201A Project; Final Writeup

Lauren Oey, Isabella Destefano, Erik Brockbank, Hayden Schill, Jamal Williams

December 13, 2018

1 Introduction

With new and exciting topics coming into existence almost every day, scientific fields are constantly fluctuating. Topics that were once popular no longer garner the same attention while topics that were recently unheard of suddenly skyrocket to the forefront of the field. What motivates these shifts in topic popularity? Do young scientists with little to lose push the field in new and interesting ways? Do veteran scientists with years of experience predict these shifts and position themselves to be ahead of the curve? This project looks to answer these questions and to ultimately understand how individual authors drive shifts in topic popularity, thereby influencing the behavior of the broader scientific community. After generating a corpus of papers from the annual Cognitive Science Society Conference Proceedings (CogSci), we intend to apply co-authorship networks and topic models to examine whether authors more central to the community initiate shifts in the field.

To assess the influence of an individual author on the behavior of the scientific community we construct an undirected, unweighted graph using the authors of papers in our CogSci corpus. Here, each individual author is a node and the co-publications between authors define the edges between nodes. This network intends to capture the connections between authors and can be used to assess the importance of an individual. Importance is a subjective concept, and there are many ways to measure how connected an individual is; in graph theory these are known as measures of centrality. We look at three simple measures of centrality: degree centrality, closeness centrality, and betweenness centrality. These measures were chosen based on previous research pertaining to the behavior of social networks (Landher, Friedl, & Heidemann, 2010). To assess changes in the field, we use a static topic model based on the principles of Latent Dirichlet

Allocation (LDA) to better understand and explore the topic space that has existed over the history of CogSci. The introduction of LDA, a generative statistical model for natural language processing, allowed for the efficient discovery and classification of unique topics from large datasets (Blei, Ng, & Jordan, 2003). Subsequent research into static topic models has primarily focused on improving LDA based models. Griffiths & Steyvers (2004) extended the standard LDA topic model by incorporating author information allowing them to model papers as sampling from the topic distributions of multiple authors instead of sampling from a simple Dirichlet prior on topic distributions over documents (for a review, see Rosen-Zvi, Griffiths, Steyvers & Smyth, 2004). More recently, changes in the distribution of topics over time have been examined through the use of dynamic topic models (Rothe, Rich, & Li, 2018). The implementation of a dynamic topic model, which tracks the emergence, disappearance, and prevalence of topics over time, is particularly interesting for this project and we intend to explore this avenue in 201B.

2 Data Scraping

Our full data set extracted from the CogSci corpus consists of the title, authors, abstracts, full text, and year of conference papers in the annual CogSci proceedings ranging from 2000 to 2018. To extract this data from the web, we use a number of Python libraries, namely the web scraping library Beautiful Soup (bs4), the PDF to text converter PDFMiner (pdfminer), the URL handling module urllib, and the regular expressions handling library re. The files from which the data were extracted are mainly hosted by two websites, eScholarship and MindModeling. In the first round of data extraction, we primarily use MindModeling and the CogSci conference website to collect titles, authors, and abstracts from 2009 to 2018. In the second round, we use full text PDF files to collect the title, authors, full text, and year for proceedings from 2000 to 2018. We map these extracted full-text data files onto the subset of data with abstracts.

After extraction of the raw text with python, we process the data using R. This processing involves removing irrelevant text (cover pages, text appearing before the abstract, punctuation, escape characters, non-ASCII characters), converting all text to lower case letters, and tokenizing each word (i.e., splitting sentences into individual words). The natural language processing package cleanNLP returns lemmas (e.g., run is the lemma of running, runs, & ran) and part-of-speech tags which aid in the filtering of function words

(e.g., determiners, prepositions, etc.). Finally, to deal with multiple name entries for the same author (e.g. Joshua Tenenbaum, Joshua B. Tenenbaum, Josh Tenenbaum, etc.) a new vector is created that contains their last name and first initial (e.g., J Tenenbaum, E Vul, etc.).

Collecting and processing the full raw text data has many advantages. For example, we will be able to further extract (or filter out) subsections of the text (e.g. references, figure captions, keywords). Once cleaned, this data is used in the implementation of two distinct models:

1. a co-authorship network to assess the relative connectedness of an individual to other authors in the network, and

2. a topic model that identifies unique topics and allows for flexible exploration of the structure of and changes in those topics.

