Emory Dissertation Network

Mert Özbay
Emory University
Computer Science
Atlanta, GA, 30322, USA
mert.ozbay@emory.edu

1 Introduction

Most scholarship doesn't happen in isolation. Collaboration has been the rule rather than the exception, and increasingly, it is taking an interdisciplinary form. This research aims to analyze how collaboration is reflected in text using quantitative and qualitative methods. This kind of work has not been conducted before on the Emory Thesis and Dissertations (ETD) library. Using the metadata obtained from the library as well as the method of topic modeling, I create networks that can represent the collaborative and interdisciplinary process of writing theses and dissertation at Emory University. This research uses abstracts, rather than the theses and dissertations themselves, as textual source, and therefore it also seeks to demonstrate how abstracts, which are more readily accessible than the dissertations, can be used in place of them.

2 Related Work

I based my inquiry on previous scholarship that uses topic modeling and networks as means of textual analysis. Previous works that used these techniques in tandem or combined these with qualitative analysis have provided a framework for my methodology.

The first work I make use of is Matthew Wilkens's "Genre, Computation, and the Varieties of Twentieth-Century U.S. Fiction" from *Journal of Cultural Analytics* (Wilkens, 2016). Wilken's article seeks to tackle the nebulous concept of genre through computational methods, focusing on a corpus of 8580 American books published from 1880 to 1990. Wilkens uses this corpus to generate an LDA model with 200 topics and then reduces that to 20 using dimensionality reduction. To those he adds stylistics features, setting, and some extra-textual features, representing the books as 27-dimensional vectors. He then reduces these to 10-dimensional vectors using PCA, and runs two different clustering algorithms, k-means and DBSCAN. He analyzes the resulting clusters, explaining how input features are reflected in the clusters by using his knowledge of American literature. I adopt this approach of representing documents as topic distribution vectors as a means of deducing similarity between them. However, I don't rely on clustering, as the dissertations and theses in my corpus are already separated into groups by department, which is far more concrete than genre. Instead, I use networks as a method of further inquiry and comparison.

Another article is Lucas van der Deijl et al.'s "The Canon of Dutch Literature According to Google," again from *Journal of Cultural Analytics* (van der Deijl et al., 2019). This article uses two different sources, internet content from Wikipedia plus Google and an academic anthology on Dutch literature to derive two different "canons" of Dutch literature. To evaluate the internet content, the authors create a network based on the connections provided by Google Knowledge Graphs displayed as search results. To interpret the network, the authors use methods such as modularity to detect communities of related authors and PageRank to measure centrality, which they then interpret and compare it to the more traditional "Canon" derived from the anthology. I also aim to use networks, specifically centrality, for analysis, but I will use betweenness centrality instead of PageRank centrality, as measures such as PageRank and eigenvector score node centrality by taking the centrality of adjacent nodes into account. In my case, this should not influence the centrality score. For example, if department A collaborates with department B, which happens to be very collaborative, this does not make department A more collaborative automatically.

"Analysis of Social Dynamics on FDA Panels using Social Networks Extracted from Meeting Transcripts" by David Broniatowski and Christopher

Magee shows us how to combine topic modeling and networks to analyze interactions (Broniatowski and Magee, 2010). The authors use the Author-Topic model, a variant of LDA, to analyze a collection of FDA expert committee meeting transcripts. Each utterance is treated as a document while each committee member is treated as an author. They later use topic distributions from the AT model to create a matrix of topic similarity between the authors, which they then use to create a network. The authors compared the results of the network to the voting results of the meetings and found committee members more closely linked based on the network created from the topic model are more likely to vote similarly. In my work, I follow this methodology to create one of my networks. However, I treat departments as author units, rather than individual authors, as virtually all thesis and dissertation authors have a single work in the ETD repository.

