Visualizing Citation Networks Over Time

Jason Portenoy
University of Washington Information School
jporteno@uw.edu

Muhammed Raza Khan
University of Washington Information School
mraza@uw.edu

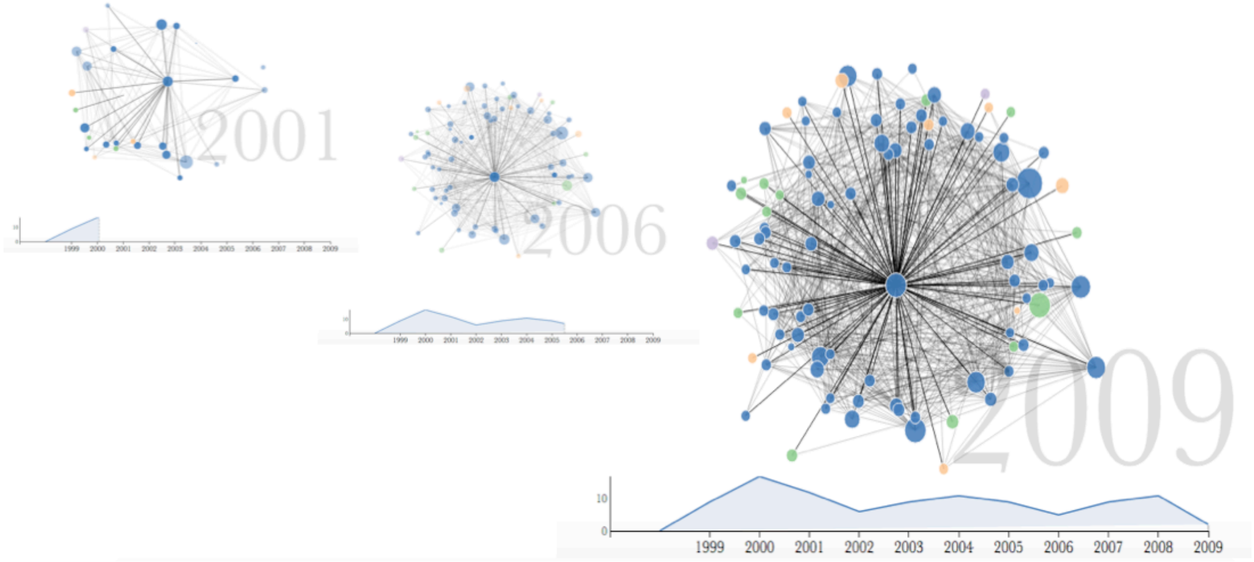


Figure 1. Three screenshots from the citation network visualization as it develops over time. On the right is the final state of the visualization.

ABSTRACT

We develop a novel adaptation of the node-link diagram for representing scholarly literature as networks of papers connected by citations. In our visualization, the influence of a particular paper is viewed over time. The viewer watches as the network builds year by year following the paper's publication, seeing the relative impact that the paper has had both within its own field and across other fields. As the narrative visualization progresses, important points---such as a citation by a paper in a different field---are identified, and annotations are automatically generated which give the viewer context.

INTRODUCTION

The quantitative analysis of scholarly publication is a dedicated field of study in its own right. The field of scientometrics, and the closely related field of bibliometrics, seek to apply statistical techniques to better understand things like the history and development of scientific knowledge and the impact of ideas [10]. One powerful technique that has been used in this space involves the conception of scientific literature as a *network,* with publications (or journals, or researchers) being the nodes and citations being represented as edges between them [9]. This takes into account the crucial ideas of information flow and the interdependence of ideas.

One problem that arises in representing scholarship in this way is how to make a meaningful visual representation of these relationships. One typical approach is the node-link diagram. However, these graphs suffer from becoming very cluttered and overwhelming when used to represent data sets of more than minimal size or complexity. Our contribution to this space is the design of a tool that visualizes the citations received by one paper over the course of time. The network surrounding a particular paper is built out over time for the viewer to watch. Using this approach, we hope to both aid in the discovery of patterns relating to the influence that paper has had, as well as provide a meaningful and emotionally resonant experience for the viewer as she follows the narrative of how related discoveries have occurred following the publication of the central paper. A related contribution is the use of dynamically calculated annotations that pause the progression of the visualization at interesting points, using heuristics to identify these points and providing information about them automatically.

