Visualizing Scholarly Publications and Citations to Enhance Author Profiles

Jason Portenoy
University of Washington Information School
Seattle, WA 98195
jporteno@uw.edu

Jevin D. West
University of Washington Information School
Seattle, WA 98195
jevinw@uw.edu

ABSTRACT

With data on scholarly publications becoming more abundant and accessible, there exist new opportunities for using this information to provide rich author profiles to display and explore scholarly work. We present a pair of linked visualizations connected to the Microsoft Academic Graph that can be used to explore the publications and citations of individual authors. We provide an online application with which a user can manage collections of papers and generate these visualizations.

Keywords

author profiles, scholarly data, scholarly influence, citation visualization, citation networks, science of science, bibliometrics

1. INTRODUCTION

The rise of large-scale collection of data on academic publications has brought with it the opportunity for new ways of analyzing and visualizing scholarly activity and impact. One area in which this can be especially useful is in author profiles, where data about publications and citations can be used to show the influence of a particular researcher.

We present preliminary work on a pair of visualizations that can be used together to provide an interactive, engaging component of a scholarly author profile. Leveraging the open academic data provided by Microsoft[9], these visualizations allow a viewer to explore the publications and citations of an author over the course of a career. We provide a fully functional online application to demonstrate these author profiles.[1]

2. BACKGROUND

The two main services for large-scale, publicly available academic data offering author profiles are Google Scholar Citations and Microsoft Academic. Previous work has looked at these profiles [6][10]; however, this is a fast-moving area and the reporting on Microsoft in particular references the decommissioned version of Microsoft Academic Search, which has since been replaced by Microsoft Academic Service and the Microsoft Academic Graph, which is being actively updated.[2]

Current author profiles on these services are relatively bare-bones. Google Scholar lists publications with citation counts and co-authors, as well as a few influence metrics such as h-index. The h-index (the maximum number h so that h of an author's papers have each been cited at least h times [2]) provides a rough measure of scholarly impact, but has received criticism for problems such as bias along academic field, academic age, and gender [3, 5]. Microsoft Academic, for its part, does not include these impact metrics in its author profiles, but does include general information such as publications, areas of study, co-authors, and work that has cited the author's papers.

Despite the shortcomings of metrics such as h-index discussed above, they are widely considered in decisions behind hiring, promotion, and grant funding. Our goal in creating interactive data-driven visualizations for use in author profiles was to design for experiences and insights with which current profiles and metrics fall short. Visualizations can make underlying patterns in data clearer [4], and act as storytelling devices [8]. In particular, we wanted to go deeper into the kind of influence that a scholar may have had through incoming citations by facilitating exploration of the papers and fields of study that these citations come from.

3. THE DATA

3.1 Dataset

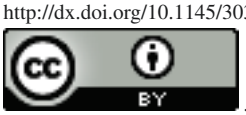
The dataset used in this application is the February, 2016 release of the Microsoft Academic Graph (MAG) [9], containing about 127 million papers and 528 million citations, as well as other information about authors, affiliations, journals/conferences, and keywords/fields. We used the citation graph to calculate article-level Eigenfactor [11] scores for all of the connected papers in the set ($N \approx 47$ million). We then loaded the data into a MySQL database that can be actively queried by the application.

To facilitate exploring influence between different fields of study, we assign a category designation to each paper using metadata supplied in the Microsoft dataset. The data

[1] `http://scholar.eigenfactor.org/`

[2] https://academic.microsoft.com/FAQ. Accessed Jan, 2017.

