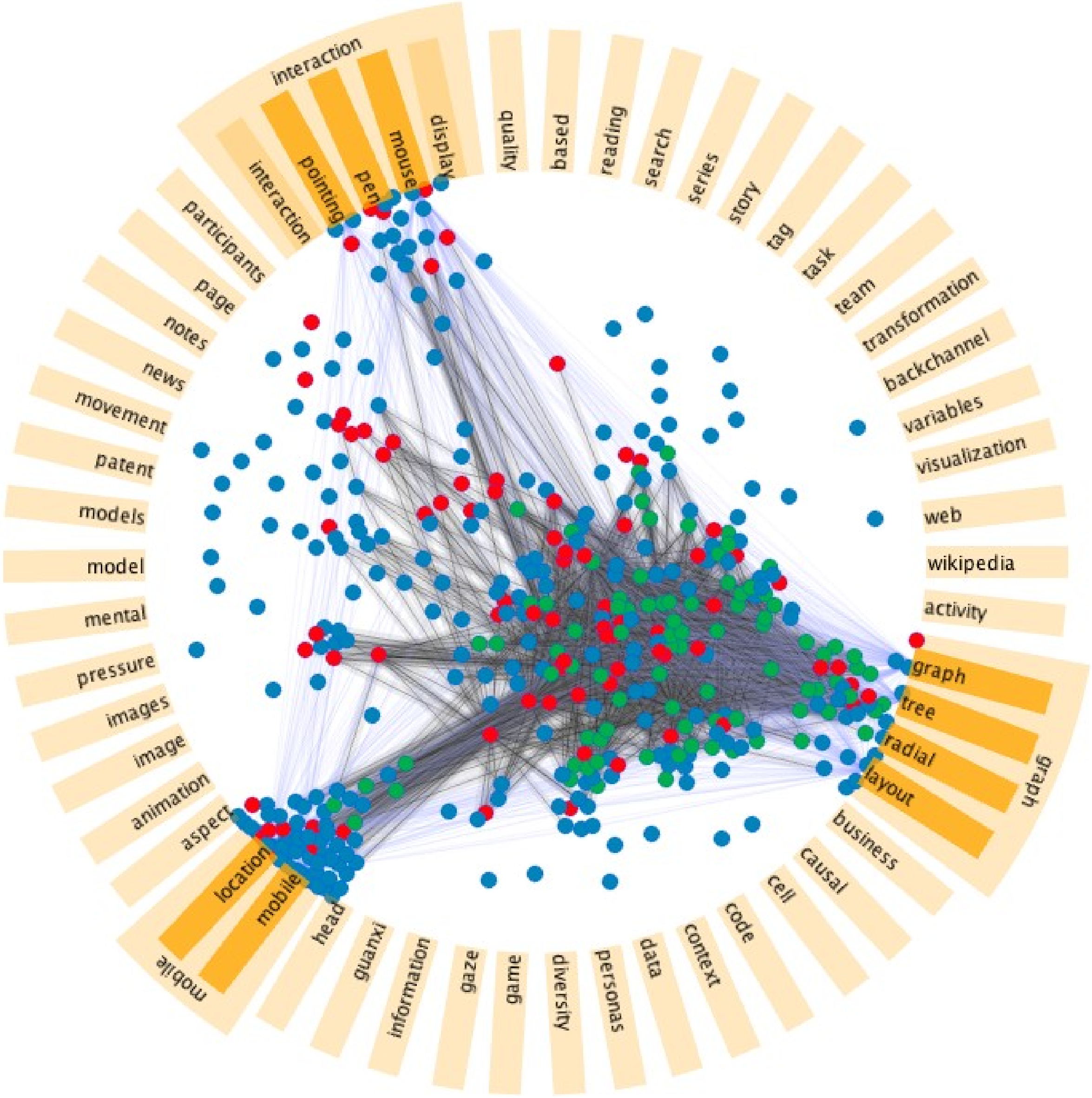

Figure 2. The screenshot displays a network of 421 nodes with 55 attributes (data set D2). The nodes are colored differently, because they have been clustered based on their attribute values. Taken from (Jusufi et al., 2013).

Figure 3 shows the D1 data set. Two attributes *InfoVis* and *VAST* specify the conference where the papers were published. InfoVis papers are shown by blue nodes, while VAST papers are represented by red nodes. The attribute group *geovisualization* has been created manually. Interesting structural features can be noticed immediately, such as a lot of unconnected nodes or subgraphs. Other noticeable structures are a number of cliques. Upon the closer examination of the two cliques marked as A and B, we are able to understand the following short story: there was only one shared author in group A, and he published articles related to geovisualization in both venues; in contrast to group B where three distinct authors collaborated and published exclusively at the InfoVis conference.

