Visualizing Connections Between Scientific Publications

Dylan Ingham
Northeastern University

Parker Brandt
Northeastern University

Jaeson Pyeon
Northeastern University

Ryan Costa
Northeastern University

Ethan Waple
Northeastern University

ABSTRACT

The following describes a tool for visualizing information about scientific articles from a Google Scholar query. This visualization will allow users to explore how papers related to each other through citation-based network diagrams or feature-based scatterplots. The visualization will also have interactive features for filtering papers and displaying additional information.

Network and embedding representations of scientific literature offer an alternative, and potentially more intuitive, research method for students. Compared to the search results of Google Scholar, this visualization offers more robust dimensions for comparing results. Also, the filters facilitate paper selection by narrowing down research results based on the user's spatial analysis of the literature network.

Index Terms: Citation engine, visualization, network

1 INTRODUCTION

This visualization leverages Google Scholar's vast database of scientific papers. Upon input, the visualization returns the top 200 papers of a custom Google Scholar query (Note: live user interaction will not be available in this project due to delayed speed of Scholarly API, so graphics will instead be derived from a prefilled dataset). These papers can then be explored in a network space, where links are derived on citations, or in an embedding space, where dimensions are derived from various characteristic columns. Both visualizations can be filtered based on user input, and selecting a datapoint will provide additional information about the paper.

Searching on Google Scholar yields a basic series of results presented in a simple linear fashion. Alternatively, this visualization presents those same results in multi-dimensional space. This allows the user to explore papers based on their broader context in the network of literature, offering new ways to compare results. Filtering provides additional methods for directed the user's exploration. Upon selection of data, the users can read the paper's abstract through annotations, which lets them make more informed decisions.

The ideal user of our visualization is a student who is researching a subject that is new to them. Such a student may query Google Scholar to find foundational literature in a subject, but become confused by the pages of identical results. This visualization provides contextual visuals through which the student may discover foundational literature faster and more intuitively, enabling faster and more effective studying.

2 RELATED WORK

Although the focus of previously completed works are distinctly different from those pursued in this paper, these pieces do provide an insight into prior usage of network graphing and academic paper citation relationships. In "Toward the Discovery of Citation Cartels in Citation Networks", Fister et al. [1] discuss the problem of "citation cartels" in the academic publishing community. "Citation cartels" are defined as "groups of authors that cite each other disproportionately more than they do other groups of authors that work on the same subject." The authors provide an enlightening visualization to exemplify this concept using a multilayer network visualization (Figure 1). In the figure, relationships between "citation cartels", "author citation networks" and "paper citation networks" are exemplified by interlayer edges between nodes. The secondary visualization also provides an example of "orphan" nodes. These are nodes that lack any relationship data linking them to other nodes. As such, they are presented as separate from the rest of the interconnected network web. In the secondary visualization, nodes (authors) are connected through edges (citations) that include numerical addendums (number of cites given by the author). This usage of network mapping to present data regarding citation relationships is a key to the use case at hand.

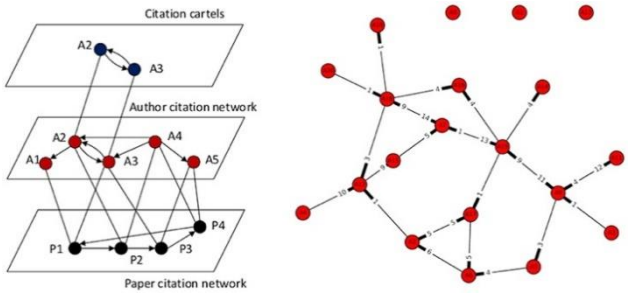
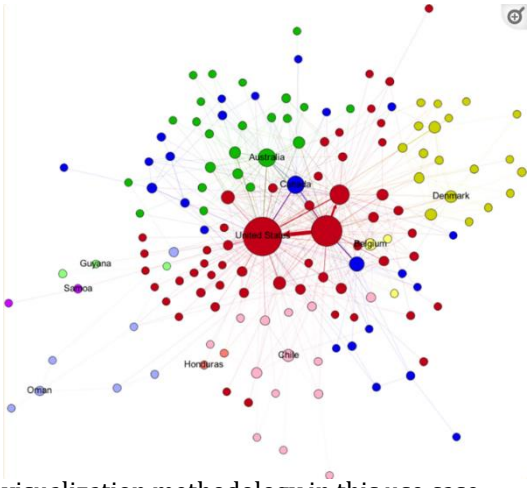


