Leveraging Citation Networks to Generate Narrative Visualizations of Scholars' Careers

Jason Portenoy[1], Jevin D. West[1]
[1]University of Washington, School of Information

Abstract
We present an approach for generating dynamic narrative visualizations of scholars' careers. This approach includes an animated node-link diagram which shows the citation network accumulated around the researcher over the course of the career, with nodes and links appearing as the representation of the career progresses. Additional data provide more richness to the narrative, including timelines of key indicators, career milestones, and excerpts from qualitative interviews with the scholars. The intended audiences for this work include the scholars, who can enjoy and gain insight from a new way of looking back on their careers, and funding agencies, who have an interest in finding ways to evaluate the impact that their scholars have had.

Keywords: Information Visualization; Dynamic Network Visualization; Citation Networks; Narrative Visualization; Evaluating Scholarly Impact
DOI: 10.9776/16572

Acknowledgements: We would like to thank the Pew Charitable Trust for funding and for allowing us to interact with program managers and the Pew Biomedical Scholars.
Contact: jporteno@uw.edu, jevinw@uw.edu

1 Introduction

The question of how best to assess the impact of scholarship is a wide area of research with many possibilities. This poster presents work and future directions in combining methods to create a narrative visualization that presents a view of a researcher's career impact while also providing the opportunity to gain new insights and ask new questions about the data being presented. The methodology combines quantitative techniques such as citation network analysis and bibliometrics with qualitative data including transcripts and audio from interviews with the scholars. The primary audiences for this work are the scholars themselves, and the organizations that fund research and have an interest in evaluating its impact.

Any organization that provides funding for scholarly research is continually faced with the question of how best to evaluate the impact that the funding has had. These organizations—including nonprofits and government agencies such as the National Institute of Health—collectively spend billions of dollars each year to fund research, and have a strong interest in making sure that their spending is having an impact. A variety of approaches to this problem have been used, some employing methods from the fields of scientometrics and bibliometrics, which seek to apply statistical techniques to better understand things like the history and development of scientific knowledge and the impact of ideas (de Solla Price, 1978). One powerful technique that has been used in this space involves the conception of scientific literature as a *network,* with vertices representing publications, journals, researchers, or organizations, and edges representing citations between the publications (de Solla Price, 1965). This takes into account the crucial ideas of information flow and the interdependence of ideas.

One challenge that arises in conceptualizing scholarship in this way is how to make meaningful visual representations of these relationships. One typical approach is a node-link diagram, which visualizes entities such as papers, authors, or organizations as nodes, and citations as links or lines between them. While these diagrams can be useful, they tend to be too abstract and overwhelming in their complexity when presented to the average viewer. Our approach, building on previous work in narrative visualizations (Segel & Heer, 2010) and dynamic network visualizations (Beck, Burch, Diehl, & Weiskopf, 2014), shows a node-link diagram with one scholar at the center, but builds the citation network out over the course of the scholar's career, allowing the viewer a sense of the developing complexity. This framework can then be combined with qualitative data—interviews with the scholars intended to have them reflect back on the course of their careers—that

provide more context and richness to the narrative. The overall effect is a data-driven method to generate an effective and even emotionally resonant narrative of a scholar's career as it unfolds.