RELATED WORK

Visualizing Citation Networks

Visualization of scholarly citation networks has been gaining attention as a research topic as it can help in identifying important research papers in a field, identifying scholarly communities, and assessing the impact of articles, authors, and journals within and across research disciplines

[4,7]. The Network Workbench (NWB) tool has been one attempt at visualizing scholarly networks. The NWB is a suite of tools to analyze networks, and it has been used to visualize co-authorship networks and citation networks in various ways [1]. The Action Science Explorer (ASE), developed at University of Maryland, is another suite of tools that includes visualizations of citation networks that can surface patterns such as clusters and identify impact based on network structure [4]. The CitNetExplorer tool allows for visualization of direct citation networks. None of these tools focus on providing a picture of how influence has developed over time [5].

The CiteSpace II tool, developed at Drexel University, uses co-citation networks to analyze the change in scientific fields over time. This approach uses article text and language processing to try to identify shifts in fields. It does not address visualizing the influence of specific papers or authors [3].

Narrative Visualizations

Segel and Heer present an overview of different techniques in narrative visualization, discussing a series of case studies of different ways of telling stories with data, and various strategies that have proven successful. Two of their findings are of particular note to us. One is the use of annotation to guide the viewer and aid in conveying a story. The other is the concept of the "martini glass structure," which involves a tight narrative path (the stem of the martini glass) which opens up to a mode of free exploration (the body of the glass). Both of these ideas were important in how we decided to approach the narrative element of our visualization [8].

METHODS

The data for this study is in the form of a database curated by Microsoft containing information on scholarly publications, and also the citation links between those papers. We pulled a subset of the data containing just a central paper of interest and the papers published since that have cited it.

As an initial proof of concept, we chose one paper to visualize using our technique. Future work will build these methods out into a more generalizable framework so that other papers can be visualized with little to no preprocessing or overhead. The paper whose influence network we chose to visualize was:

Robert V. Skibbens, Laura B. Corson, Doug Koshland, and Philip Hieter. 1999. Ctf7p is essential for sister chromatid cohesion and links mitotic chromosome structure to the DNA replication machinery. *Genes & Development* 13, 3, 307–319.

The visualization begins with just this paper, represented as a single node. An annotation appears alongside the node giving information about the paper and informing the viewer that he will be shown the impact it has had since it was published. As the years progress, new papers appear around the central paper and send out links to the central node as well as to other papers that appear in this network.

Visual Encodings of Influence

We chose three dimensions related to the influence of the central paper to encode visually. One is the *number of citations* the paper receives. This is encoded in two ways as the visualization proceeds: (1) the number of nodes that appear, and (2) a line graph showing the frequency of citations year-by-year that tracks along with the temporal development of the network graph above it.

The second dimension of influence is a numeric metric of the influence of all of the papers in the network called the Eigenfactor. The Eigenfactor takes into account the relative importance of each node in the network based on the relative importance of the nodes that link to it, using a recursive approach [11]. This method is similar to the PageRank algorithm that Google employs to rank the relative importance of web pages [6]. In our visualization, the radius of each node is calculated as a fixed value plus a scaled Eigenfactor value.

In addition to looking at the relative impact an article has had in general, we can also look at to what extent this impact is isolated to the same field as the article versus whether the influence spreads to other disciplines. Our third dimension takes this into account using color encoding. To assess whether an article in the network belongs to a different research field, we use a community detection algorithm based on the map equation objective function, which yields a multilevel hierarchical mapping of the full citation network [7]. This algorithm assigns a community address to each node based on the structure and information flows within the network; we then examined some sample papers from the different communities to qualitatively determine the appropriate label for each one that appears in our data set. We determined that our central paper is in the field "Genetics" (sub-field "DNA Replication"), and that the two other top-level clusters represented in our data correspond to the fields "Immunobiology" and "Cancer and Genetics".

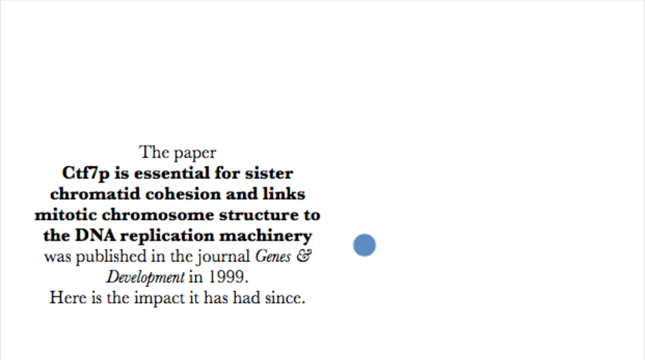

The paper
Ctf7p is essential for sister chromatid cohesion and links mitotic chromosome structure to the DNA replication machinery
was published in the journal *Genes & Development* in 1999.
Here is the impact it has had since.

