# Visualizing citation networks of academic fields over time as a tool for novice researchers

David Mora*
Stanford University, CS 448b Fall 2017

## ABSTRACT

Understanding of how research papers, authors, and applications fit together remains a central task for academic researchers, yet this information remains largely trapped inside the minds of experts with years of experience. This visualization of academic citations over time aims to tackle the challenge of giving novices researchers exactly that sort of expert perspective on a field's "knowledge landscape," and a way to discuss and refine it with with others. Selecting the field of positive psychology as a test case, I created a network visualization of academic citations which allows users to dynamically filter across time and author in hopes of illustrating how the field relates and evolves. As this visualization aimed to provide a initial, not exhaustive, perspective, the vis only also shows highly influential papers which cite or are cited by the currently selected authors, with influenced determined by a SemanticScholar. Initial feedback from both experts and novice researchers was positive and users reported higher confidence and much faster time in finding papers to read and gaining insight into a new field.

**Keywords**: graph visualization, citation networks, graph evolution over time

## 1. INTRODUCTION

Scientific visionary and administer Vannevar Bush wrote his 1945 essay *As We May Think,* "scientific publication has been extended far beyond our present ability to make real use of the record," calling on the necessity of improved data retrieval and interaction technologies [1].

Today, this is challenge proves tougher than ever, especially for research newcomers today who, entering new fields with little any prior context, face some subset of the 2.5 million new papers each year, as well as the burden of understanding existing research. This overwhelm proves more than an an annoyance: as fields continue to grow, researchers risk wasting precious grant money on research out of touch with the ecosystem, or even redundant to existing work.

Beyond merely avoiding costly pitfalls, presenting novel, more effective ways to understand the evolving landscape of a scientific publication holds potential to accelerate research reviews, give novel insights, and spark creative connections between works of research which would have otherwise been missed.

*davidmora@cs.stanford.edu



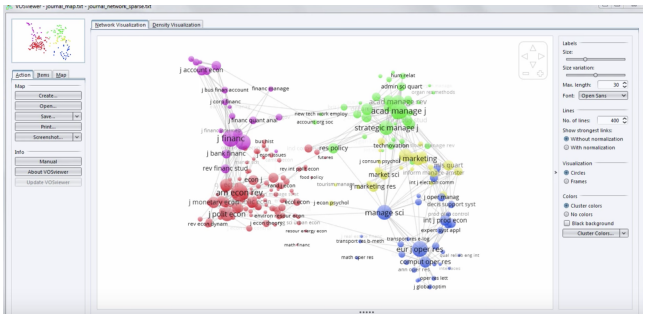
## 2. RELATED WORK

Several works regarding visualization and aggregation of scientific citation networks exist, however, as I will discuss, none offer views or interactions with the network accessible and amenable to the needs of new researcher to spot trends, understand relationships across times, and identify core papers and authors.

### 2.1. VOSviewer

VOSviewer is a general purpose, open source tool aiming to make it easy to import huge networks (particularly from major sources such as Web of Science, copus, Crossref, PubMed) and visualize co-authorship networks, word co-occurrence networks, and citation-based networks [2].

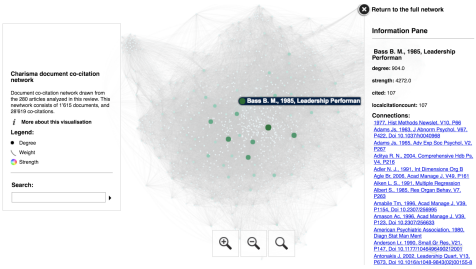
While efficient and flexible, VOSviewer is optimized for synthesizing huge networks but doesn't provide resolution to understand or interact with the finer details like a novice researcher might want to. Even more, VOSviewer provides only static graphs, hiding how academics networks grow and shift over time.



1. VOSviewer sample screenshot

### 2.2. Charisma Study Graph

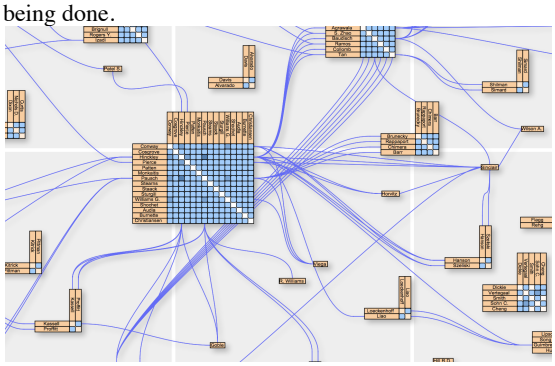
An interactive visualization attempting to show, in finer resolution than VOSviewer, show studies relating to the attribute of charisma displays a dense cloud of citations across 1,615 documents, and 28,619 co-citations [3]. While presenting an attempt to show the structure of a field, the vis feels largely cluttered and edges obscure any discernible structure, preventing effective navigation without prior expertise of the papers. Similar to VOSviewr, the charisma papers vis (beyond a user manually inspecting each node) hides the chronological order of citations and their distribution across time. Further, it fails to identify which authors have contributed to what degree, a core need of new researchers orienting themselves to major contributers.