3 Corpus

This dataset is a compilation of abstracts from the Emory Theses and Dissertations database containing the PhD dissertations as well as Masters and Honors theses of people who received their degrees from Emory University. Each instance is the abstract plus the metadata associated with it (but not the thesis author's name). There are 9981 instances. The dataset does not contain the instances of the abstracts whose owners embargoed their paper (plus their abstract) and did not accept to share it for the purposes of this research. Each instance contains several features: the title of the dissertation , school (College of Arts and Sciences, Laney Graduate School, Public Health, etc.), department, degree(BS, MA, etc.), language (English, French, Spanish — the abstract may still be in English), committee chair, committee members, graduation year, and the abstract itself, among other things. The 9981 abstracts from these instances serve as my main (textual) data. This dataset has been provided to me by the Scholarly Repository Coordinator of the Emory Library. I can't share this dataset since the abstracts are subject to copyright.

The dataset needs preprocessing to run topic modeling and to create networks. Firstly, the abstracts contain HTML tags that need to be removed. I use the BeautifulSoup library, used usually for parsing scraped website data, to clean up the HTML tags (Richardson, 2020).

The other preprocessing needed is to standardize the names of the thesis committee chairs and other members. These are irregular, sometimes including information about the member's Emory network ID, institution affiliation (mostly Emory University), and titles (Dr., etc). Sometimes, the last name of the committee member precedes the first name, separated by a comma. I use regular expressions to remove all titles, institution names, IDs, and put all names into first name-last name format. Still, this data is not always consistent, with some instances lacking a committee chair while others have multiple committee chairs. Therefore, I don't distinguish between the chair and other members when making analyses.

Every student graduates from a specific department, which means each dissertation is associated with a specific department. Once again, there is a variety in department names. I use my own judgment to group similar department names together while also making sure no detail is lost. Different public health programs are all combined under "Public Health." When one department appears to be a subset of the other, I combine them. For example, theses and dissertations from the Russian department are added to the Russian and East European Studies department. In the case of a joint degree (e.g., Mathematics and Economics), I leave them as it is, as it is not possible to pick either of the departments without losing valuable information. After this process, the dataset had 83 distinct departments.

After the grouping, we can observe that theses and dissertations are unevenly distributed with regards to department affiliation. Here are the 15 departments with the highest count of dissertations and theses:

Department	**# of Theses/ Dissertations**
Epidemiology	1412
Global Health	1050
Biological and Biomedical Sciences	961
Chemistry	455
Psychology	440
Public Health	353
Environmental Health	349
Biostatistics and Bioinformatics	335
Behavioral Sciences and Health Education	314

Biology	292
Religion	275
English and Creative Writing	269
Neuroscience and Behavioral Biology	238
History	216
Clinical Research	198

Epidemiology, Global Health, and Biological and Biomedical Sciences are overly represented, while some other departments at the bottom of the list have less than 10 dissertations or theses.

4 Process and Methods

4.1 Topic Modeling — LDA

After preprocessing, the first step I took was to derive a topic model using Latent Dirichlet allocation (LDA), using the MALLET toolkit (McCallum, 2002). I preferred MALLET, as it produced more intelligible topics. In addition to stopwords, I chose to remove numbers altogether, as replacing them with the placeholder 'NUM' affected my results. For the number of topics, I chose 25. Having a large number of topics carries the risk of each discipline or department being associated with a specific topic or, at the very least, the risk of having limited topic distribution overlap between fields. However, with a very low number of topics, we might not get meaningfully distinct topics that represent the entire corpus. For 83 departments, I found 25 to be an ideal number.

The topic model produces a representation of each abstract as a topic distribution vector. By taking the average of all dissertations/theses that belong to a particular department, I am also able to get department topic distribution vectors. I use these department topic distributions in my analysis.

4.2 Network — Department Connections

There are multiple ways to represent this dataset as a network. Different abstracts, committee members, and departments can all be represented as nodes, resulting in three different networks. In each case, the edges between the connections are derived thanks to the fact that committee members serve in multiple dissertation/thesis committees. Since each dissertation/thesis is associated with a department, they end up serving in committees associated with different departments as well. I use the networkx package to create, visualize, and analyze these networks (Hagberg et al., 2008).