Figure 1: Schematic presentation of citation networks using a multilayer network diagram (Fister et al., 2016)

This technique is further exemplified in "Using Visualizations to Explore Network Dynamics". In this piece, Chu et al. [2] discuss the usage of various network analysis techniques in an exploratory network study of the online tobacco-control community GLOBALink. This exploratory network study yields a great visual aid (Figure 2) that includes several notable aspects. In the visualization, nodes are used to represent member countries of GLOBALink. Each directed edge symbolizes one member country vouching for another. The node

size is used to exemplify the "out-degree" – i.e. the total number of target country referrals made from the source country. Node color is used to show the modularity class of the member country – a classification system used to separate the member countries into 10 distinct communities. For example, the modularity class shown in Light Blue contains countries such as Oman, Saudi Arabia, Yemen, and Egypt. The characteristic used in this classification is the geographic region of the Middle East. In comparison, the modularity class shown in Dark Blue contains countries such as Canada, France, Congo, and Algeria – countries distinguished by the characteristic of "French influence". To further aid in user understanding of the visualization, the highest "out-degree" node in each cluster is also labeled. Edges also carry data in this figure with edge size representing the number of referrals between countries and edge colors deriving from the color of the source node. While the paper and this specific figure have value in their own right, the inclusion of many datapoints in one comprehensive network graph is notable for its ability to be leveraged



as a visualization methodology in this use case.

Figure 2: GLOBALink referral network used in network analysis of membership referral data (Chu et al., 2013)

3 Use Case

The following description will demonstrate a use case for this tool in a research context. Suppose an undergraduate student wishes to create a list of research papers to read based on a specific keyword search, such as "epigenetic profiling in single cells". The student could simply browse Google Scholar, but those results appear in a simple linear format and don't provide any context about how the papers relate to each other. Instead, this tool will query the top 200 Google Scholar results for the specified search term and display them in a network format, where links are created based on papers that cite each other. Using this visualization, the student can locate papers that are central to this subset of literature based on prevalence of links.

In order to dig deeper, the student can hover over papers and see information about their name, abstract, number of citations, Journal, and Authors. If the user wishes to compare the papers based on different contexts, they can view them as points in a scatterplot where the axes are chosen from among several metrics, such as Number of Citations, Number of Authors, Citations per Year, Year of Release, and Publication Journal. The user can learn which papers are outliers based on certain metrics of interest.

3.1 Domain Tasks

One domain task for the user to complete will be selecting papers that are contextually important to their field, based on citation interactions, so they can learn more about the paper and potentially read it in full.

Another domain tasks for the user to complete will be exploring the way that papers relate based on various axes of comparison, so that they can select papers to read with specific outlier or trend-related characteristics (such as publication in a rare scientific Journal or a high rate of citations per year).

3.2 Task Abstraction

This visualization supports the high-level domain task of annotation, because nodes in the network can be hovered over to learn more details about the scientific paper of interest, including its name and abstract, which can help the user decide if they want to read the paper.

This visualization also supports the mid-level domain task of explore, because the papers can be viewed as points in a scatterplot among various axes. This allows the user to casually browse through different arrangements of the data and, based on coordinate locations, uncover specific outlying papers or common trends in the search results.