Figure 2. The Initial annotation introducing the central paper

The Network Visualization

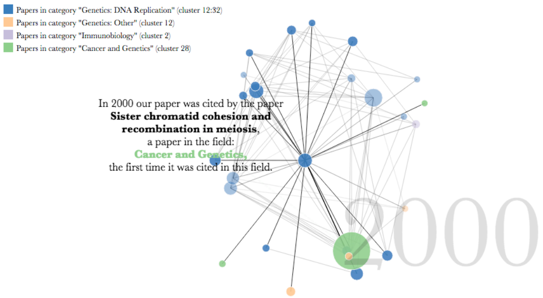
We used D3, a popular open-source JavaScript library, to build our visualization. This library takes advantage of existing web frameworks to produce data visualizations that can be dynamically created on a client's device, allowing for easy dissemination and minimal overhead for a user [2]. We used D3's force layout to position the nodes, which updates positions based on dynamic calculations of repelling forces between nodes and attracting spring forces between links.

When the web page is first loaded, the nodes and edges of the network are plotted, but all of the elements of the graph are hidden from the viewer. The force layout algorithm is run 1,000 times, and then animation begins. Year by year, each node is revealed one at a time based on year of publication. (The resolution of our data is only at the year level, so the order of nodes appearing within a particular year is arbitrary. Future work can seek to order the appearance to minimize nodes that send out links to nodes that haven't yet appeared.) After each node appears, it sends out a link to the central node. It also sends out links to any other papers in the graph that it cites; these links are faded to reduce the overall density of the visualization. After each year completes, the nodes and links for that year are faded to yield more emphasis to the year currently being plotted. When the final year has been plotted, all of the nodes and links are faded back in to their original state, and the nodes may be manipulated by clicking and dragging. Details about the individual papers appear below the visualization upon mouseover.

An experimental feature---"Disconnect Ego Node"---removes the central node from the network and yields a free-floating force network of the remaining nodes and links. The idea behind this is to show the network around the center paper had that paper never existed. Future work will experiment with ways of making this feature more meaningful and relevant.

Responsive Annotations

Another technique we used to increase engagement with the viewer and counteract the complexity of the network visualization was the use of annotations that appear at important points in the animation to provide context. These relevant points and the associated text are determined automatically based on the data. At the beginning of the animation, an annotation introduces the central paper, including its title, year of publication, and journal or venue of publication (see Figure). Another annotation appears whenever the central paper is cited by an outside field. The animation pauses,[1] the node size increases temporarily, and text appears to explain what is going on and identify the paper and its field (see Figure). In addition, the new field, encoded by a new color, is added to the legend at the top of the graph at this point. Future versions of the visualization will expand this feature and include more heuristics for identifying other important points in time, such as the year the central paper was most heavily cited and citations by highly ranked (high Eigenfactor) papers within and outside of the central paper's field.



🟦 Papers in category "Genetics: DNA Replication" (cluster 12:32)
🟧 Papers in category "Genetics: Other" (cluster 12)
🟪 Papers in category "Immunobiology" (cluster 2)
🟩 Papers in category "Cancer and Genetics" (cluster 28)

In 2000 our paper was cited by the paper
Sister chromatid cohesion and recombination in meiosis,
a paper in the field:
Cancer and Genetics,
the first time it was cited in this field.

Figure 3. An annotation that appears when the central paper is cited by the first paper from a different field.

RESULTS

Figure 1 shows a series of screenshots of our visualization, ending at the final view for this paper. However, as this is an interactive and animated visualization in which movement and transitions are very important, the reader will have a better sense of it by viewing it at http://cse512-15s.github.io/fp-jporteno-mraza/.

Formal user testing will be very valuable in terms of assessing whether our visualization achieved its goal of improving the interpretability of the citation network graphs and providing an engaging and emotionally resonant viewing experience. It will also help to guide future directions of development by asking viewers---especially authors of research papers---what features they would like to see, and how they might respond to different ways of showing the data.

Experimenting with data sets of varying complexity, we did find that performance can suffer with larger, more complicated networks. Having our graph layout be static rather than continually having the force algorithm update the layout helps with this. However, it should be noted that besides considerations of computation, large and complicated network graphs have a separate but correlated problem---that these larger networks are much more difficult for a viewer to interpret. This was the rationale behind our limiting the data set only to the central paper and the papers that directly cited it. We also set a cutoff at a maximum of 100 nodes total---past this point we remove low-Eigenfactor nodes from random years. The resulting visualization runs smoothly on a Macbook Pro with 4GB of RAM.

[1] Because JavaScript is meant to be run in a web environment with many concurrent asynchronous events, it can be difficult to coordinate timed events. We ran into difficulty executing the annotation alongside a pause of the animation. In its current state, the animation completes the year it is currently running through before pausing. Future versions will pause the animation right when the relevant node appears.