WWW 2017 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3038912.3038914



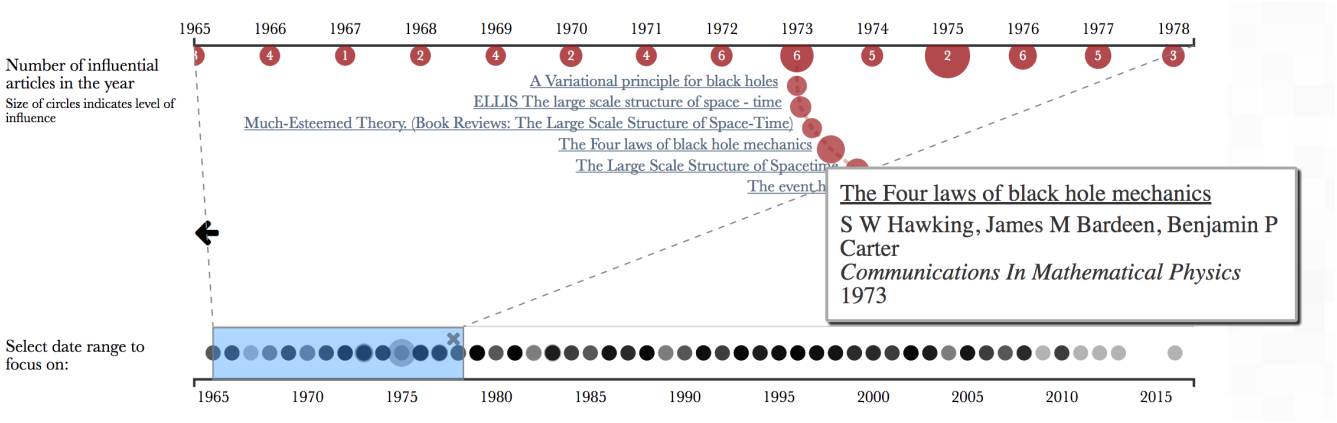


Figure 1: Timeline visualization for English theoretical physicist Stephen W. Hawking. The bottom displays all of the papers spanning his full career. The main display above shows the years within the blue brush at the bottom. The year 1973 is expanded to show the papers in that year, and more information is displayed for one of those papers via tooltip. The numbers within the year circles show the number of papers in each year, and the size of each circle is proportional to the influence (Eigenfactor) ranking.

contain keywords for many papers that have been mapped to a hierarchical Field of Study designation. Our current approach takes the second level of this hierarchy for each keyword of a given paper and assigns the majority Field of Study designation to that paper. In the event of a tie, all of the top categories are combined to make a new category (e.g. "Physics, Quantum Mechanics"). This provides an array of categories corresponding to an author's incoming citations that can be assigned colors when visualizing the data—the variation in color tends to correspond to the degree with which an author has had influence in different fields. We have also experimented with using different methods to assign these categories, such as journal and citation clustering [12].

3.2 Creating and managing paper collections

The MAG dataset is largely constructed from crawling the web for academic publications, and as such it can suffer from problems such as author identification and disambiguation. These errors can be problematic when visualizing patterns and notable features related to citation-based influence in the data. To counteract this problem, we allow the user to create, save, and manage collections of papers to use with the visualization tools. These collections typically represent a single author, but they can also represent groups of authors (e.g. labs), departments, topics/keywords, etc.

To build a collection, a user may log in to the system, then perform a search such as an author name. The search queries the Microsoft Academic API.[3] The search returns a list of unique Paper IDs, which we match to the papers in our database. This list of papers is often itself a good representation of the author's work, but the user may manage the collection herself, performing additional searches and adding and removing papers as necessary. In addition to the list of papers, the user can include metadata such as an author name and associated image.

Once the user is satisfied with a collection, the system can generate the citation data needed for the visualizations. This processing step is fast for smaller collections, but can take several minutes for collections with many citations. Once the processing is complete, the results are cached and can be used to generate the visualizations from then on until the user modifies the collection.

[3] `https://www.microsoft.com/cognitive-services/en-us/academic-knowledge-api`. Accessed Jan, 2017

4. PROFILE VISUALIZATIONS

A paper collection representing a scholar's work is used to generate a pair of linked visualizations of the author's publications and important papers that have cited the author's work. These visualizations were built using the open-source JavaScript library D3 [1]. A demonstration can be found at `http://scholar.eigenfactor.org/demo`

4.1 The Author Timeline Visualization

The author timeline visualization (Figure 1) shows the papers produced by an author by year. The smaller display at the bottom shows every paper; papers in the same year are sized by influence score (the article-level Eigenfactor score [11], a citation-based metric similar to PageRank) and overlaid on top of each other so that years with more papers appear darker. The user can brush along this view to select a region to view more closely in the main view on top.

