Measuring Breadth and Depth of Study Using Neural Embeddings

Annaliese Paulson[1]
University of Michigan
School of Education

[1]I am grateful for support from the PR/Award R305B200011 from the Institute of Education Sciences, U.S. Department of Education. I employ data and research from the College and Beyond II project. College and Beyond II is supported by a grant from the Andrew W. Mellon Foundation to the Inter-university Consortium for Political and Social Research (grant #1802-05485) at the University of Michigan under the auspices of the Mellon Research Forum on the Value of Liberal Arts Education. The findings and conclusions contained within are my own and do not necessarily reflect positions or policies of College and Beyond II, the Mellon Foundation, or the Department of Education.

Introduction

Contemporary curricular structures in American post-secondary education have remained relatively stable since their emergence at liberal arts colleges in the early 20th century (Wells, 2016). At most institutions, a student develops depth of study by devoting approximately one third of courses to curricular experiences in a major, devotes and approximately one third of courses to breadth of study through general education programs with the remaining courses devoted to electives (Brint et al., 2009; Lattuca and Stark, 2009, Wells, 2016). However, this division of the curriculum has emerged through historical processes rather than as the result of empirical research (Bok, 2006). Indeed, concerns have been raised about some of these curricular structures for nearly 50 years, with the Carnegie Foundation for the Advancement of Teaching calling general education "a disaster area" in 1977. However, 76% of institutions affiliated with the American Association of Colleges and University continue to rely, in part, on these same structures to ensure breadth of study (Hart Research Associates, 2016).

As states and institutions look to make more equitable institutions, it is natural that they consider whether the contemporary division of coursework motivated by ideas of breadth and depth of study continue to serve students and society well in the 21st century. We might wonder, for instance, whether working parents attending college should be required to pursue a similar curricular program inspired by liberal art students of the 1920s and 30s. For instance, both Texas (Undergraduate Education Advisory Committee, 2011) and Missouri (Gwaltney, 2020) have considered reducing the share of the curriculum devoted to general education and breadth of study in public institutions. This impulse to decrease requirements for breadth of study typically comes from concerns that additional requirements are making degree programs more difficult to complete in a timely fashion and thereby raise time to degree or detract from students developing specialization in a single field and depth of study. Arguments can be made, however, for both increasing and decreasing requirements for breadth and depth of study; ultimately, determining the optimal balance of breadth and depth will depend on research assessing the effect of increasing and decreasing breadth/depth on desired outcomes. Unfortunately, this necessary research is stymied by a failure to adequately measure depth and breadth of study.

In this exploratory paper, I develop measures of depth and breadth of study that rely on notions of course similarity. On this account, a student that enrolls in many highly similar courses is developing depth of study while a student that enrolls in many dissimilar courses is developing breadth of study. To do this, I draw on two approaches to learning neural embeddings of courses from natural language processing and learning analytics and present evidence that these embeddings capture intuitive notions of course similarity and content. In the first approach, I use the doc2vec algorithm from natural language processing (Le and Mikolav, 2014) to learn neural embeddings of course descriptions that capture the semantic similarity of courses. In the second approach I use the course2vec algorithm from learning analytics (Jiang and Pardos, 2020; Pardos et al., 2019; 2020; Pardos and Nam, 2020) to learn neural embeddings of courses from the structure of student transcripts. This second approach captures the structural similarity of courses. Using these learned neural embeddings of courses, I calculate the similarity of all pairs of courses students enroll in and aggregate these into measures of student level course-taking depth/breadth. Finally, I explore the relationships these measures have with salient social identities and fields of study. While exploratory, these initial results suggest that variation in breadth and depth of study is as strongly with students from some social backgrounds as it is with some fields of study. This exploratory work lays the foundation for future research into the mechanism through which students with various social identities navigate the postsecondary curriculum and provides quantitative

measures of breadth and depth of study that can be used to test relationships between the constructs motivating substantial portions of the American postsecondary curriculum and the desired outcomes of a postsecondary education.