The approach that first comes to mind is to represent each abstract as a node. In this approach, there is a connection between two nodes if a professor has served in the committees for both of the thesis/dissertation committees of the nodes represented. The shortcoming of this approach is that committee membership has inconsistent documentation, especially for earlier instances in the dataset, meaning some instances show up to 4 or 5 committee members while some others show only 1, which will result in a completely disconnected node. This network has 9980 nodes.

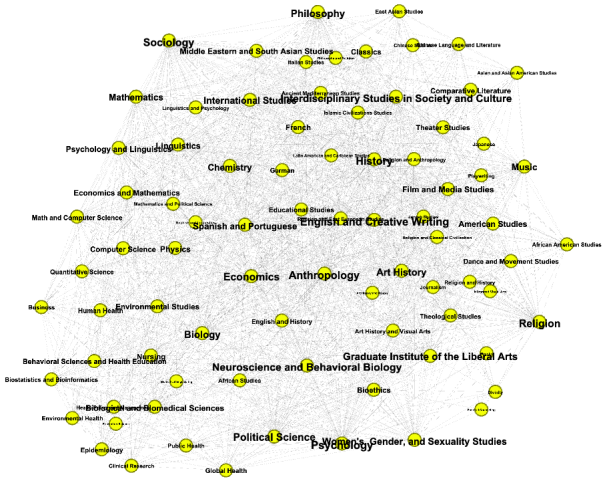
The second approach is to represent each committee member as a node, which can be thought of as the inverse of the previous network. This time, an edge is drawn between two nodes if two people served in the same committee. This approach has its shortcomings too, as people who served in more committees will inevitably end up with more connections and seem more collaborative, resulting in a more connected node and higher centrality. In reality, however, a person might have appeared in fewer committees but in more diverse disciplines and might have collaborated with more people in proportion to the number of times they served in a committee. Furthermore, this network can be harder to interpret as data regarding the department affiliations of committee members isn't available to me. This network has 5554 nodes.

The approach I chose to ultimately follow was to represent each department as a node. In this case, an edge connects two departments if a person served in two thesis/dissertation committees affiliated with two departments. For example, if Professor P (whose department affiliation we don't know) served in an Honors Thesis committee for a Sociology student and in a PhD dissertation committee for a Public Health doctoral candidate, an edge connects the nodes representing the Sociology and Public Health departments. In this kind of a network, a better connected node represents a more interdisciplinary department. As discussed before, the data is highly skewed with regards to the number of abstracts affiliated with each department. However, compared to the case of the committee member network, this does not influence the centrality measure as significantly. A department might have many disserations/theses associated with it but still might have collaborated with a limited number of other departments. This network

has 83 nodes, which is computationally much more manageable than the previous ones and easier to interpret visually.

With this network, we can use centrality measures to determine how well-connected — and therefore collaborative and interdisciplinary — a department is. As discussed before, measures such as eigenvector or PageRank centrality are not appropriate, as a high score (and therefore high interdisciplinary) of an adjacent department node does not automatically mean a department node itself should get a higher centrality score. I use betweenness centrality, which is scored according to the number of shortest paths passing through a node. A well-connected node will have more shortest paths passing from it and a higher betweenness centrality. With the topic distribution vectors and the centrality score for each department, I ran a multiple linear regression with each topic as an independent variable and the centrality as the dependent variable to see if any particular topic is related to a higher or lower centrality score. For the linear regression, I excluded departments with less than 10 theses/dissertations associated with it.

Figure 1: Department Network



4.3 Author-Topic Model Network

Making use of Broniatowski and Magee's paper, I created another network, this time using a topic model instead of the committee membership metadata to derive networks. Following the paper, I used the Author-Topic model, which is "a generative model for documents that extends Latent Dirichlet Allocation to include authorship information" (Rosen-Zvi et al., 2004). I represent each department as an author, and this model automatically provides a representation of authors as topic distribution networks. I made use of the Gensim library to train my model (Rehurek and Sojka, 2011). I again use 25 topics.