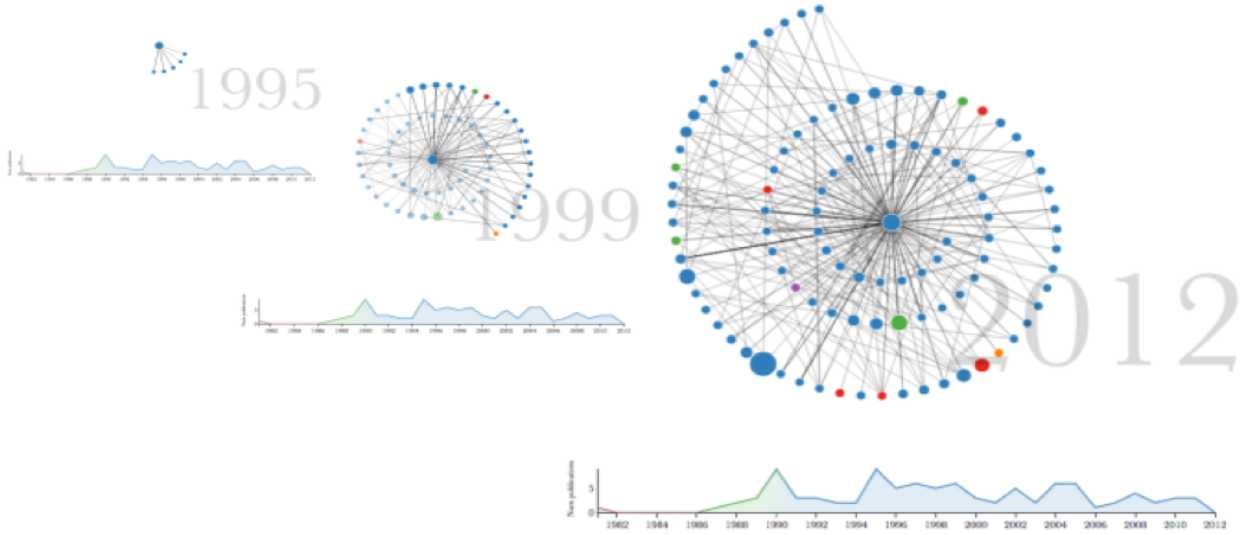

Figure 1: Three screenshots from the citation network visualization as it develops over time. On the right is the final state of the visualization.

2 Methods

2.1 Case Study: Pew Scholars Program

The Pew Scholars Program in the Biomedical Sciences provides four years of early-career funding to young researchers in health-related fields. This program and its subset of scholars provides a useful case study for exploring narrative visualizations of scholarly impact that tell a compelling story while also providing some useful insights. The Pew program is highly selective, and its scholars tend to have a great deal of impact in the biomedical sciences. The program has been in existence for thirty years, and so claims a large number of scholars with a rich history. These details make the Pew program an ideal case to explore narrative visualizations of scholarly impact; however, the methods we develop are generalizable to other funding organizations and other scholars in general.

2.2 Data

The data we use come from three different sources. The first is a database curated by Microsoft which contains information on 49 million scholarly publications and 262.5 million citation links between them. The second is the Pew program, which provided supplementary data on its subset of scholars such as funding periods and demographic information. The third is the audio and transcripts of interviews conducted individually with these scholars over the past several decades by the Chemical Heritage Foundation, with whom the Pew program has partnered.

2.3 The Visualization

In our current approach, each visualization we construct centers on one scholar, although the framework can also be used with an alternative focus, such as individual papers or groups of authors. We use D3, a popular open-source JavaScript library which can create dynamic visualizations of complex datasets (Bostock, Ogievetsky, & Heer, 2011). The main section is a node-link diagram with the central node representing all of the papers that the scholar has written. The other nodes radiate out in a spiral formation and represent

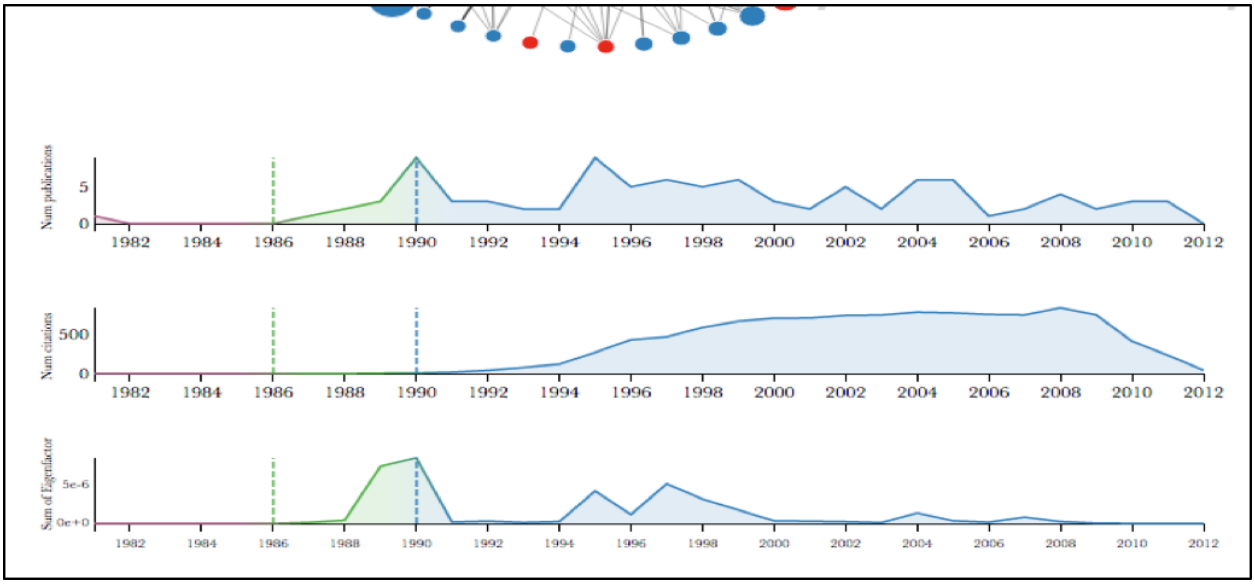

Figure 2: Integrated timeline charts below the network visualization. Top: Number of publications by the central author by year. Middle: Number of citations received by the central author by year. Bottom: Sum of the Eigenfactor (see section 2.3) for all of the publications published by the central author in each year.