2. *Charisma Graph* sample screenshot

### 2.3. Nodetrix

Nodetrix looks to help synthesize dense graphs by transforming dense components into adjacency marrixes while maintaining edges between clusters [4]. While an effective technique for de-cluttering dense graphs, the tool conveys very little meaning to someone unfamiliar with author names, and also fails to elucidate the direction or order of citations, or where the most recent work is being done.



3. *Nodetrix* sample screenshot

### 2.4. Animated Transitions

Jeffrey Heer and George Robertson revealed that animated transitions can significantly improve graphical perception [5]. Particularly relevant to the challenge of introducing new researchers to an evolving field is two types of animations applications identified in the paper: *filtering operations* and *viewing changes of data over time*. To understand and investigate a field, researchers must both filter among the contributions as well as understand how contributors change tactics and focuses across time.

## 3. Methods

### 3.1. Design and Planning

To investigate visualization of academic fields over time, I selected positive psychology for several key reasons: first, it's a fairly young field and excellent documentation exists for its entire lifespan; second, while significant in size, the number of authors is reasonably small, allowing for manageable assessment of results and improvements. After consulting with several experts and educators in the field of positive psychology, I create a cross-reference list of the most influential researchers which those wishing to understand the field should begin by studying. Those 19 authors are referred to henceforth as "core authors" [see Appendix 1 for the complete list].

### 3.2. Data Collection

While the current gathered data focuses on the 19 positive psychology research authors, all data collections methods were written as general python scripts, so any arbitrary list of authors could be input and the data constructed, conceivably done server-side in the future.

In collection the data, I first scraped Semantic Scholar to match author names to their unique Semantic Scholar identification numbers, then used a series of Semantic Scholar API calls do download every paper written by each author, and compiled a list of all papers which cite or are cited by those compiled core author papers. The total 31 MBs of citations spanned 93,0000 papers.

This network was huge, and incredibly dense, and afforded nothing but a mess of links and edges, similar to the *Charisma* visualization. To remedy, I filtered down only to nodes who were marked by Semantic Scholar as "highly influential" on each other, a technique they developed specifically for academic papers combining machine learning, graph analysis, and NLP [6]. The filtered graph in the end contains 9,000 nodes 11,000 edges with metadata such as year and authors for each node, though in practice a majority of nodes and edges are filtered away as the vis is used.

### 3.3. Visualization

D3.js's force directed layout visualizes the citation network, dynamically satisfying the constraints of nodes trying to repulse each other, links drawing nodes together, a force pulling all nodes toward the center of the canvas, and a constraint maintaining nodes within the viewport.

## 4. Results


4. Academic Field visualization, showing the filtered dataset for 19 core authors of positive psychology across 1950 to 2017

The final visualization allowed you to filter any permeation of core authors's publications, show all papers one (undirected) hop away from core authors' papers, filter to display only papers published during a specific period of years, drag nodes to inspect dense clusters, and hover to inspect paper title and core author.

Testing the visualization with its target audience, I found two prominent improvements the vis provide over traditional methods of exploring and assessing features of an academic field new to you:

1. Novice researchers in positive psychology or related fields were able to identify the most important 5 papers for themselves to read based on their current research question within two minutes, and reported having a higher confidence level in their

choices as opposed after previously to do the same by spending hours, even days, reading papers individually and searching keywords online.
2. Users with no prior knowledge of the core authors were successfully able to identify the originators of specific fields merely by inspecting the visualization by cluster, author color, and paper titles. For example, users correctly identified Carol S. Dweck and Kristin D. Neff as the founding researchers and current leading experts of their respective sub-fields, mindset and self-compassion.

I also found several key issues from testing:
1. **Lack of clear, differentiating names:** while colors indicate author, it's inefficient to have to hover every node to ascertain its paper name, and after hovering nodes become easily confused, particularly as the graph evolves and nodes move.
2. **Tracking nose over time:** similarly, while overall graph structure is clearly displayed over time, it's difficult to track individual papers in relationship to all the changes, and older nodes look indistinguishable from more recent nodes.