In the main view, each year has an associated circle containing all of the papers for that year. When the viewer points her mouse at a year's circle these papers are revealed along with their titles. More information for these papers is revealed on mouseover, and the user can click to be taken to a page to view or purchase the paper.

4.2 The Citation Influence Visualization

The citation influence visualization (Figure 2 left) condenses all of the author's papers into a single central node, and shows the citation-based influence that these papers have had by showing this node surrounded by important papers that have cited the author's work. The design process by which this visualization was developed is described in [7].

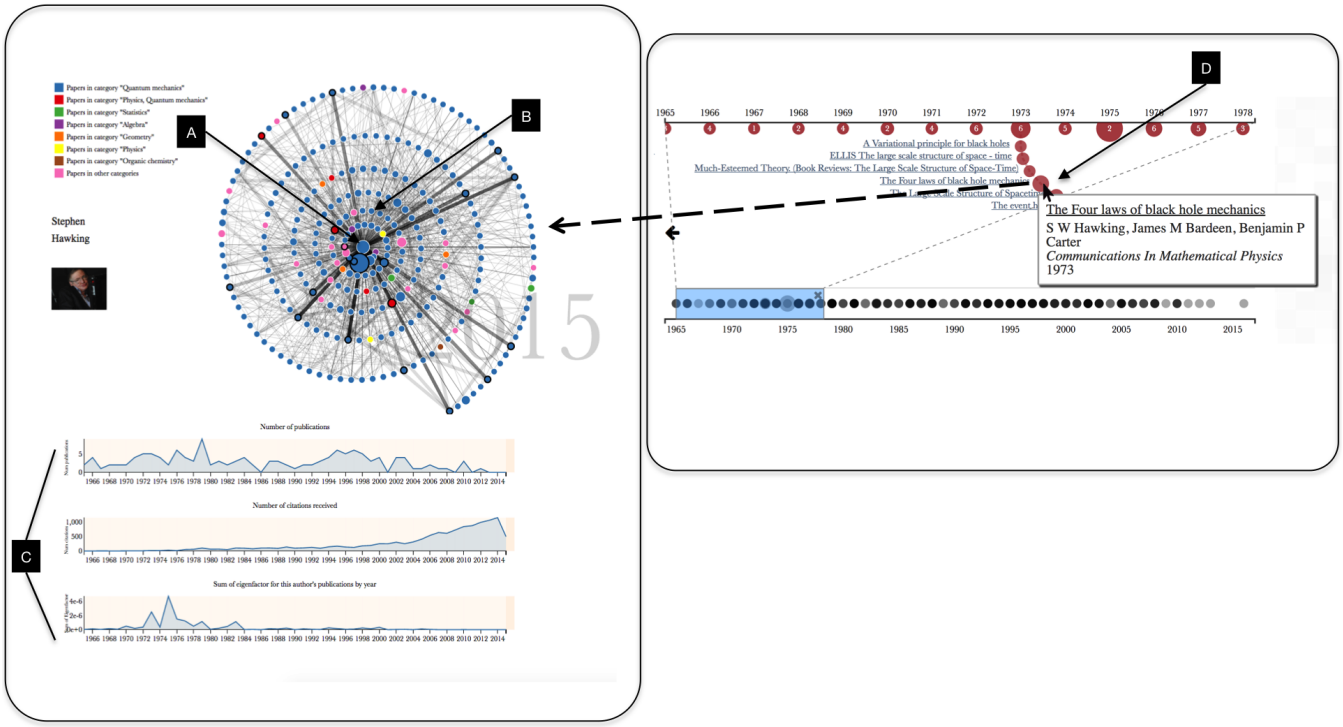


Figure 2: Linked visualizations. To the left is the citation influence visualization. (A) The center node represents all of Stephen Hawking's publications. (B) Nodes that appear in the spiral around the center are important papers that have cited Dr. Hawking's work, colored by category (field of study) and with size proportional to influence (Eigenfactor) score. (C) Below the graph display are three timelines of key indicators by year: number of publications, number of citations, and sum of Eigenfactor. On the right is the timeline visualization of Dr. Hawking's papers. Selecting one paper in the timeline (D) highlights the papers on the left that have cited that particular paper.

