Exploring InfoVis Publication History with Tulip

Maylis Delest* Tamara Munzner† David Auber‡ Jean-Philippe Domenger‡

ABSTRACT

We show the structure of the InfoVis publications dataset using Tulip, a scalable open-source visualization system for graphs and trees. Tulip supports interactive navigation and many options for layout. Subgraphs of the full dataset can be created interactively or using a wide set of algorithms based on graph theory and combinatorics, including several kinds of clustering. We found that convolution clustering and small world clustering were particularly effective at showing the structure of the InfoVis publications dataset, as was coloring by the Strahler metric.

1 INTRODUCTION

We use the Tulip [2] system to investigate the InfoVis 2004 contest dataset of relationships between published papers. Tulip is a highly scalable open-source system for exploratory visualization of graph and tree that offers an extensive set of algorithms based in combinatorics. We use the following Tulip features extensively:

- interactive navigation, layout and selection
- subgraph hierarchy navigation
- extracting the induced subgraph for a selection set
- controlling color, rendering order, and label drawing priority with the Strahler metric [1]
- convolution clustering [4]
- small world clustering [3]
- using lightweight *plugins* for additional computation

2 TULIP CAPABILITIES

We created a large directed graph from the XML file provided by the contest organizers. The graph edges are oriented: there is an edge (Y,X) if paper Y is cited by paper X. All nodes and edges have properties that can be used for computation. Our original graph contained all the properties such as year of publication specified in the XML file, and Tulip can compute many other graph-theoretic properties, for example the arity of nodes (total number of incoming and outgoing edges) or their ranking by several metrics. Tulip is extensible through its *plug-in* architecture, and we wrote plugins to do application-specific computations. For instance, to create a graph containing only authors and conferences, we created new links between authors based on paper citations and assigned a citation count property for author nodes based on all their papers. We also added coauthorship links, then deleted the paper nodes. Although we experimented with creating separate graphs for paper coauthoring and paper cocitation, we found the combined graph of both coauthorship and cocitation to be much more informative.

We created many subgraphs of the entire dataset using the built-in capabilities of Tulip. The general interaction paradigm in Tulip is that the user loads a graph and then creates many subgraphs while interacting with the graph. These subgraphs form a hierarchy shown in an auxilliary window, and internal storage of data is handled efficiently using inheritance so that only data unique to a subgraph uses memory. In contrast to simply filtering away edges and nodes so that they are not drawn, subgraph manipulation changes the complexity of the underlying dataset, so that algorithms intractable for the full dataset can be run on its subgraphs.

We can interactively select some set of nodes and edges using regular expression search on a string property like author names, or an arithmetic operation on numeric data such as year of publication. We can simply delete a selected set, or carry out more sophisticated graph-theoretic operations such as finding the induced subgraph or reachable subgraph for that set. For instance, we found induced subgraphs to show the evolution of the graph over time, because simply selecting all the papers for a particular year would lose the links to their authors.

We use the Strahler metric extensively. The Strahler metric takes into account the global branching structure of the dataset as follows. For each node in the graph, we compute the branching of the DFS spanning directed acyclic graph (DAG) and the number of nested cycles induced by edges that are not in the spanning DAG. In contrast, a simpler measure such as the citation count is simply a local computation at each node. We can use this metric to control color, label drawing priority, and progressive rendering order [1]. The effects of the latter is not directly visible in the still images, but can be seen in the accompanying video. We use a greedy algorithm to decide which labels to draw using the ordering given by the Strahler metric of the nodes. The visual density of the labels was manually optimized for each image using an interactive slider; we typically used maximum density for small subgraphs but turned down the density when showing overviews of large components. Finally, the titles of all conferences and some papers were manually shortened to make browsing easier.

In some of our examples, we compute a few simpler metrics for author nodes. One is "longevity": the number of years between their first publication in this dataset and their last. Another is "prolificness": the number of publications. We also use the eccentricity metric, which directly measures whether nodes are peripheral or central. This metric is $O(n^3)$, so we use it only after simplifying the graph via clustering.

Convolution clustering is an approach to partitioning a graph that gives the user interactive control over how many clusters to create. Tulip calculates a density function based on the chosen metric, displays a convolution of its histogram, and partitions the graph according to the humps in the histogram. The user has a slider that controls the extent that the convolution kernel blurs the histogram; that is, whether there are a small number of wide low humps, or a large number of sharper humps. When we use the Strahler metric for convolution clustering, we can quickly segment the dataset by importance, and focus our attention on the relatively small number of nodes that have high value.

Small world clustering is a very different approach that creates a a recursive subdivision of the dataset, providing a simplified overview that shows the graph's high-level structure [3]. This approach uses heuristics to extract near-cliques; that is, nearly complete subgraphs where most nodes are connected to each other by edges. These near-cliques are collapsed into supernodes, and we can then create a quotient graph where only a single edge is drawn

*LaBRI, Université de Bordeaux I, {maylis,auber,domenger}@labri.fr

†University of British Columbia, tmm@cs.ubc.ca

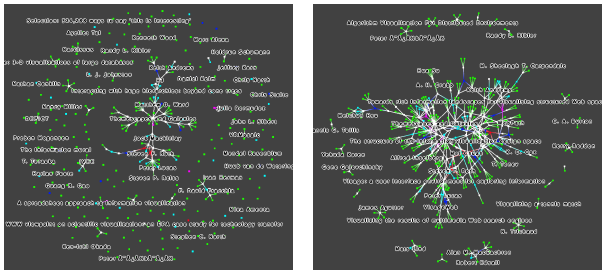

Figure 1: Author-conference evolution over time. **Left:** In 1997, only a few papers published at InfoVis cited each other. **Right:** By 2002, most were connected through co-citations

between two of these higher level supernodes if any exist between the two near-cliques. We argue that this clustering provides a useful mental map of the graph because once inside a supernode, everything is "nearby"; that is, there are only a very small number of hops between any of the nodes it contains. We compute the small world clustering a *strength* metric that finds the number of cycles of length 3 and 4 passing through each edge, normalized by the maximum possible value. Moreover, we automatically find the optimal number of clusters by computing all possible clusterings, measuring their quality, and picking the best result. Tulip allows interactive exploration where one can either see the nested subgraphs drawn inside the screen extent of a supernode, or jump inside to see just the supernode's subgraph. Small-world navigation is useful when exploring an unfamiliar graph to quickly find the structure of complex components.