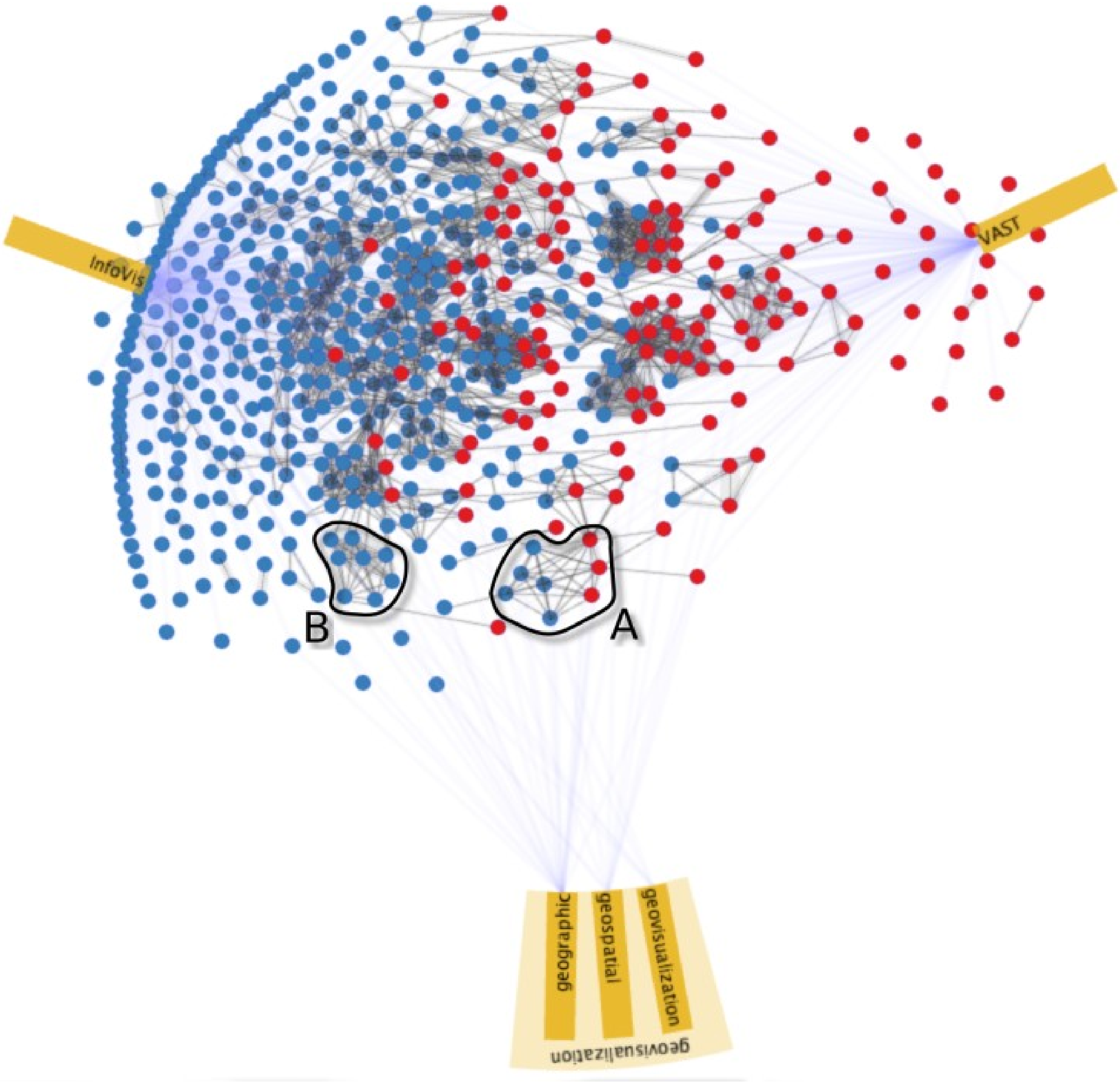

Figure 3. The screenshot displays papers from the D1 data set. Taken from (Jusufi et al., 2013).

CASE STUDY 2

Now, we present a different approach for exploring papers and co-authorship (Jusufi et al., 2014). A k-means clustering algorithm has been used to cluster the papers based on the concept terms from data set D1 (cf. Figure 4). Nodes are placed in different clusters in a circular layout. The edges are shown in two ways: internal co-authorship within a single cluster and external co-authorship going to other clusters. Additionally, co-authorship patterns regarding publication years within clusters can be observed as well by a green color gradient. The most often-occurring concept terms are drawn as tag cloud into the center of each cluster. In Figure 4, we show one of the graphical layouts we implemented. Here, the edges have been bundled together to avoid clutter.

