Case Study: Visualizing Visualization

Frank van Ham*

Department of Mathematics and Computer Science
Technische Universiteit Eindhoven
P.O. Box 513, 5600 MB Eindhoven
The Netherlands

ABSTRACT

In this case study we attempt to visualize a real-world dataset consisting of 600 recently published information visualization papers and their references. This is done by first creating a global layout of the entire graph that preserves any cluster structure present. We then use this layout as a basis to define a hierarchical clustering. The clusters in this hierarchy are labelled using keywords supplied with the dataset, allowing insight into the clusters semantics.

CR Categories: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Graphical User Interfaces; H.2.8 [Database Management]: Database Applications—Data Mining I.5.5 [Pattern Recognition]: Clustering—Algorithms

Keywords: Graph Visualization, Graph Drawing, Clustering, Small World Graphs, Information Visualization

1 INTRODUCTION

Cross referenced document collections such as the one presented in the 2004 Information Visualization contest are very commonplace in today's world. Other examples include large encyclopedias or dictionaries and even the Internet itself can be thought of as a document collection. Insight in the structure of these networks can tell us a lot about the prevailing global topics and their interrelations. However, visualizing these networks can be quite a challenge because of their often highly interconnected nature.

A useful concept in this respect are small world graphs, first discovered by Milgram [1] in 1967. Small world graphs are graphs that have a small average shortest path length (the average length of the shortest path between all pairs of nodes). The only property that distinguishes them from random graphs is the fact that small world graphs are clustered to some extent. That is, we can identify groups of nodes that have a lot of interconnections (connections to other nodes in the same group), while having a comparably small number of intraconnections (connections to other nodes in different groups). The amount of clustering in a graph can be expressed by the clustering coefficient [3]. The clustering coefficient is 1 for a complete graph and 0 for a graph with no edges. We computed clustering coefficient of 0.064 for this graph, an equally sized graph with a random edge distribution has a coefficient of 0.005, indicating that some clustering is present in the contest graph. Since this clustering is the only thing that separates this graph from a random graph our visualization will focus on extracting these structural clusters. For this we employ a modified force directed layout technique to create a global layout, and then cluster the graph based on this layout. We discuss this layout technique and the clustering in section 2. An overview of the analysis process is given in section 3. Finally we discuss results in section 4.

*e-mail: fvham@win.tue.nl

2 LAYOUT AND CLUSTERING

One way to extract cluster structure from the given graph, is by explicitly clustering the graph based on its structure. Clustering in itself is a fuzzy problem however, there is no single clear-cut definition of a cluster and based on the structure of a graph, it is very well possible that nodes can belong to multiple clusters. Even when given a clustering, creating a satisfactory layout of a small world graph is no trivial task, because of the large number of edges. Instead we use a different approach and base the clustering of a graph on its layout. This is only possible if the layout accurately reflects the clustering. We adapted a force directed layout technique with a modified force model from [2]. The result is a layout of the entire graph in which the geometric distance between two clusters of nodes is inversely proportional to the coupling between the two clusters. The coupling between two clusters A and B is defined as E(A,B) / (Size(A) * Size(B)) where E(A,B) represents the number of edges connecting A and B. In these layouts strongly connected groups of nodes show up as visually dense clusters of nodes, and the closest (groups of) nodes are those which are coupled the strongest.

Based on the layout generated in the previous step we can generate a clustering. First, we assign all nodes to singleton clusters. We then iteratively select and merge the geometrically closest clusters in the layout, assigning this newly created cluster to the weighted average positions of the previous clusters. This process is repeated until we are left with one single cluster, resulting in a binary hierarchy of clusters. Each cluster is assigned a *level* l in the hierarchy that corresponds to the distance between its two subclusters.

We can then abstract from the layout by defining a minimum abstraction level A, and only displaying the part of the cluster hierarchy with $l > A$. We can even define an increasing function $A(d)$ in which the abstraction level depends on the distance to a focal point f, to obtain a view with local detail inside a global context.

3 ANALYSIS

As a basis for the analysis of the data we first created a global layout using the force directed algorithm described in the previous section. The resulting layout showed a large central cluster with a number of satellite clusters. Most satellite clusters could be easily identified as side interests of Information Visualization by examining the paper titles, classifying them as algorithm animation, volume visualization and image similarity. Inside the main cluster there was some smaller scale clustering present, but manual inspection did not reveal any immediate clues for their shared content (if any). We created a hierarchical clustering based on this layout and coarsened the graph to about a quarter of its original size (150 nodes). Since each paper has a number of keywords attached we can use these to scan for common keywords in a clustered set of papers to get

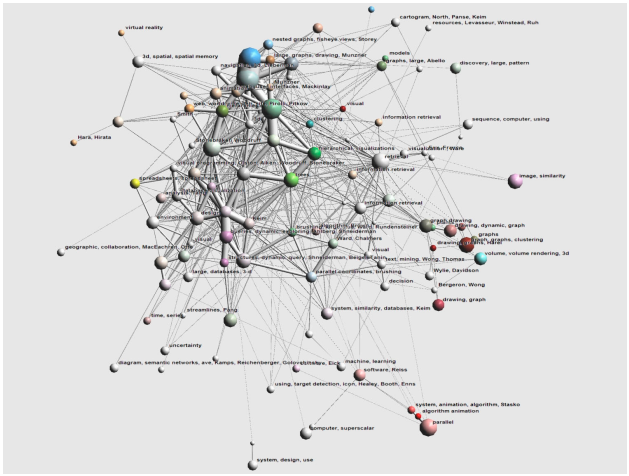

Figure 1: Overview image of the InfoVis research field.