Conceptualizing Breadth and Depth of Study

Traditionally, research has conceptualized depth and breadth of study along departmental boundaries. In the limited research examining the role of breadth of study for instance, a student's breadth of study is determined by coursework in departments. Seah et al. (2020 examine the effect of a reform that allowed some, but not all, students to enroll in more courses outside of their major departments at the National University of Singapore. They argue that increased coursework outside of a student's major represents breadth of study and, using a difference-in-difference design, find that enrolling in a broader course of study has no effect on student's short term labor market outcomes. Similarly, Goldhaber et al. (2015) examine the association between breadth of study and labor market outcomes using administrative data from the state of Washington. They first define a distance between two courses as the correlation between the number of courses student took in each of the two departments the courses are offered in. They then operationalize a student's breadth of study based on the average correlation between the average distances of all pairs of courses. Breadth of study in some terms of student's career is associated with increased earnings, but higher breadth of study over a student's entire course-taking career is associated with decreased earnings. Goldhaber et al. conclude that breadth of study may be best in moderation. In both cases, the department is the primary driver of breadth. The assumption seems to be that departmental boundaries meaningfully track the development of the diversity of knowledge, values, and skills that breadth of study is thought to entail. However, such an assumption obscures the relations between individual courses. On these measures, a pre-medical biology student enrolling in an English course on Shakespeare or an English course on the literature of disease are experiencing an identical amount of breadth. However, it seems that these represent quite different experiences. Although both courses would likely teach humanistic and literary modes of inquiry, the former is a more drastic move away from a pre-medical biology students' area of expertise while the latter seems to fit into the student's overall course-taking and might be regarded as furthering depth of study.

With regards, depth of study, a large literature has examined the effects of majors or credits accumulated in specific fields (for instance, Altonji et al., 2012; 2014; Bleemer and Mehta, 2022; Kinsler & Pavan, 2015; Stange, 2015). Again, this largely assumes that depth of study is comprised of specific courses in a department or major rather than in the kinds of knowledge, values, and skills a student accumulates through courses. While completing a major is one way to develop depth of study it is not the only way: majors are an instrumental tool to ensure depth of study and not equivalent to the concept itself. In both the research on breadth of study and that of depth of study, the assumption is that the knowledge, values, and skills learned through coursework can be neatly divided into departmental or major categories.

However, approaches that measure breadth and depth of study as coursework in departments or majors do not track the intuitive notions of breadth and depth. Intuitively, depth and breadth of study refer to the concentration/dispersion of knowledge, skills, values, and modes of reasoning a student acquires in their study while these threads of research largely rely on course enrollment in specific departments or majors. While departments are loosely organized around knowledge, skills, and values, they do not perfectly divide them. In a qualitative study of ideas of breadth and depth, Brady (n.d.) asks academic

advisors to evaluate the breadth and depth of student transcripts and discuss their reasoning. Two broad themes emerged from these interviews. First, advisors note that depth of study extends across departments, contradicting the notions of depth and breadth employed in previous studies. According to these academic advisors, a student majoring in environmental studies and the earth sciences is developing depth of study through complementary coursework in both of their two majors. Second, advisors noted that breadth and depth of study are difficult to understand in isolation from one another. From this perspective we may think that a pre-med student majoring in biology may appear to be pursuing a substantial breadth of study across many departments but, in reality, is completing a prescribed set of courses that prepare them for graduate study. Without the context of a student's major, it is difficult to know whether a student is truly developing breadth of study. As such, I argue that exploring the role depth and breadth play in the American curriculum requires us to identify measures of course similarity that examine the concentration and dispersion of knowledge, values, and skills learned across a transcript. Students develop depth of study by studying courses that are meaningfully similar to one another and breadth of study by exploring a variety of courses that are meaningfully dissimilar from one another.