In all the images we create, unless otherwise indicated the graph layout used is Frick's GEM algorithm [5] as implemented in Tulip. Tulip also supports constrained motion, where moving nodes also moves the edges attached to them, so in some of our final images we made minor manual adjustment of the placement of graph components to maximize label readability in the final images. (There is no manual adjustment in the accompanying video.)

3 Results and Discussion

In Task 2.1 we look at the graph of all papers. Our results show that simply using the Strahler metric for coloring leads to the visual popout of highly cited work: Generalized Fisheye Views, Cone Trees, and Tufte's book. In Task 2.2 we use the graph of both papers and authors, looking at evolution over time by separating the data by year. All authors are always shown, but new papers are added year by year. In 1986, most of the literature is disconnected, with a small connected component featuring books by Tufte and Cleveland. That component gradually increases, until it encompasses the entire dataset by 2002. In Task 2.3, we filter the dataset to only include the papers published at InfoVis itself, along with their authors, as shown in Figure 1. In 1995, there is of course no cocitation between InfoVis papers. In 1996 a few cocitations appear, with the largest connected component around Themescapes and Galaxies. Figure 1 Left shows that this central core grows much larger in 1997, retaining the Themescapes paper at its heart, with only a few scattered cocitations that are disconnected from it. This pattern continues in later years, with the intercitation connection structure in the central core growing more complex. Figure Figure 1 shows that in the final year of 2002, only 20 papers drawn around the periphery are completely disconnected from any other InfoVis paper, and a 21st connected component near the bottom, with the label of Alan MacEachren visible, shows connections between three geographic visualization papers.

For Task 2.4, we show the full dataset of authors, conferences, and papers, and again separate out the data year by year. Using the Strahler metric for convolution clustering is a very powerful way to track the evolution of central topics over time. We find that Focus+Context started strong and became even more dominant. Dynamic queries, ZUIs, brushing/statistical graphics, and high dimensionality are four more strong topics. In general, we see that a first influential paper appears, and the topic expands to include more papers as time went on. Tufte's first book is an interesting exception to the pattern followed by the other highly-cited items: it stays singular through the entire dataset, never forming a connected component with others. The clusters we create also show the relationship between the top authors and which of these authors publish in which area, as described in our detailed entry.

In Task 3.2.2, we show an overview of all authors and conferences, coloring the data by three different metrics. In Task 3.2.3, we find the top authors according to Strahler metric using convolution clustering. The top tier has Card and Shneiderman; the next has Mackinlay, the third highest is Roth, Robertson, Keim, and Stasko; the fourth is Chi, Bederson, Eick, Rao, Pirolli, Ward, and Brown; the fifth has 26 authors, and the last has all the rest. In Task 3.2.4, we show how small world clustering and coloring the quotient graph edges by eccentricity metric yields an easily navigable overview of the data.

4 Open Problems and Conclusion

The most pressing need that we found when exploring the contest graphs is a scalable algorithm for force-directed placement. Although there has been a flurry of recent work in scalable force-directed placement, the literature unfortunately only contains algorithms suitable for mesh-like graphs. The graphs typically found in infovis applications are not meshes. One of the reasons we made heavy use of clustering, both convolution and small-world, is that we could not create a useful drawing of the entire graph at the lowest level of detail.

Tulip's ability to find structure in this dataset shows that using combinatorics to guide visualization is a very powerful approach. The Strahler metric is useful for many functions: coloring, progressive rendering, label drawing, and clustering. We found that both convolution clustering and small world clustering provided insight into the structure of the dataset. We also extensively used graph-theoretic features such as finding the induced subgraph or reachable subgraph for some set of selected nodes and edges.

5 Acknowledgements

Munzner's visit to Bordeaux was funded by Université Bordeaux 1. The other authors were funded by ACI masse de Donnés.

References

[1] D. Auber. Using Strahler numbers for real time visual exploration of huge graphs. In *International Conference on Computer Vision and Graphics*, pages 56–69, September 2002.

[2] D. Auber. Tulip : A huge graph visualisation framework. In P. Mutzel and M. Jünger, editors, *Graph Drawing Software*, Mathematics and Visualization, pages 105–126. Springer-Verlag, 2003.

[3] D. Auber, Y. Chiricota, F. Jourdan, and G. Melançon. Multiscale visualization of small world networks. In *Proc. IEEE Symposium on Information Visualization (InfoVis '03)*, pages 75–81, 2003.

[4] D. Auber, M. Delest, and Y. Chiricota. Strahler based graph clustering using convolution. In *8th International IEEE Conference on Information Visualisation, London, England*, to appear 2004.

[5] Arne Frick, Andreas Ludwig, and Heiko Mehldau. A Fast Adaptive Layout Algorithm for Undirected Graphs. In *Proc. Graph Drawing '94, LNCS 894*, pages 388–403. Springer-Verlag, 1994.