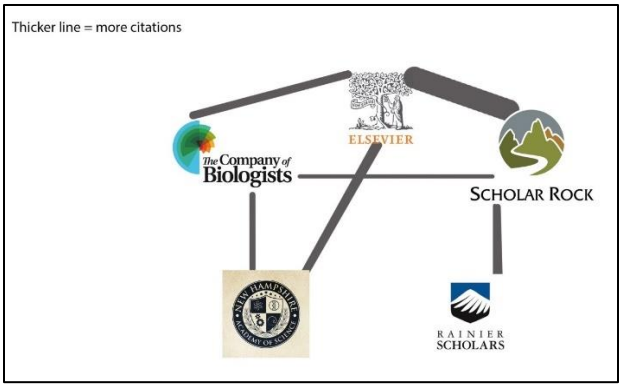
4 Data

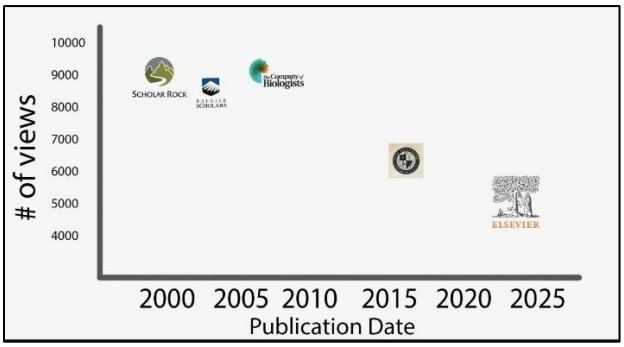
Throughout the data collection process, there were issues in extracting data off the data source, Google Scholar. This was due to restrictions from Google Scholar against web scraping, which was bypassed through implementing a Python script using scraperapi. As a result of the data being scraped, there are some ethical considerations to consider in terms of potentially filtering out largely cited papers, quality control, and respecting the terms of service. The issue of quality control was addressed throughout the data cleaning process. The issue of potentially missing data points can be handled to an extent by continually updating data throughout the course of the project to make sure that visualizations are backed by the most up to date data. In terms of potential bias, the results of the dataset are heavily influenced by the Google algorithm as these selected results are the ones that have been scraped and then compiled into a dataset. There is also selection bias during the web scraping process if the search terms used to scrape are biased or incomplete in any way. This would also lead to potentially missing important research papers that would alter the results of the project.

Based on the initial web scraping results, we were able to derive attributes for author name, number, names of the sources that cited the paper, title, publication year, venue, number of citations, pages,

cite_per_year, and abstracts. During the data cleaning process, column entries for page numbers had to be normalized to convert page ranges to integers as there were entries with dashed lines or letters in front of it.

5 Design Process





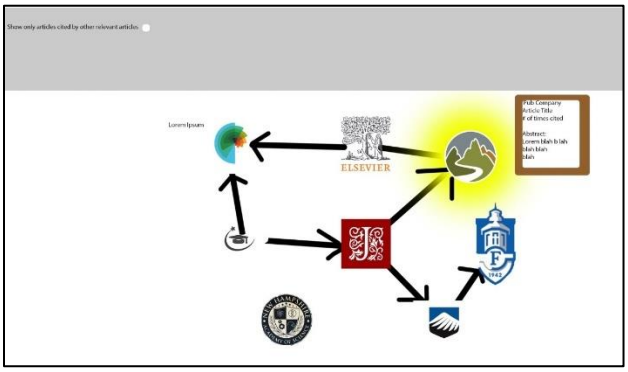


Figure 3: Rough and partial sketches. (Top) This initial draft inspired the idea to utilize a network-style visualization. As well as having each node be distinguished by a distinct publishing company. (Middle) A scatterplot visualization was conceptualized to show correlation. This draft inspired the alternative view that will be implemented in our final tool. (Bottom) Continuing with the network visualization, this draft inspired the use of directed arrows to indicate which article cited which. In addition, this draft also inspired the idea to add a hover feature which would display more information about a selected node.

