Title

Student1, Major; Student2, Major; Student3, Major; Mentor, Department

Research Statement

To develop computational methods for analyzing, visualizing, and cleaning the dataset of all 17 million Library of Congress publishing records, specifically using methods from graph theory and social network analysis.

Project Goals and Objectives

The primary goal is to contribute to a growing body of digital humanities scholarship by providing an analysis of the publishing industry over time, through an initial holistic, topohraphical survey of the data followed by specific case studies. Our particular objectives are as follows: (1) Use Python programming techniques to quantify numerical properties of the full dataset based on the relationships between publishers, authors, places, and books. (2) Determine how to remove duplicate records by merging matching titles using edit distance measures as well as modern machine learning-based predictive models. (3) Augment dataset where possible, filling in unknown values using various online resources. (4) Visualize data using the Javascript D3 library to provide graphics that inform viewers of the dataset's key facets.

Background & Related Work

Broadly, book history is the study of patterns of authorship, genres, and practices of reading, buying, or other use of books in the past. Within book history, the field of historical bibliometrics is the quantitative study of these phenomena, for example measuring production or consumption. Usually historical bibliometric studies work at the scale of a single genre or market slice in a certain period, and occasionally on a large scale, as in the large Swiss publisher of the Société Typographique de Neuchâtel.[2] Our mentor has acquired the *complete dataset* of records of the Library of Congress (LOC). While there is certainly much dirty and missing data, this data set of 10 million records notionally contains complete metadata for *all cataloged books,* or all books in the history of printing, from the late 15th century to the present. Our investigations this year have so far discovered no extensive analysis of this data set, much less with the methods we propose to apply to its study (for an exception see Al et al. (2012)). The only work we have found in our review has been some Machine learning categorization of the data.

Our specific approach employs social network analysis (SNA), or the use of mathematical graph and network theory and principles of sociology to detect patterns that stem from social interaction (see e.g. Otte & Rousseau (2002)). Some examples include studies on the spread of disease, friendship networks, and business networks. These all have the commonality that there is some attribute, or vertex, that is connected to other vertices by some relationship, or edge. Viewing social interactions in this context helps to detect the impact that individuals or groups have on the whole. SNA researchers often work with large datasets, for example in Facebook or Twitter. In terms of methods, SNA researchers often write their own analytic methods, but the field is huge. (See the examples of Skinner (2016) and Van Eck (2014).) We justify our approach by observing that our chosen tool (the SNAP Python library) is well regarded, widely used, supported, stable, and can scale up to the size we need.

Project Impact, significance, or purpose

Network analysis is sometimes used in humanities research, but usually at small scale, and often only to illustrate arguments made through traditional textual interpretation. This project will apparently be the first to apply the rigorous quantitative methods from SNA to the extremely large data set of all LOC records, notionally containing all information for all books ever published. SNAP highlights datasets that have used their library, and the vast majority of them fall within the fields of either biology or social media. The publishing industry is typically analyzed manually and the current literature base focuses on one particular country or era at a time, so *this project would be unique in its holistic review of the entire industry over a long period of time.* This project would contribute to the fields of literature and history by identifying interesting phenomena about the publishing industry. It would also provide meaningful contributions in Computer Science in the fields of text processing, network analysis and categorical data visualization. Finally, the clean data set we will produce will be of enormous use to other researchers in history; library and information science (LIS); and data science.

Project Design

This project would be a continuation of our current roles on the project; S1 and S2 have been working on this project for 2 semesters, and S3 has been working for 1.5 semesters. The data was obtained in Fall 2017 by an undergraduate team, also supervised by [Mentor], doing a Capstone senior Computer Science project. We three spent most of Fall 2018 tackling how best to make queries and metrics on this large dataset, which is computationally expensive. We also experimented with

---

1 [about title]

2 Respective examples of these are Darnton (1982) and Burrows (2015).

different programming libraries. We finally decided on the Stanford Network Analysis Project's (SNAP) free code module, which is a Python library developed for modeling networks with millions of records. By early Spring 2019, we received resources from the university's High Performance Research Computing Institute, which allowed us to feasibly create a working model of the 17 million records. Given that we have access to the data and resources to conduct rigorous analyses, we intend to concurrently pursue three areas of study: quantitative analysis (S1), text deduplication (S2), and data visualization (S3). For both semesters, methods will be developed and tested in the first half, and the last half will be spent preparing writing and presentation of this information.

In the fall, S2 (who graduates in December) will be working to reduce the number of duplicate place names, authors, publishers, and editions using traditional edit distance measures such as the Levenshtein distance, experimental predictive models based on modern machine learning techniques, and statistical methods where appropriate. He will also be augmenting the dataset, filling in unknown properties using online resources such as geographical databases and the WorldCat Library Catalog.