## 5. Discussion

Amid the extensive work in graph analysis, layout, and interaction already published, this paper contributes to the space by presenting a novel visualization and interaction techniques for evolving citation networks. After posing an initial visualization, this paper helps raise core questions in moving forward with more refined visualizations of this sort, specifically:
1. How do you do effectively display and track node information (such as paper title) across time in an evolving graph?
2. While the time slider was a useful for showing global evolution, as well as for isolating papers published within a specific time frame, it invites the question: how might you track and filter, not just a single node over time, but the preceding and post-ceding papers who's influenced depends on or extends to that node? For example, selecting node A might highlight nodes B and C which heavily influenced it in one encoding, and nodes which cited it in another encoding.

These questions help point future work toward richer, more effective tools for understanding academic fields.

## 6. Future Work

From my design, testing, and feedback with novice researchers and experts in various fields including and beyond positive psychology, several key opportunities for future work presented themselves:
1. **Filtering by paper keyword** would allow for richer exploration, particularly in letting users zero in on specific topics within authors' work, and quickly gage where different authors have focused their energies. Further, combined with a more generalized search across multiple fields, keyword filtering would allow for discovery of diverse connections between otherwise seemingly disconnected fields. For example, being able to look at the role of social support as discussed, say, both in positive psychology and in immunology might reveal techniques, paradigms, and perspectives otherwise missed or non-existent without comparison.
2. **"Longest path" discovery**: while identifying dense clusters proves illuminating, researchers I spoke with also emphasized the usefulness of identifying historically disconnected or unrelated sub-fields or concepts when evaluating assumptions and looking for new research opportunities. Identifying blindspots might well be a more powerful contribution of the visualization than just identifying existing clusters, which seasoned researchers are well aware of.
3. **Searching by study techniques:** From speaking with researchers, the ability to aggregate and filter papers not just by author, keyword, or time, but by the techniques used to measure and conduct the research itself seemed extremely useful. In particular for beginning researchers, identifying how to go about measuring research direction proves crucial, but still currently quite difficult.

All in all, I hope this paper helped lay the ground work for visualizing evolving scientific networks, as well as make claims for the usefulness of these networks to novice researchers.

## 7. Acknowledgements

Special thanks to Eleanor Collier, Donnovan Yisrael, and megaMegan Meyer's Darmouth social neuroscience lab for guidance, feedback, and inspiration. Much gratitude to code wizard Albert Sławiński for his javascript mentorship and debugging prowess. A humbled bow to the late, great Hans Rosling — this entire project is likely merely a subconscious attempt to recreate the thrills of his legendary TED talk.

## 8. Appendix

1. Core authors list, with their respective areas of expertise in positive psychology
    1. Martin Seligman - overall happiness research
    2. Sonja Lyubomirsky - Components of happiness, happiness habits
    3. Fred Luskin - forgiveness
    4. Kristin Neff - self compassion
    5. Kelly McGonigal - "stress mindset"
    6. Carol Dweck - growth mindset
    7. Mihaly Csikszentmihalyi - flow
    8. Angela Lee Duckworth - "grit"
    9. Richard Davidson - mindfulness + other stuff
    10. Jon Kabat-Zinn - meditation, MBSR (mindfulness based stress reduction)
    11. Barbara Fredrickson - happiness in general, broaden & build theory, positivity to negativity ratio
    12. Corey Keyes - flourishing vs languishing (dual-continua model); key requirements for wellbeing
    13. Emma Seppala - compassion, altruism
    14. Zindel Segal - mindfulness
    15. Dacher Keltner - emotion & social interaction
    16. Daniel GIlbert - book Stumbling on Happiness (has some interesting research on "affective forecasting")
    17. Jennifer Aaker - meaningful life vs happy life (don't know a ton about her but she seems legit)
    18. Dan Siegel - "the mind"
    19. Robert Emmons - leader on gratitude

2. …..

## References

1. Bush, Vannevar. "As we may think." The atlantic monthly176.1 (1945): 101-108.
2. N.J. Van Eck, L. Waltman. Software survey: VOSviewer, a computer program for bibliometric mapping,
3. Antonakis, J., Bastardoz, N., Jacquart, P., & Shamir, B. Charisma: An Ill-Defined and Ill-Measured Gift. Annual Review of Organizational Psychology and Organizational Behavior, 3(1).

4. Henry, N., Fekete, J. D., & McGuffin, M. J. (2007). NodeTrix: a hybrid visualization of social networks. IEEE transactions on visualization and computer graphics, 13(6), 1302-1309.
5. Heer, Jeffrey, and George Robertson. "Animated transitions in statistical data graphics." IEEE transactions on visualization and computer graphics 13.6 (2007): 1240-1247.
6. Valenzuela, Marco et al. “Identifying Meaningful Citations.” AAAI Workshop: Scholarly Big Data (2015).