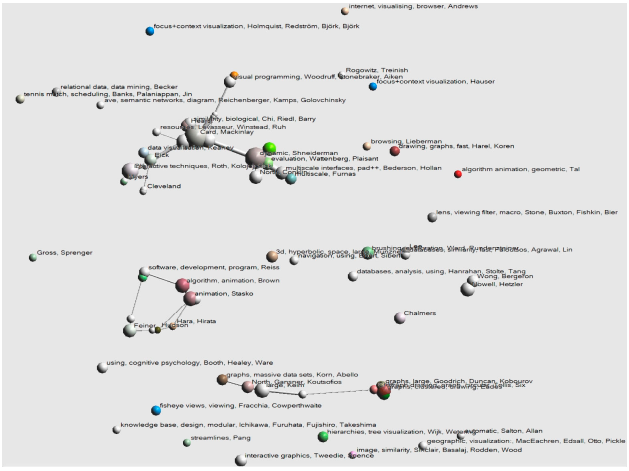

Figure 2: Cooperations between researchers in the InfoVis field.

an impression of the clusters content. There were three problems however.

First, not all papers had keywords, we solved this by adding the words in the papers title to the keyword set of a paper. Secondly,not all keywords are meaningful. For example, some keywords are too general ('information visualization'), and some words in the paper title are not relevant ('a','an','with'). To solve this we manually created a list of keywords that are to be ignored.

A final (bigger) problem is that there is no uniformity in keywords: fisheye techniques are keyworded with (amongst others) '*fisheye*', '*fish-eye*', '*fish-eye technique*' and '*fisheye views*'. The same goes for treemaps ('*treemap*','*tree-maps*','*tree-map*' and '*tree-map visualization technique*'), multidimensional ('*multidimensional*','*multi-dimensional*','*multi-dimensional information space*','*multidimensional information space*') and many others. These differences in spelling reduce the accuracy in mapping keywords to clusters. Since we expected to find clusters of common authors, we also included author names in the keyword set. Unfortunately, author names suffered from the same problem, although this was compensated to some extent by an (incomplete) list of equivalent authors.

We then mapped common keywords to clusters, only displaying them when the relative amount of keywords or authors in a cluster was over a preset threshold. For keywords we used a threshold of 0.3, for authors 0.5. The resulting image (Figure 3) gives a rough overview of the information visualization research field. To emphasize major research areas we manually applied a color to each node in the graph based on their keywords.

Other coloring options we used were the possibility to mark papers based on author (to see what research field an author is in) or by year of publication (to see what research fields were particularly active in a given time interval).

To visualize relations between authors we used a different edge set than the references given in the dataset. Instead, we defined an edge between two papers if there is an overlap between their respective author sets. We used a similar procedure to the one outlined above to compute a layout and attach keywords (Figure 3). Here we see a much stronger clustering structure present because most authors work with a fixed set of coauthors within a fixed research area. The combination of author and topic keywords allow us to quickly asses the research field an author is in.

4 Discussion

The tool we used was specifically designed to be able to deal with small world graphs such as this one. Most other tools do not perform very well on graphs with a fairly small diameter. The fact that we used a less well known force model that tries to map structural properties to distances, allows us to infer structural properties from the resulting layout. Using explicit clustering to identify research areas is a difficult route since one does not know the number of clusters beforehand, and it does not remove the obligation to create a meaningful global layout of the entire graph afterward. Our approach has the following advantages

- The ability to interactively select any desired granularity in the layout, makes it easy to select a semantically meaningful clustering by looking at the keywords that appear.
- The clustering is based on a global layout of the entire graph, meaning we can preserve the users mental map when changing the granularity.

and disadvantages

- Since the clustering is directly based on the layout, the quality of the clustering is dependent on the quality of the layout. The layout results from an iterative optimization procedure and is not necessarily the global optimum.
- Our layout algorithm currently still uses a naive $O(N^3)$ implementation of a spring embedder and does not scale well.

The resulting images give an impression of the Information Visualization field. The fact that most of the papers reside in one strongly connected cluster, is probably a consequence of the lack of specialization in InfoVis. Because Information Visualization is an applied science, good solutions will often apply different combinations of techniques, meaning that researchers in the field have to be knowledgeable in different areas.

References

[1] S. Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

[2] A. Noack. An energy model for visual graph clustering. In *Proc. 11th Int. Symp. on Graph Drawing*, pages 425–436. Springer-Verlag, 2003.

[3] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature 393*, pages 440–442, 1998.