CogSci has grown over the last few years, the general positive trend is demonstrated in Figure 1. There are several leading authors in the field, the publications of these leaders increase over the years as demonstrated in Figure 2. A visual comparison between sex of the leaders seems that the top most productive authors in CogSci are predominantly male.

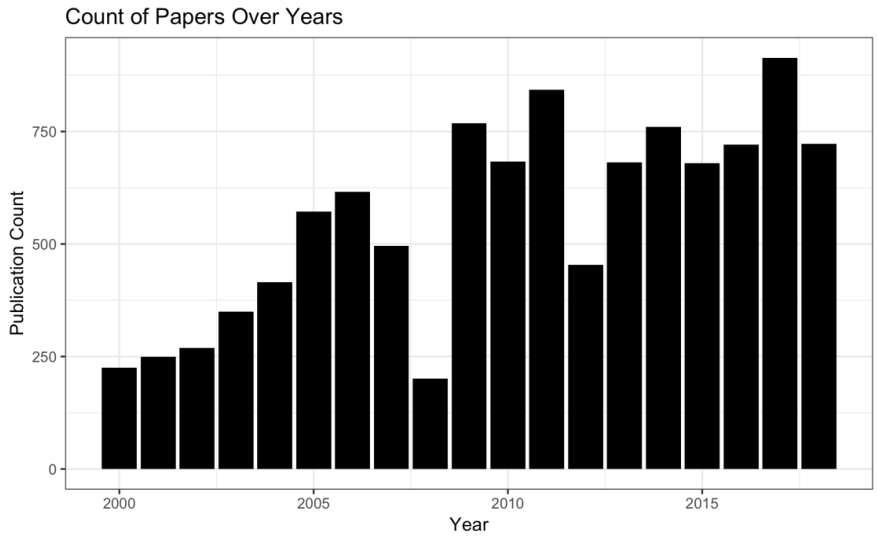

Figure 1: Figure 1 Number of papers publish per year

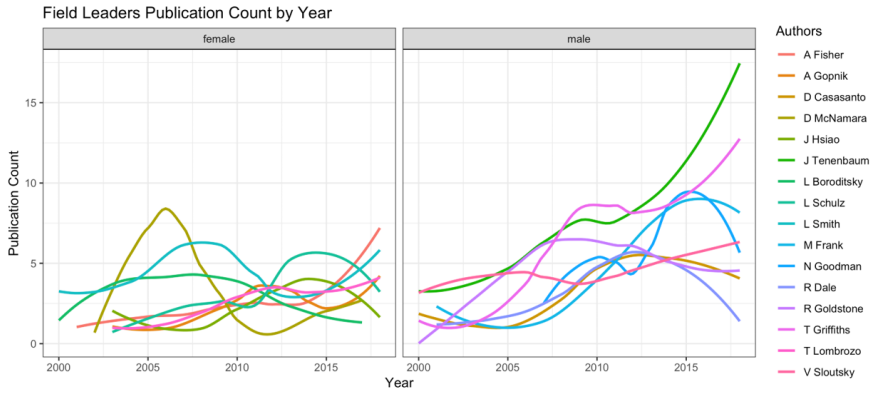

Figure 2: Figure 2 Publication trends of top 8 most published authors by sex.

3 Author Network

Concepts from graph theory that are often used to analyze other kinds of social networks facilitate analysis of the co-authorship network. To make the co-authorship graph tractable, we assume that pairs of authors with multiple co-publications have only one edge connecting them, making the graph "simple" which is a requirement for calculating many properties of graph. Calculations of centrality were performed using the R-package igraph.

A graph with n nodes can be represented as an adjacency matrix, $A = (a_{ij}) \in \{0;1\}^{n \times n}$, where the value of $a_{ij} = 1$ just in case there is an edge connecting nodes i and j. Properties of the graph can be derived from the adjacency matrix.

Degree centrality (DC) σ_D is the simplest centrality measure that enumerates the nodes that a given node is directly connected to. Formally, using the adjacency matrix we define DC as:

$$\sigma_D(x) = \sum_{i=1}^{n} a_{ix}$$

In the co-authorship network this measure can be interpreted as the number of authors on which a given author has direct influence; more influential authors will have a higher DC.

