Computational Analysis of Social Science Research Heather A. Haveman 7 September, 2018

Background Description for Undergraduate Research Apprentices

Research in many academic fields requires reviewing the literature to document what we know about a phenomenon and what gaps exist in our knowledge. I am developing a new computational method for reviewing academic literature that takes advantage of online collections containing nearly all articles published in academic journals (e.g., JSTOR, MathSciNet, Web of Science, MEDLINE). This method stands in sharp contrast to the time-honored practice of reading a small number of texts, usually far less than 1,000, which is prone to biases stemming from researchers' training and social networks. This computational method is dictionary-based, in contrast to citation-based, cosine-similarity, and topic modeling methods. After generating the descriptive results I explain in this document, we will validate this method by comparing it to other methods.[1]

The focus for this project is the field of organizations, which spans multiple disciplines (e.g., sociology, political science, management) and which is largely published in articles, rather than books. This literature has been dominated for the last four decades by three paradigms: demographic, relational, and cultural.[2] These paradigms (aka perspectives) have different conceptions of social structure[3], and thus different conceptions of what creates opportunities for and constraints on the actions of individuals, groups, and organizations. The *demographic perspective* holds that social structure is constituted by distributions of social actors along salient dimensions of social and physical space (individual age, race, and gender; workgroup size and composition; organizational strategy, size, and location). The *relational perspective* holds that social structure is constituted by webs of social connections among people, groups,

[1] As I explain below, we have data on the full texts of all published articles in the two subjects that are at the core of organizations research, so we will be able to conduct more complex text analyses for these subjects. And we have data on the citations used by each article in the dataset we will analyze, which means we will be able to conduct a citation network analysis as well.

[2] A paradigm is a framework within which scientists think, an intellectual perspective shared by a group of scientists that guides their formulation of theories and their empirical research.

[3] Social structure means patterns in social life that (a) cannot be reduced to the sum of its parts (e.g., to the actions of individual people) and (b) guide people's actions (e.g., the requirements for declaring a major at UC Berkeley).

and organizations (e.g., friendship networks, workgroup interactions, buyer-supplier ties between organizations). The *cultural perspective* holds that social structure is constituted by widely shared norms, values, expectations, roles, and rituals, which generate understandings of what is possible and reasonable. These perspectives also have different conceptions of identity ("who am I? what do I stand for? what do I want?"), and therefore motivations for action. For demographers, identity derives from position, absolute or relative, along salient dimensions of social life. For relational scholars, identity derives from ties among individuals, groups, and organizations. For cultural scholars, identity derives from social interaction. All three perspectives have been applied to explain both *micro-level behavior* (of individuals like professors and students, small groups like campus clubs and fraternities/sororities, or organizational subunits like academic departments), and *macro-level behavior* (of entire organizations like UC Berkeley, organizational populations like American public universities, and fields like American higher education).[4]

I begin by reading foundational texts in each paradigm (mostly journal articles) to generate a <u>dictionary of concepts</u> (consisting of single words, bigrams [2-word pairs], and trigrams [3-word triplets]) for each paradigm, based on my deep knowledge of this field. Each concept dictionary contains the language that situates scholarly work within its paradigm: the words and phrases in a paradigm's dictionary express the ideas in that paradigm. Each concept dictionary I have created so far contains between 100 and 200 concepts; these are excel spreadsheets, with full references for the source texts on one worksheet and the concept list on a second worksheet. To the extent that a published article uses the concepts in a paradigm's dictionary, the article is engaged intellectually with that paradigm. For example, if an article talks mostly about employees' social ties to coworkers, relationships outside of work, and connections to bosses, it is engaged with the relational paradigm. But if an article talks mostly about employees' age, educational background, gender, race/ethnicity, or functional

[4] For more details on theses three paradigms, see the working paper (Haveman and Wetts 2018) that I am sending you along with this background note.

background, it is engaged with the demographic paradigm. And if an article talks mostly about employees' understandings about acceptable workplace behavior, it is engaged with the cultural paradigm.

For each paradigm, we will write code in Python (preferably in Jupyter notebooks) to measure how much articles published in academic journals use the focal paradigm's concepts, scaled by article length. I label this measure *engagement*. My current (work-in-progress) definition of engagement with paradigm *p* by articles published in subject *s* at time *t* is as follows:

$$engagement_{pst} = \sum_{a=1}^{n} \frac{engagement_{psta}}{n} = \frac{\sum_{i=1}^{m} paradigm\ concept_{ipsta}}{\sum_{j=1}^{q} article\ concept_{jsta}}$$