other papers that have cited the scholar of interest, ordered by year of publication so that the radial distance encodes time. The spiral positioning allows the visualization to encode time as a function of distance while also condensing the placement of nodes, allowing for the inclusion of more nodes than would be possible were the nodes placed linearly. While the total set of the nodes represents any paper that has ever cited a paper authored by the central scholar, in order to reduce the complexity of the graph we choose a sample of these nodes to visualize. We choose this sample based on the Eigenfactor metric, which ranks papers by the relative impact they have had in their own citation network. The Eigenfactor takes into account the relative importance of each node in the total network of papers and citations based on the relative importance of the nodes that link to it, using a recursive approach (West, Bergstrom, & Bergstrom, 2010). This method is similar to the PageRank algorithm that Google employs to rank the relative importance of web pages (Page, Brin, Motwani, & Winograd, 1999). The nodes are sampled, taking the most influential (highest Eigenfactor) nodes first, and more complete statistics are visualized in timeline representations below the network graph.

We explore representing the idea of influence in several different ways, and so the network diagram features several different visual encodings of influence. The *size of the nodes* is scaled by the Eigenfactor metric of each paper, so that larger nodes are easily identified as more influential papers. The *color of the nodes* represents a different aspect of influence: it encodes the research field of the paper. This allows a view not only of the impact that an author has had in general, but to what extent this impact is isolated to one field versus whether the influence spreads to other disciplines. An author with a more monochrome network will have had most of his or her influence within one particular field, while a network with more color means more citations from researchers in other fields.

Presenting a network diagram to a viewer in this way can be overwhelming, as the mass of nodes and links tends to elicit a "hairball" effect. This complexity is important, and we want to give the viewer a sense of how this complex network develops over the course of the scholar's career. We present the diagram as a dynamic visualization, with the nodes and links appearing as the narrative moves forward in time. The year-by-year progression also tracks along with the key indicator timelines below the graph (see Figure 2).

3 Results and Future Directions

Figure 1 shows a series of screenshots of the visualization for one author, ending at the final view for this paper. Figure 2 shows a static view of the integrated timelines underneath the network diagram.[1] With this framework in place, we can explore ways of incorporating different types of data into the visualization to provide more richness to the overall narrative. For example, in Figure 2 we have added colors to indicate the funding period of the Pew program, showing the periods before, during, and after this early-career funding. Future work will bring in more relevant data, such as text or audio from the interviews with the scholars. Interactive elements could appear at certain points during the narrative which would access relevant interview excerpts relating to pivotal career periods or collaborations.

This framework also provides opportunity for adding elements of *automated storytelling.* It is possible to, for example, link the citation network with the interview transcripts to pull in excerpts of the interview where the scholar has mentioned one of the papers in the network. Data in the network can also be analyzed to provide responsive annotations, highlighting to the viewer important moments such as when the central scholar is cited by a paper in another field.

One limitation of this work is that it is difficult to compare narratives. For funding agencies especially, it is important to be able to compare the careers of scholars against baseline measures. A key area of future work will be exploring ways to provide more context for these visualizations, and aggregating representations to compare different groups.

References

Beck, F., Burch, M., Diehl, S., & Weiskopf, D. (2014). The state of the art in visualizing dynamic graphs. *EuroVis STAR*. Retrieved 2015-09-21, from http://www.visus.uni-stuttgart.de/uploads/tx_vispublications/eurovis14-star.pdf

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

de Solla Price, D. J. (1965, July). Networks of Scientific Papers. *Science*, *149*(3683), 510–515. Retrieved 2015-06-11, from http://www.sciencemag.org/cgi/doi/10.1126/science.149.3683.510 doi: 10.1126/science.149.3683.510

de Solla Price, D. J. (1978, September). Editorial statements. *Scientometrics*, *1*(1), 3–8. Retrieved 2015-06-11, from http://link.springer.com/article/10.1007/BF02016836 doi: 10.1007/BF02016836

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999, November). *The PageRank Citation Ranking: Bringing Order to the Web.* [Techreport]. Retrieved 2015-06-11, from http://ilpubs.stanford.edu:8090/422/

Segel, E., & Heer, J. (2010, November). Narrative Visualization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 1139–1148. Retrieved 2015-06-11, from http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5613452 doi: 10.1109/TVCG.2010.179

West, J. D., Bergstrom, T. C., & Bergstrom, C. T. (2010, May). The Eigenfactor MetricsTM: A Network Approach to Assessing Scholarly Journals. *College & Research Libraries*, *71*(3), 236–244. Retrieved 2015-06-11, from http://crl.acrl.org/content/71/3/236 doi: 10.5860/0710236

[1] An interactive demonstration of the visualization is available online at http://scholar.eigenfactor.org