Each year is set to take a fixed amount of time (5 seconds) to complete, so years with more citations have nodes that appear quicker than years with fewer citations. In its current version, the animation takes approximately 1.5 minutes to run before the final state is reached.

DISCUSSION AND FUTURE WORK

Overall, our visualization achieves its goal of providing an engaging experience along with some insight into the impact of a research paper within and outside of the paper's discipline. Although a full user study is outside the scope of this paper, informal feedback suggested that viewers respond positively to seeing the network build over time and provide context through dynamic annotation.

One idea that will be refined in future iterations of this work is the use of spatial encoding. In this version, the location of nodes around the center node is mostly arbitrary, placed according to the force algorithm. Many users commented that they expected the placement of the nodes to have some relationship with the data or the insight that could be gained. Experimenting with different spatial encodings---for example, making sure papers in related fields are grouped together, and distancing nodes radially from the center based on year of publication---could make for a more effective visualization.

Another feature that could enhance the effectiveness of the story is a temporal encoding of how the Eigenfactor changes over time. This shows another metric of the paper expanding its influence year by year after being published. Currently, the Eigenfactor is calculated for each node as a static value as of the time the data were collected, but it is possible to go back and recalculate the Eigenfactor at each point in time (i.e. each year). Each node will start out with a fixed radius with a halo around it showing its final Eigenfactor (or rather its final Eigenfactor in the database---the Eigenfactor could increase in the future). As the animation plays, the radius will expand to fill this halo, and the viewer will get to witness the impact of the paper growing over time.

Although the current version only shows one paper, we designed the framework to be scalable and generalizable. Building on what we have done, this tool will be able to generate a narrative visualization for any research paper---although more influential papers that have been cited a fair amount will tend to produce more interesting visualizations. This is certainly a feature that many of our informal users wanted to see, and we are excited at what we could surface by viewing different types of papers, comparing multiple papers at the same time, or viewing multiple papers by the same author at the same time in order to show more of the overall influence of a particular researcher.

Future work will include formal user testing with different types of users, in order to determine how effective these techniques are and what other features various users would want out of a visualization like this. Of particular interest is testing the visualization with the author of the paper being visualized. To test whether we achieve our goal of providing an emotionally resonant viewing experience, it would be very valuable to have this author watch the story of her paper as it unfolds over time, especially for a career-defining paper of which she is particularly proud. Perhaps she will recognize citation events that the annotations point out, or perhaps she will gain new insights on her own work by reliving the story.

REFERENCES

1. Katy Börner, Weixia Huang, Micah Linnemeier, et al. 2010. Rete-netzwerk-red: analyzing and visualizing scholarly networks using the Network Workbench Tool. *Scientometrics* 83, 3, 863–876. http://doi.org/10.1007/s11192-009-0149-0

2. Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D3: Data-Driven Documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*. Retrieved from http://vis.stanford.edu/papers/d3

3. Chaomei Chen. 2006. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology* 57, 3, 359–377. http://doi.org/10.1002/asi.20317

4. Cody Dunne, Ben Shneiderman, Robert Gove, Judith Klavans, and Bonnie Dorr. 2012. Rapid Understanding of Scientific Paper Collections: Integrating Statistics, Text Analytics, and Visualization. *J. Am. Soc. Inf. Sci. Technol.* 63, 12, 2351–2369. http://doi.org/10.1002/asi.22652

5. Nees Jan van Eck and Ludo Waltman. 2014. CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *arXiv:1404.5322 [physics]*. Retrieved May 20, 2015 from http://arxiv.org/abs/1404.5322

6. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank Citation Ranking: Bringing Order to the Web. Retrieved June 11, 2015 from http://ilpubs.stanford.edu:8090/422/

7. M. Rosvall and C. T. Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* 105, 4, 1118–1123. http://doi.org/10.1073/pnas.0706851105

8. E Segel and J Heer. 2010. Narrative Visualization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics* 16, 6, 1139–1148. http://doi.org/10.1109/TVCG.2010.179

9. D. J. de Solla Price. 1965. Networks of Scientific Papers. *Science* 149, 3683, 510–515. http://doi.org/10.1126/science.149.3683.510

10. D. J. de Solla Price. 1978. Editorial statements. *Scientometrics* 1, 1, 3–8. http://doi.org/10.1007/BF02016836

11. Jevin D. West, Theodore C. Bergstrom, and Carl T. Bergstrom. 2010. The Eigenfactor MetricsTM: A Network Approach to Assessing Scholarly Journals. *College & Research Libraries* 71, 3, 236–244. http://doi.org/10.5860/0710236