The topic distributions of the authors can be compared to derive connections. Broniatowski and Magee derive the probability that two authors are speaking about the same topic via the formula:

$$P(X_1 \cap X_2) = \sum_{i}^{T} P(Z = z_i \mid X_1) * P(Z = z_i \mid X_2)$$

This means, in order to get the probability that the authors are "speaking" about the same topic, we sum up, for all topics, the probability that both authors are talking about that specific topic (Broniatowski and Magee, 2010). In my case, this is a measure of shared topics between departments, implying possible collaboration or similarity in discipline. Using the values derived from this formula, we can construct the edges of a graph: If the shared topic value between a pair of authors is greater than a certain cutoff point corresponding to random chance, an edge is drawn between nodes. Broniatowski and Magee use 1 (one) divided by the number of topics as this cutoff value, which I also adopt. Broniatowski and Magee do this process iteratively while creating the model to compare author-pair joint probability to the cutoff value in each iteration (Broniatowski and Magee, 2010). I couldn't do this because of the limitations of the library I used. Broniatowski and Magee end up using unweighted edges anyway based on the distribution of edge weight values. Furthermore, they use multiple samples to derive multiple models in order to produce a more statistically sound network, which I couldn't do due to computational limitations. Based on the edges I have, I produce a network.

After creating the network, I again use betweenness centrality to calculate the centrality score for each department. The logic behind is that departments that interact more will have more similar topic distributions, and more collaborative and interdisciplinary departments will be better-connected due to the topic distribution similarity to a larger number of departments. Using this centrality score derived from the Author-Topic model as the dependent variable and the centrality score derived from committee membership connections between department as the independent variable, I run a simple linear regression. The assumption behind this is that a committee member that served in

two different committees will result in similar theses/dissertations with regards to topic distribution. For the linear regression, I excluded departments with less than 10 theses/dissertations associated with it.

5 Results and Discussion

5.1 Department Connection Network

Looking at the betweenness centrality scores for the department network created by committee membership metadata, we can observe the most collaborative departments:

Department	Centrality Score
Religion	0.034226
Psychology	0.034015
English and Creative Writing	0.028924
Biology	0.028030
History	0.027485
Neuroscience & Behavioral Biology	0.027169
Anthropology	0.025055
Sociology	0.024127
Women's, Gender, and Sexuality Studies	0.021204
Interdisciplinary Studies in Society and Culture	0.020782

It is interesting to see that while there are many departments with a relatively high number of dissertations at the top of this list (e.g., Psychology, Biology), the list isn't completely dominated by them, and the most over-represented departments don't appear at the top of the list. We can look at the topic distribution of specific departments to interpret them and learn more about the reasons for these centrality scores.

The most prominent topic for the Religion department is Topic 22, which contains words such as 'religious', 'moral', 'god', 'human', and 'church'. This is very clearly a topic that related to religious life and religion. Other departments that have a high share of this topic are Religion and Classical Civilization, Divinity, Philosophy and Religion, and Theological Studies. Another prominent topic, Topic 7, characterised by words 'social', 'american', 'identity', 'racial', 'political' show us the cultural side of the study of religion. Departments with the largest share of this topic are African American Studies, Women's, Gender, and Sexuality Studies, and Philosophy and Religion. Here, we see the unique position of the Religion department in Emory University. It is connected to the departments in the School of Theology, some of which focus on religion itself, with issues of ethics and theology in the forefront, while also being connected to other departments that end up studying the cultural and historical context around religion. This provides the Religion department with a unique central position in the network. It should also be noted that in the case of Religion, we see a lot of joint/double majors (e.g., Philosophy and Religion), which underlines its interdisciplinary nature and definitely led to a higher number of connections in the network.