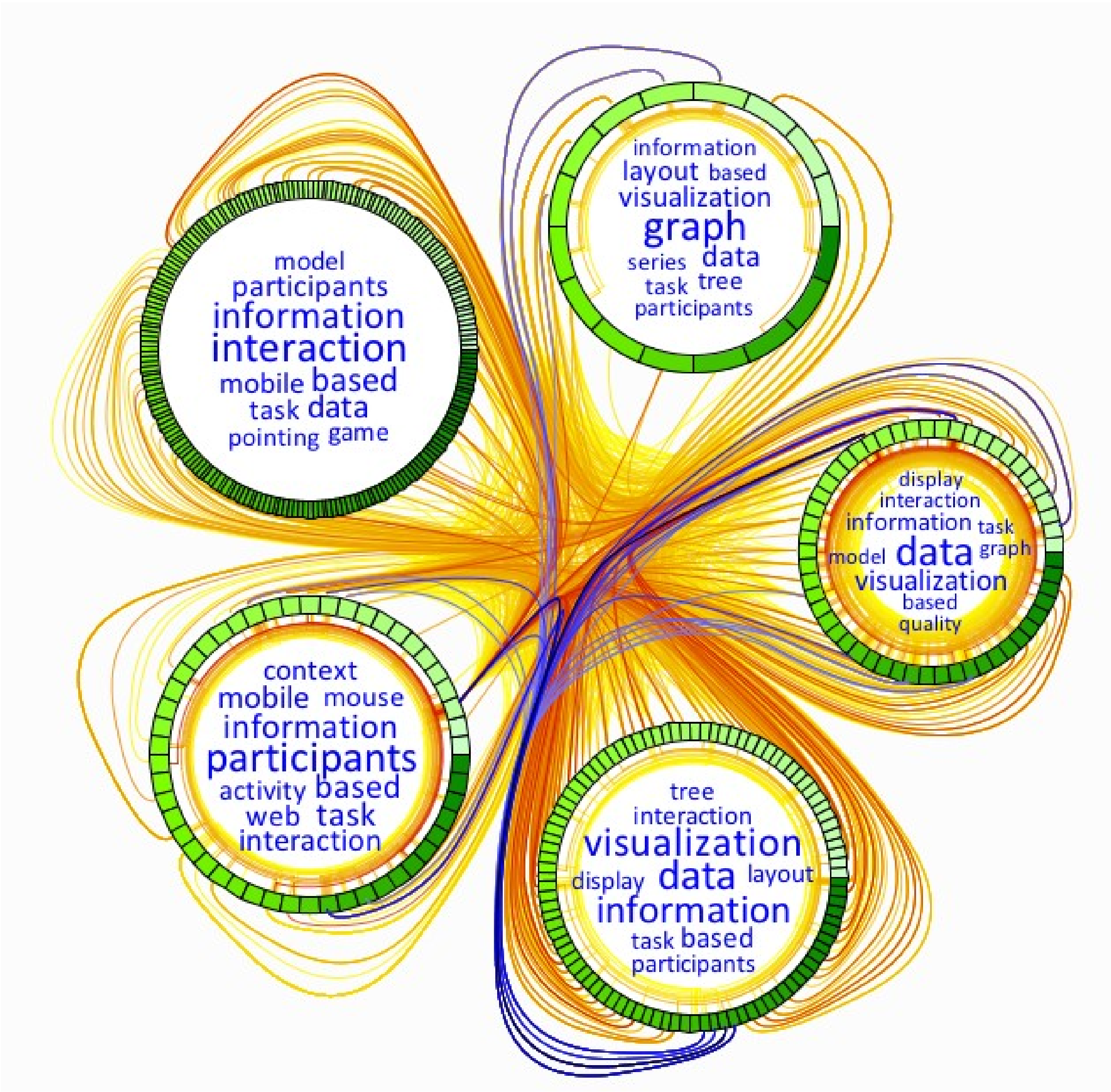

Figure 4. A screenshot of our approach representing five distinct clusters. Nodes (i.e., papers) are represented around each cluster with a green color saturation specifying the publication date. The edge saturation represents the number of shared co-authors. Selected edges are highlighted in blue. Taken from (Jusufi et al., 2014).

As an example how interaction supports the analysis, we might ask if people publish within the same topic, or if the papers are topically distributed within different clusters, see Figure 5. After selecting the only paper in cluster A (highlighted in orange) that has edges to other clusters with more than three co-authors, we can immediately identify related terms, because the concept terms unrelated to the paper are grayed-out. Upon selecting the other two papers in the clusters B+C, we can learn that these papers have almost no similarities.

CONCLUSION

The challenge of analyzing many hundreds or thousands of related text documents increases with the daily growing amount of new data. We have presented two different visualization approaches that could help researchers in DH to investigate the content and metadata of publication corpora. Both approaches are highly interactive. Some insights that were gained by using the presented approaches are not possible with traditional computational methods, especially when users are not aware of what they want to find, i.e., for exploratory analyses.