S1 and S3 will be working for the full 2019-2020 academic year. S1 will spend the Fall semester writing Python methods that work alongside the preexisting SNAP functionality to deliver quantifications of properties of the data. She intends to determine where significant communities and clusters of data exist by conducting comparisons over time between different publishing hubs around the world. In the spring, her work will be focused on two specific case studies based on the information gleaned from the previous semester's work in identifying influential cases.

S3 will be working on making the large data set digestible by creating visualizations of key components of the data set. In the fall, S3 will be developing visuals like standard network visualizations, moving charts, and beeswarm frequency histograms from meaningful slices of data based on location, time period, and specific publishing communities. They will use the Javascript D3 library to make these visualizations interactive and flexible for different chunks of data (for example, the publishing trends in Germany in the first half of the 20th century). In the spring, their work will be focused on visualizing the two specific case studies identified by S1.

Expected Project Timeline

Fall 2019 - all students; *Spring 2020* - S1 & S3

Tasks	**Aug**	**Sept**	**Oct**	**Nov**		**Jan**	**Feb**	**Mar**	**Apr**
Develop methods in respective areas (statistical analysis, data visualization, text deduplication [in Fall only]) to understand dataset's properties	■	■	■			■	■	■	
Conduct testing to ensure methods are reliable and accurate	■	■	■			■	■	■	
Produce article on semester's work, prepare presentation			■	■				■	■
Create and update a Wordpress website to host data visualizations and notable findings			■	■					■

Anticipated Results

By December 2019, we expect to have a dataset with a majority of the duplicates removed, usable Python methods for holistic network analysis, data visualizations to clarify viewers' understanding of publishing, and an article written (co-authored with [mentor]) about our process and findings, perhaps similar to Michel et al. (2011). By May 2020, we expect to apply the methods created in the fall to specific places or eras of publishing to develop visualizations and quantifications of interesting phenomena in our two case studies, and to have co-authored again with [mentor] a second article. The latter will be a case study in his area of research expertise, 17th-18th century Germany. We hope to identify significant trends about the publishing industry and find notable case studies to pursue in the future, which would be a likely avenue for Honors theses for S1 and S3 . We want to make our findings available through presentations at Discover USC and to classes and student groups within the computing community whenever possible. We will pursue publication in journals and magazines like *Digital Humanities Quarterly, Digital Scholarship in the Humanities, Caravel,* and *Carolina CrossTalk*.

Personal Statements

S1: I am excited to build an important bridge between the separate, but connected worlds of the social sciences, literature, and computing. I volunteered at various local public libraries for about seven years when I was younger, and I love working on a project where I can see that our future findings will be an important contribution to a field that shaped my childhood. I am majoring in computer science and statistics to attain my ultimate career goal of becoming an applied data scientist, and receiving a grant for this project would aid me in developing a solid foundation for a research-oriented future.

S2: While working on this project, I have developed a much greater appreciation for the field of text processing and machine learning. Before being hired by CDH to work on this project, my interests were mainly in cybersecurity. After learning about the importance of having a clean data set and the challenges involved, I became really interested in the field of data science. Receiving a Magellan Grant will help me to continue working on this project, enabling me to further develop skills relevant to my future career and potentially graduate school.

S3: As a computer science major with a philosophy minor and a practicing artist in Columbia, SC, I joined this project to explore and discover the many ways in which these fields interact and support each other. This semester I am taking a CSCE 567, Data Visualization Tools, where I am learning to master using the Javascript D3 library to aid my work on this project. Receiving this grant will get me the support I need to keep studying and developing interactive data visualizations and how they can be applied to the digital humanities.

Short Bibliography

Al, Umut, Zehra Taşkın, and Güleda Düzyol. "Use of social network analysis in bibliometric researches." e-Motion (2012): 40.

Burrows, Simon. "Bibliometrics, Popular Reading, and the Literary Field of an Enlightenment Publisher." *Annuaire d'Etudes Francaises* (2015): 15-43.

Darnton, Robert. *The Literary Underground of the Old Regime*. Cambridge: Harvard University Press, 1982.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (January 14, 2011): 176-82.

Otte, Evelien, and Ronald Rousseau. "Social Network Analysis: A Powerful Strategy, Also for the Information Sciences." *Journal of Information Science* 28, no. 6 (December 2002): 441–53. doi:10.1177/016555150202800601.

"Research Guides: Bibliometrics and Altmetrics: Measuring the Impact of Knowledge: Bibliometrics." University of Maryland Research Guides. Accessed February 10, 2019. https://lib.guides.umd.edu/bibliometrics/bibliometrics.

Skinner, Jason. "Bibliometrics and Social Network Analysis of Doctoral Research: Research Trends In Distance Learning." (2016). https://digitalrepository.unm.edu/oils_etds/32

Van Eck, N.J., & Waltman, L. (2014). Visualizing bibliometric networks. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), Measuring scholarly impact: Methods and practice (pp. 285–320). Springer.