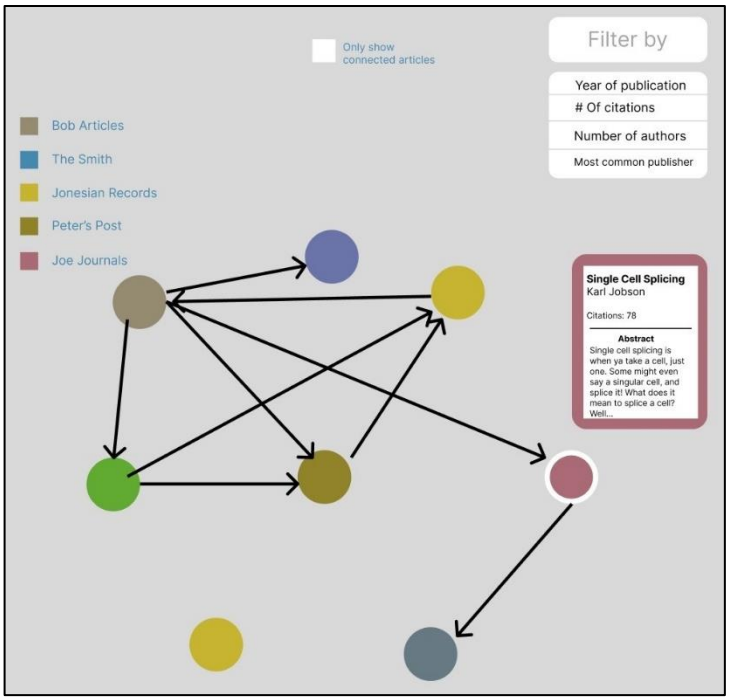


Figure 4: Final, polished sketch. Interactive components of the visualization include filtering the articles(nodes) based on certain criteria, which will shift their position in the network accordingly. The user is also able to remove nodes that lack connections to only view cited articles. Upon hovering over a node, a tooltip will popup which will include the title of the article, its author, how many times it has been cited (globally) and its abstraction. We utilize color to categorize each article as well as its position to determine relevance to specific filter query. The lines and nodes themselves serve as the marks for this visualization.

In some of our initial designs we thought to utilize the article's publishing company's logo as a mark on the visualizations. However, this introduced some scalability issues as well as some confusion when multiple articles from the same company appeared on the visualization. We then sought to utilize different colors for each article, but this made the visualization look hectic when lots of points were added. We eventually settled on using a singular color whose color intensity changed based on the number of citations it received. This design decision worked best with our visualization as we were primarily focused on frequency of citations. We initially wanted to include "orphan" nodes. However, they created a lot of excess white space and weren't really solving the problem of finding

relevant articles, so we decided not to include them in our final design.

For our usability testing, we asked users to identify specific points on the scatterplot as well as tell us which data points were the most or least relevant on the network graph. Thanks to our detailed filtering tools and legend, most users were able to find the correct point within 7 seconds, and many reported that our visualization was easy to use. So long as we stick to the general principles of our current visualization, even as the data scales up and we add more filtering options, we anticipate users will still have a positive and intuitive experience with our visualization.

6 *Final Design*

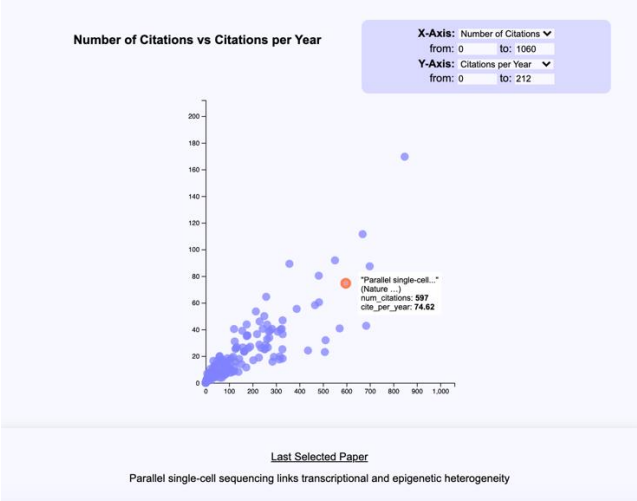


Figure 5: Scatterplot. This scatterplot helps visualize data regarding the content of queried papers. The user is able to explore the queried data by browsing various arrangements and identifying outlying papers or trends in the papers. Each point on the plot represents a paper from the query. Hovering over the point reveals a tooltip that provides information like the paper's title and where it was published. The user can select the variable represented on the X and Y axis through the dropdown menus. This updates both the layout of the graph as well as the tooltip in real time.