Closeness centrality (CC) $\sigma_C(x)$ is defined in terms of the distance between a node and the other

nodes in the graph. This distance is quantified as the minimum number of edges to travel from one node to another. The distances between a given node and all other nodes are summed, and inverse of this total distance gives CC. Formally, we define CC as:

$$\sigma_C(x) = \frac{1}{\sum_{i=1}^{n} d(i,x)}$$

In the co-authorship network CC can be interpreted as how productively an author can spread information through the network of authors; authors who can proliferate information will have a higher CC.

Betweenness centrality (BC) $\sigma_B(x)$ is defined by how many of the shortest paths between pairs of nodes a given node is located. Formally we have:

$$\sigma_B(x) = \sum_{i=1, i\neq x}^{n} \sum_{j<i, j\neq x}^{n} \frac{g_{ij}(x)}{g_{ij}}$$

where g_{ij} represent the number of shortests paths from i to j and $g_{ij}(x)$ represents the number of these paths that contain node x.

In the co-authorship network, we assume that the interaction between two authors is facilitated by the authors in between them. We can interpret BC as the influence that an author has on the communications between other authors in the network.

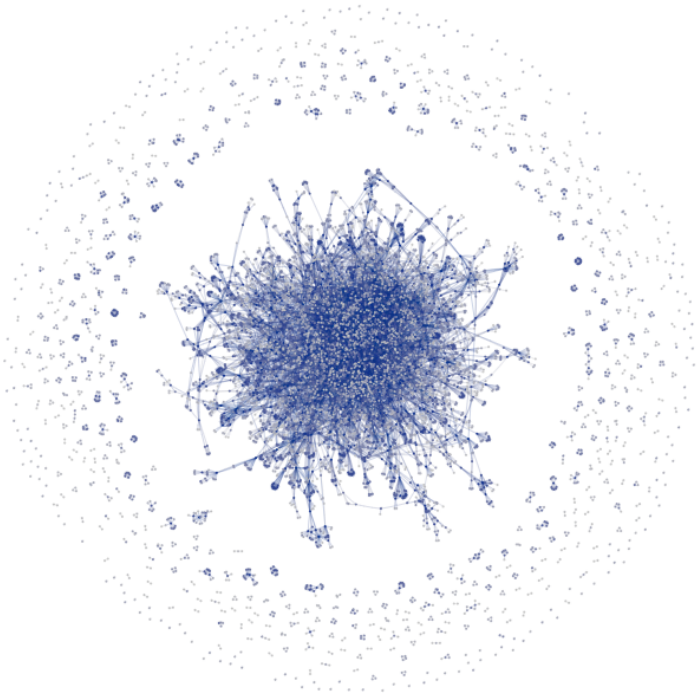

Figure 3: Figure 3 Full Co-authorship network

We used the R package visNetwork to generate a visualization of the co-authorship network shown in Figure 3. This illustration reveals a problem with our network, which is the fact of multiple subgraphs surrounding the main graph, making the network disconnected.

Using our measures of centrality we found the top 10 authors for each of the meausres (Table 1). This revealed that the centrality measures were relatively successful; they all agree on the top two authors, but differentiate between authors ranked lower than that.

rank	betweenness	closeness	degree
1	J Tenenbaum	J Tenenbaum	J Tenenbaum
2	T Griffiths	T Griffiths	T Griffiths
3	R Goldstone	R Goldstone	N Goodman
4	D Gentner	T Gureckis	D Gentner
5	R Dale	N Goodman	M Frank
6	C Yu	C Yu	R Goldstone
7	N Chater	M Jones	R Dale
8	N Goodman	L Smith	C Yu
9	B Love	B Love	T Gureckis
10	L Smith	D Gentner	J Hu

4 Structural Topic Model

We focus on exploring available topic modeling tools and assessing the feasibility of implementing these topic models based on the abstracts before moving on to full text documents. We use the Structural Topic Model (stm) package in R which can be easily modified to implement basic, or advanced, topic models. More advanced models can be implemented with stm by including covariates such as author(s) or date as part of the generative process for topic and word distributions (Roberts, Steward, & Tingley, 2015). Models estimated with stm include topical prevalence covariates on the distribution of topics in a document and topical content covariates on the distribution of words in a topic. Therefore, this package is a powerful tool for exploring topic dynamics in CogSci papers.