Looking at the Psychology department reveals that the most important topic is Topic 19, which includes words 'mental', 'depression', 'students', 'university', and 'sleep'. Psychology is the department with the highest percentage of this topic. Other notable departments include Journalism (now defunct), Nursing, and Educational Studies (now defunct). This topic shows us an intersection of issues of mental health and college life. Very close in share to Topic 19 is Topic 14, with words 'memory', 'cognitive', 'learning', language', 'social', 'children', and 'monkeys'. Departments with the highest share of this topic are Linguistics and Psychology, Linguistics, and Spanish and Linguistics. This topic reflects the more cognitive and developmental side of the discipline of Psychology. Furthermore, we again see joint degrees, reflecting the interdisciplinary nature of this department, similar to the Religion department. The fourth most important topic (admittedly not as significant) for the Psychology department is Topic 15 with words brain', 'neurons', 'mice', 'motor', 'dopamine', and prominent departments Neuroscience and Behavioral Biology, Biological and Biomedical Sciences, Biology. This reflects the part of Psychology that relates to Neuroscience and other branches of Biology. Thus, we observe again that a department is central because it bridges two distinct clusters (Social Sciences and Biology in this case) in the academic realm.

The multiple linear regression yields no significant results. No specific topic seems to be strongly correlated to a high or low centrality. The most statistically significant result is for Topic 12, with a p-value of 0.071, which is admittedly greater than the common cutoff value of 0.05. It is slightly correlated with a negative centrality score. Public

Health and Global Health, which are large departments with relatively low centrality scores, have a high share of this topic.

From these results, a high centrality score appears to be a result of not a specific share or combination of topics but a highly particular position of a discipline that bridges the academic field.

5.2 Author-Topic Model Network

Comparing the two centrality scores, one from the network based on the committee membership metadata and the other from the network based on the Author-Topic model topic distributions, gives me a way to quantitatively assess the correlation between topic distribution and centrality. However, the result of the simple linear regression with these two variables showed only a weak positive correlation between the two centrality scores. The p-value was 0.092, and the coefficient for the independent variable (committee membership-based department network centrality score) was 0.1013. Furthermore, rerunning the topic model and trying different number of topics did not result in better or even equally significant results. However, our previous qualitative analysis showed us departments with the highest centrality scores actually were located at interdisciplinary crossroads when it came to topic distribution. Therefore, we should discuss the shortcomings of the methodology.

Firstly, the Gensim library produces worse results than MALLET, and this is evident in the less coherent topics I obtained from the Author-Topic model even though this model takes authorship into account and therefore supposably should produce more coherent topics. A better library could produce more reliable results. Furthermore, using many instances of the model as different samples, as done by Broniatowski and Magee, will lead to more statistically reliable results.

One other limitation is the limited size of the Emory academic community. While Emory University is not a small university by most definitions, it is still significantly smaller than large academic institutions, both with regards to the number of departments and the number of theses/dissertations produced. Having more than 83 departments (without the overlaps I removed) in analysis would lead to more accurate results.

6 Conclusion and Next Steps

This analysis used computational and qualitative methods to show collaboration and interdisciplinarity in the Emory University academic community. It built on the existing scholarship that use network analysis by providing a new approach to creating networks given the metadata. It supplemented Broniatowski and Magee's approach to combining topic modeling and network analysis by adding a new way to validate and quantify the accuracy of the results. My analysis showed that most interdisciplinary departments link otherwise unrelated (or rather, unconnected) communities of academic disciplines.

Future work should use other academic communities and more refined methods to assess the results of this study. Furthermore, there is still a large part of the metadata that this research did not use. The time feature can be used to conduct a diachronic study. The names of the advisers can be used for a gendered analysis of the textual data. The titles, another textual source in the dataset, can be incorporated in future research. All of these can be used to inspect questions this dataset is waiting to answer.

References

David A. Broniatowski and Christopher L. Magee. 2010. Analysis of social dynamics on fda panels using social networks extracted from meeting transcripts. In *2010 IEEE Second International Conference on Social Computing*, pages 329–334.

Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. Http://mallet.cs.umass.edu.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Leonard Richardson. 2020. Beautiful Soup.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, page 487–494, Arlington, Virginia, USA. AUAI Press.

Lucas van der Deijl, Antal van den Bosch, and Roel Smeets. 2019. The canon of dutch literature according to google. *Journal of Cultural Analytics*, 4(2).

Matthew Wilkens. 2016. Genre, computation, and the varieties of twentieth-century u.s. fiction. *Journal of Cultural Analytics*, 2(2).