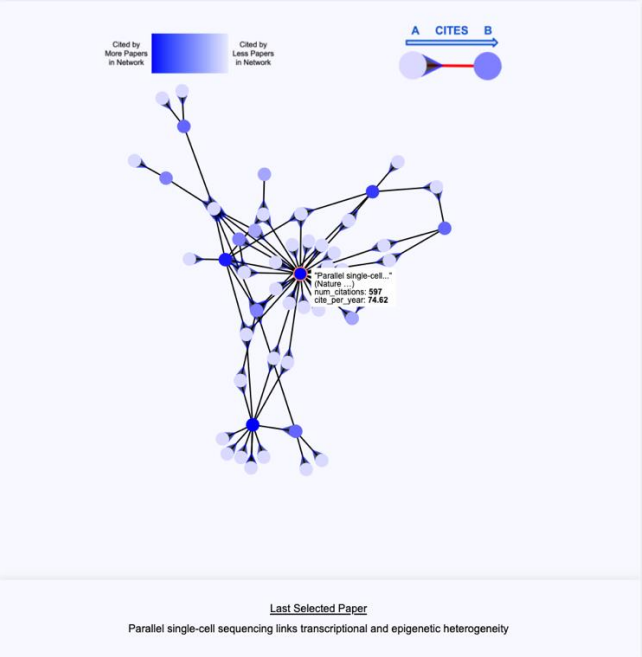


Figure 6: Network Graph. This network graph helps visualize data regarding the context of queried papers. The visualization incorporates the high-level domain of "annotate" to help users contextualize a specific paper within its citational network and decide whether that paper is the right one for their research purposes. Each node on the graph represents a queried paper with spanning edges representing citational connections between papers. Tooltips akin to those in the scatterplot are also present. The shape of the edge and hue of the node yield information on the order of the citational connection and the relative number of citations respectively.

7 *Discussion*

All in all, our visualization does address the domain problem of difficulties in finding quality papers while doing research. By implementing a network graph visualization showing the context of papers relative to its citational network and a scatter plot of paper content related attributes, a user can gather enough information to pinpoint individual papers and then research them further. Future improvements to enhance the user experience could include a zoom in and out feature to make viewing nodes in the network easier as the dataset of network of papers grows larger. Moreover, having a feature for users to produce a list of their most relevant sources with the source list automatically cited would be another necessary improvement. Given the scope of the project, the major limitation of our current work is that this visualization provides a great proof of

concept but not a fully reusable framework that can take in any number of sources and organize them. However, with the improvements listed above and further development, we believe that the insights leveraged in the network visualization can be a powerful way for students to intuitively discover foundational literature.

8 Conclusion

In this paper, we have presented an interactive visualization device that enables users to navigate scientific articles on Google Scholar by exploring contextual relationships through a citation-based network diagram or content-based associations via a mutable scatterplot tool. We identified a specific use case for this device to simplify the process of literature analysis for academic researchers and students. To achieve this goal, we have established domain tasks and task abstractions that present a traditionally large and linear dataset across two distinct and separately informative two-dimensional spaces, providing functionality for accessing more descriptive, paper-specific information when desired. We hope that the resulting visualization improves both accessibility and user experience by enabling efficient exploration of the scientific literature in a unique spatial analysis, benefiting academic research within the field of science and in disciplines beyond.

References

[1] I. Fister, I. Fister, and M. Perc. Toward the Discovery of Citation Cartels in Citation Networks. *Frontiers in Physics, 4*, 2016. doi:10.3389/fphy.2016.00049

[2] K.H. Chu, H. Wipfli, and T. W. Valente. Using Visualizations to Explore Network Dynamics. J Soc Struct. 14:4, 2013. PMID: 25285051; PMCID: PMC4184104.

Appendix

Task Abstraction

Selecting papers for deeper review based on context with other papers.	High-level – Produce – Annotate
Looking for outliers or trends in the way that papers relate based on various axes of comparison	Mid-level – Search – Explore

Data Abstraction

Attribute Names	Attribute Type	What does it represent?
Cited Names	Text	Title name of articles the paper was cited in
Author name	Text	Name of author(s) of publication
Author number	Discrete Numerical	Number of author(s) involved
Title	Text	Name of paper
Publication Year	Discrete Numerical	Year of publication
Venue	Text	Name of publisher or group involved
Citation Number	Continuous Numerical	Total number of recorded citations
Pages	Discrete Numerical	Page range denoted with dashes in between start and end
Cite Per Year	Continuous Numerical	Average citations per year