where *engagement*$_{psta}$ is engagement by article *a* (published at time *t* in subject *s*) with paradigm *p* and time *t* spans a single calendar year. It has two components: the numerator, $\sum_{i=1}^{m} paradigm\ concept_{ipsta}$, is the total number of times the concepts in paradigm *p* are used in article *a* (i.e., we will count each instance of each dictionary concept in article *a* separately), and the denominator, $\sum_{j=1}^{q} article\ concept_{jsta}$, is the total number of concepts in article *a* (i.e., we will count each instance of each article concept separately, by summing up the frequencies for all the concepts in an article). Note that *engagement*$_{pst}$ is the arithmetic average across articles in a subject and time period (year).[5] Because a "concept" can be a single word, a bigram, or a trigram, we will first calculate numerator counts for words, bigrams, and trigrams separately for each article. Then we will add up these three counts to create $\sum_{i=1}^{m} paradigm\ concept_{ipsta}$.

Engagement is a continuous variable with a theoretical range of [0,1], where 0 indicates that the articles in the focal subject area in the focal year use none of the paradigm's concepts and 1 indicates that all articles use all of the paradigm's concepts. *Engagement*$_{pst}$ will equal 1

[5] Depending on the shape of the typical distribution for *engagement*$_{psta}$, we may also calculate the median, standard deviation, and/or skewness.

only when every word in every article in that subject published at that time is part of the paradigm's concept list.[6] Since no article will consist entirely and exclusively of a paradigm's concepts, the observed range for $engagement_{psta}$ is [0, 1), meaning that 1 is not included in the range.[7] Indeed, the observed maximum is likely to be a very small fraction. Therefore, the observed maximum for average $engagement_{pst}$ is also likely to be a very small fraction. To show 3 significant digits (3 digits after the decimal point), we will likely have to multiply *engagement* by 1,000 or more. We will determine the scaling factor after we get preliminary results.

The data we will analyze come from the JSTOR Data-for-Research corpus from 1970 to 2015. The data include all subjects in the social sciences and selected subjects in business and economics (specifically, business, economics, finance, labor and employment relations, management and organizational behavior, marketing and advertising), plus public health, health policy, and history (N≅3.4 million). (You can see the subject list for JSTOR here). For each article, JSTOR has provided counts of Ngrams (1-3 words) and metadata (e.g., publication date, author name, journal title). JSTOR has also provided machine-readable full texts for all articles published 1970-2015 (N≅400,000) in two subject areas that are at the core of organizational research: (1) sociology and (2) management and organizational behavior. We will analyze these full texts after we analyze the Ngram and metadata files. Methods for that analysis will be determined after we conduct the descriptive analysis of the Ngram and metadata files, so I do not discuss them here.

We will have to map the metadata file for each article onto the 3 Ngram data files. We will also have to map the title of each journal onto the JSTOR subject(s) under which it is classified, using the JSTOR subject list. Some journals are listed under 2 or more subjects; for example, *Administrative Science Quarterly*, one of the top journals in the field of organizations,

[6] Note to myself: Do I need to alter the formula so that $engagement_{psta} = 1$ iff (i) article *a* uses all of the concepts in paradigm *p* and (ii) all of the concepts in article *a* are in the concept dictionary for paradigm *p*? The current definition requires only (ii). Talk with David Bamman in the I School about this?

[7] Even the texts I used to create the dictionary for each paradigm will not yield an engagement score of 1 because each dictionary captures only the text concepts most relevant to that paradigm and ignores text concepts, such as organization, employee, job, and performance, that are common across paradigms.

is listed under both sociology and management and organizational behavior. My initial inclination is to allow subject areas to overlap (so 1 journal may be assigned to 2 or more subjects), but I may change my mind about this later and force each journal into a single (primary) subject area, based on either my knowledge of the field or some computerized text analysis of the articles published in that journal.

Measuring *engagement* will allow us to trace how the impact of each paradigm has ebbed and flowed, as well as to discern where (in terms of subject areas) its impact has been most intense. We will also be able to calculate correlations between the concepts associated with each pair of paradigms, to capture the extent to which they are deployed together.

Because it is based on comprehensive (computer-assisted) reading of the literature, this method is likely to reveal unexpected findings – things I and other organizations scholars wouldn't have predicted based on our own readings of the literature. The ability to reveal unexpected findings could make this method useful to scholars in multiple fields, especially those that span multiple disciplines.

I will (soon) upload the data to the Savio supercomputer cluster that the Lawrence Berkeley National Lab has shared with UC Berkeley. As I mentioned above, this descriptive analysis is to be done using Python, preferably in Jupyter notebooks. I will create a Google drive folder for us to share work. Eventually, the code will be deposited in GitHub, to make it accessible to other scholars, but for now Google drive will suffice (I think). I expect your code to be well documented, so others on the team (including me and any PhD students who come to work on the project) can scan and understand it easily. All code, research papers, and books that result from this project will credit undergraduate research apprentices for their contributions.