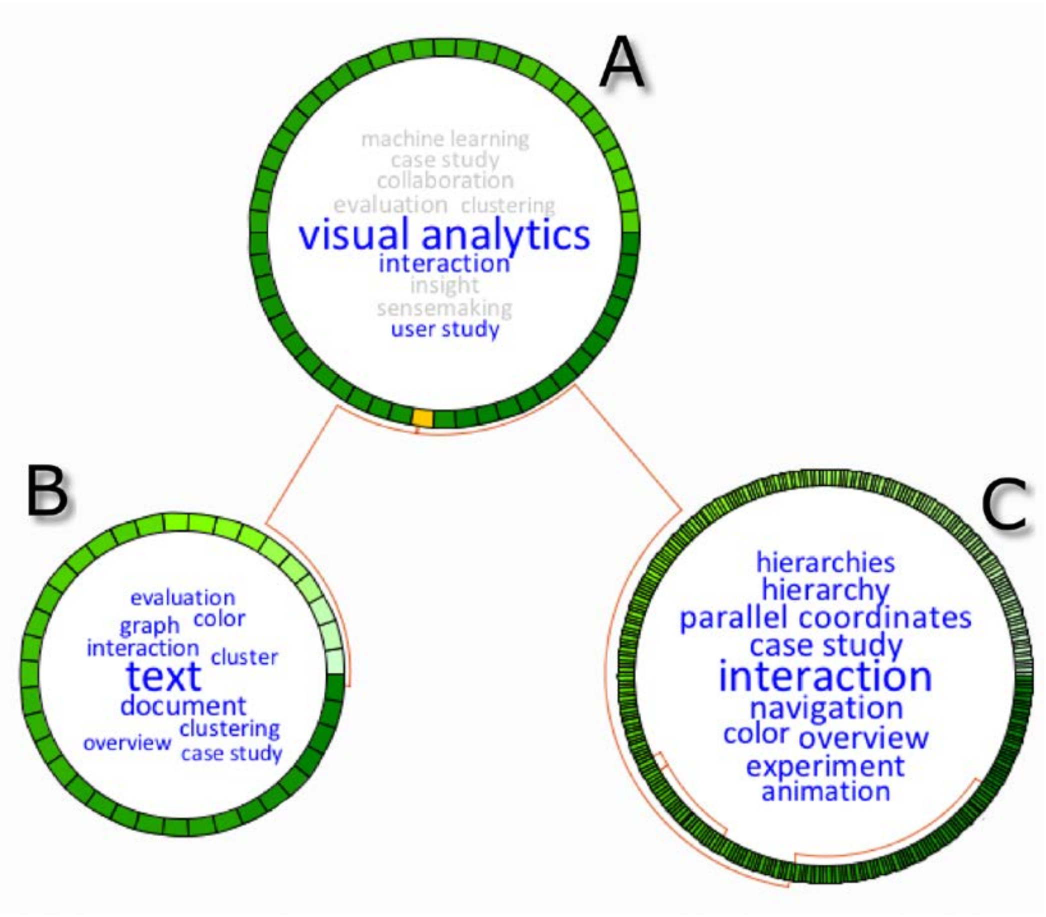

Figure 5. A group of authors who published the selected paper (highlighted in orange) wrote papers that again were grouped into separate clusters. Taken from (Jusufi et al., 2014).

REFERENCES

Jusufi, I., Kerren, A., and Zimmer, B. (2013) Multivariate Network Exploration with JauntyNets. In Proceedings of the 17th International Conference on Information Visualisation (IV '13). IEEE Computer Society, 19–27.

Jusufi, I., Kerren, A., Liu, J., and Zimmer, B. (2014) Visual Exploration of Relationships Between Document Clusters. In Proceedings of the 2014 International Conference on Information Visualization Theory and Applications (IVAPP '14). SciTePress, 195–203.

Kerren, A., Purchase, H. C., and Ward, M. O. (Eds.) (2014) *Multivariate Network Visualization*. Volume 8380 of Lecture Notes in Computer Science (LNCS) State-of-the-Art Survey, Springer.

Kucher, K., and Kerren, A. (2015) Text visualization techniques: Taxonomy, visual survey, and community insights. In Proceedings of the 2015 IEEE Pacific Visualization Symposium (PacificVis '15). IEEE Computer Society, 117–121.

Stasko, J., Görg, C., Liu, Z., Pratap, S., and Sainath, A. (2014, Jan 1) Jigsaw: Visual Analytics for Exploring and Understanding Document Collections. Retrieved from http://www.cc.gatech.edu/gvu/ii/jigsaw/.