The visualization begins by displaying the central node representing all of the author's work. Over time, nodes appear around the center representing important papers (those papers with higher Eigenfactor) that have cited the author's work. These nodes send out links to the center as well as to other nodes that appear in the visualization. This is an egocentric network, with alter (non-ego) nodes placed radially around the center in order of time. In this way, spatial placement encodes time; in addition, color encodes category (field of study) and size encodes influence (Eigenfactor score). In order to reduce the visual complexity of the graph, the number of nodes in the visualization is restricted to 275, giving preference for the alter nodes to highly ranked papers that have category information available. See [7] for more detail on this visualization.

4.3 Linking the visualizations

The citation visualization shows the influence of a scholar's entire body of work (condensing all of the papers into a single central node), while the timeline visualization shows the individual papers authored by the scholar. By opening the visualizations in separate browser windows (with a dual monitor display, or in side-by-side windows), the timeline can be used to drill down deeper into the citation visualization. Pointing the mouse at a paper on the timeline causes the papers that cited this selected paper to be highlighted in the citation visualization (Figure 2).

5. CONCLUSIONS AND FUTURE DIRECTIONS

We presented a pair of linked visualizations generated from the open Microsoft Academic Graph data to explore a scholar's influence through citations to her work. These visualizations could be included in author profiles to offer an interactive tool to explore any author's publications and citations.

We plan to further develop these tools, in particular improving paper selection and management and allowing the use of different category groupings, so that papers can be colored by journal, citation-based community, etc. We would also like to collect and incorporate user feedback to evaluate how these profiles can be used and improved. Finally, we plan to make these profiles easier for authors to share and present among colleagues and evaluators.

6. ACKNOWLEDGMENTS

We thank Microsoft Research for allowing open access to their academic graph. We also thank JSTOR, the Pew Charitable Trust, and the Chemical Heritage Foundation for their support in this work, and three anonymous reviewers for their helpful feedback.

7. REFERENCES

[1] M. Bostock, V. Ogievetsky, and J. Heer. D3: Data-Driven Documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011.

[2] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572, Nov. 2005.

[3] C. D. Kelly and M. D. Jennions. The h index and career assessment by numbers. *Trends in Ecology & Evolution*, 21(4):167–170, Apr. 2006.

[4] J. H. Larkin and H. A. Simon. Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science*, 11(1):65–100, Jan. 1987.

[5] L. Leydesdorff. Caveats for the use of citation indicators in research and journal evaluations. *Journal of the American Society for Information Science and Technology*, 59(2):278–287, Jan. 2008.

[6] J. L. Ortega and I. F. Aguillo. Microsoft academic search and Google scholar citations: Comparative analysis of author profiles. *Journal of the Association for Information Science and Technology*, 65(6):1149–1156, June 2014.

[7] J. Portenoy, J. Hullman, and J. D. West. Leveraging Citation Networks to Visualize Scholarly Influence Over Time. *arXiv:1611.07135 [cs]*, Nov. 2016. arXiv: 1611.07135.

[8] E. Segel and J. Heer. Narrative Visualization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1139–1148, Nov. 2010.

[9] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, and K. Wang. An Overview of Microsoft Academic Service (MAS) and Applications. pages 243–246. ACM Press, 2015.

[10] J. Ward, W. Bejarano, and A. Dudás. Scholarly social media profiles and libraries: A review. *LIBER Quarterly*, 24(4), May 2015.

[11] I. Wesley-Smith, C. T. Bergstrom, and J. D. West. Static ranking of scholarly papers using article-level eigenfactor (ALEF). *arXiv preprint arXiv:1606.08534*, 2016.

[12] J. West, I. Wesley-Smith, and C. Bergstrom. A recommendation system based on hierarchical clustering of an article-level citation network. *IEEE Transactions on Big Data*, (in press), 2016.