In our current analysis, we use the basic topic modeling features of stm to explore the feasibility of estimating topic models with our data set. Using the text from 7,844 abstracts, reduced to 2,859 unique lemmas, we are able to successfully fit a topic model with 5, 10, 20, and 50 topics. Qualitative analysis of the most probable words in each topic confirms that the topic divisions are meaningful and provide useful proof of concept for the topic modeling approach (Figure 4). Next, we collect 9,165 full text papers from 2000 to 2018 which, compared to abstract only data, significantly increase the number of unique lemmas to 19,947.

This new model converges on qualitatively interesting topics based on the most probable words in each topic. In 201B, we will look at more quantitative ways of evaluating this model for varied numbers of topics by using prevalence and content covariates.

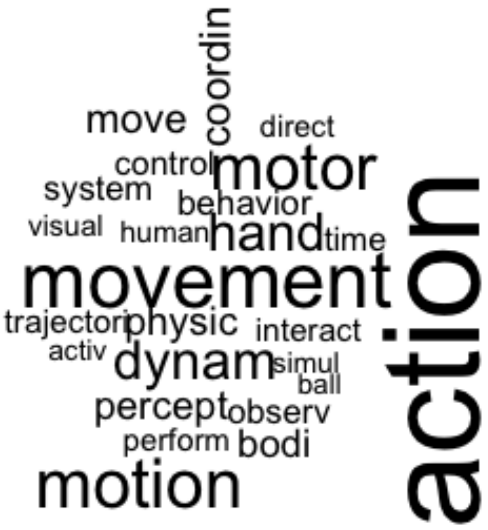

Figure 4: Figure 4 Wordcloud for topic 2.

5 Discussion

To identify how individual authors influence the behavior of the scientific community we apply a topic model and a co-authorship network to two large datasets of empirical papers and abstracts. At this intermediate stage, we generate a co-authorship network which applies multiple measures of centrality to determine the prominence of authors in the broader CogSci community. Furthermore, we have produced multiple topic models that generate quantitatively robust topic groups (i.e. developmental, vision) based off of recurring patterns of words and themes in the text.

Future efforts intend to increase our dataset to include the entire history of CogSci and to continue working through the logistics of wrangling the data and creating comprehensive models. Data processing issues include: optimized filtering of irrelevant words, dealing with missing data from erratic and idiosyncratic website formatting, updating the author vector to take into account individuals with the same last name and first initial, and resolving inconsistencies across text within PDF files (e.g. some articles have centrally labelled sections while others do not). A problem about the co-authorship model is that the graph has unique connections between authors, which fails to account for how the number of publications between two authors reflects the strength of their relationship. One potential solution is to implement a weighted graph. However, the reliability of centrality measures in weighted graphs is a current topic of debate and, while this might solve the problem of connectedness, it might incorporate new indomitable problems to our network. Another issue with the graph is that it is disconnected, there are dozens of sub-graphs with no edges between any of their constituent nodes. This poses a problem for the calculation of betweenness and closeness centrality. A possible solution is to identify the sub-graphs and perform centrality analyses on only the largest. While the exact form of our author topic model and co-authorship network will be dictated to some degree by computational resources and limitations on analytic techniques for defining an algorithm (i.e. deriving closed form equations to use when approximating the posterior distribution), there is still much room to improve on existing topic models.

On a broad level, we intend to unify both models such that we are able make key inferences about the data and answer some of the questions posed earlier. Specifically, we intend to explore any correlations that arise from the outputs in the two models. For instance, are there similar patterns that arise that might

allow us to make inferences about which authors spark trends in the field? Such inferences might then be able to be extrapolated to predict up-and-coming authors, as well as up-and-coming trends. This information is of critical importance to journal publishers and funding agencies.

References

[1] Blei, D., Jordan, M., & Ng, A. Y. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

[2] Griffiths, T. L., Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences, 101(suppl 1)*, 5228-5235.

[3] Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04* (p. 306).

[4] Rothe, A., Rich, A. S., & Li, Z.-W. (2018). Topics and Trends in Cognitive Science (2000-2017). *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 979–984. Retrieved from https://github.com/blei-

[5] The igraph core team (2015). Igraph-the network analysis package. Retrieved from https://igraph.org/redirect.html

[6] Roberts, Margaret E., Stewart, Brandon M., Tingley, Dustin (2015). Stm: R Package for Structural Topic Models. Retrieved from cran.r-project.org/web/packages/stm/vignettes/stmVignette.pdf