However, Brady's qualitative data also made clear that a wide variety of definitions of breadth and depth were employed by advisors, suggesting more exploratory work needs to be done to clarify these concepts. This ambiguity in definitions motivates my empirical approach as I pursue what Grimmer et al. (2022) call an inductive exploratory approach to applying machine learning to my research question. They argue that the traditional approach to quantitative research emerged in a time when data was relatively scarce. This lack of data motivates researchers to theorize and develop hypotheses prior to examining their data and leads to a research paradigm in which exploratory data analysis of the kind I pursue is perceived as undermining the validity of inferences. However, with the rise of digital and administrative data, Grimmer, Roberts, and Stewart argue we no longer operate in a data scarce environment and, as such, can use statistical methods from machine learning to explore and organize a subset of our data and develop new inductive theory and measures. In this paradigm, we use statistical methods to help us explore data and suggest organizations of data that can be generative for future research. These measures can then be applied to unseen data and test the validity of our theories without undermining our inferences. In a world in which researchers can access tens of thousands of course transcripts or conceivably collect the course catalogs of most institutions online, researchers can afford to explore a portion of their data using quantitative methods and develop inductive theories before testing these theories deductively on new data. This explorative approach to quantitative research aligns itself more closely with the inductive paradigm of qualitative methods such as grounded theory than paradigms of quantitative research that stress forming testable hypotheses before examining data.

Given the preceding, I can now more precisely state the objectives for the remained of this paper. My objective is to develop statistical measures of breadth and depth of study that capture the intuitive notions discussed by advisors in Brady (n.d.). I develop these measures as an exploratory tool to help us organize student's course-taking with the belief that future research can use them as tools to examine the effect of breadth and depth of study on student outcomes. In what follows, drawing on neural embedding approaches from natural language processing and learning analytics, I develop two measures of course similarity that allow us to organize student course-taking into more and less similar coursework without relying on departmental boundaries. First, I develop measures of similarity that capture course similarity through the semantic content of course descriptions. Second, I develop measures of similarity that capture similarity through the structural similarity in student course taking. I argue that these measures provide a

new and rich way to organize the post-secondary curriculum that captures the intuitive notions of depth and breadth of study that organize the American post-secondary curriculum.

Data

To develop measures of course similarity and breadth/depth of study I draw on three sources of data: text data from course catalogs at many institutions including the University of Michigan, administrative data from the University of Michigan, and records of course equivalencies from the University of Michigan's Registrar. I describe each of these data sources below.

Text Data

Course catalog text data for this project comes from a variety of sources. The primary text data comes from the University of Michigan's institutional course catalog API and contains 8,821 unique course records with descriptions. Because the quality of word embeddings tends to improve when provided with greater amounts of training data, I supplement the course descriptions from the University of Michigan with course descriptions from three additional sources. First, I include 34,056 course descriptions from six partner institutions collected as part of the College and Beyond II project. Second, I include 32,411 course descriptions from three institutions that were generously provided to me from Coursicle, a for-profit company that provides college students resources from course catalogs and course schedules. Finally, I include 345,568 course descriptions collected from 96 institutions by the Incite research group at Columbia University. In total, this gives me a full dataset of 420,856 unique course descriptions.

In addition to these large course description datasets, I rely on a smaller set of course descriptions for validation. Drawing on the National Center for Education Statistics College Course Map, I identify ten courses present at 11 institutions across the country. The College Course Map provides a typology of courses based on nationally representative samples of student transcripts (Bryan and Simone, 2012). For each College Course Map code in my validation sample, if the institution offers a course that meets this code, I collect the associated course description from publicly available course catalogs. This generates a validation dataset of 988 course pairs that have similar course content across 11 institutions. Table 1 summarizes the validation course content and provide example descriptions. Supplementing course descriptions that cover the entirety of the University of Michigan, I use a limited list of 2,792 courses that meet general education requirements within the College of Literature, Science, and the Arts as a validation method.

Table 1:

Example Validation Course Descriptions

CCM Code	CCM Description	N	Example Description
54.0102	American History United States	11	Introduction to the nature and methods of historical study and examination of specific topics focusing on significant periods in the development of the U.S. and considering them in the light of certain elements shaping that history. Among these elements are the constitutional and political system; and the society's ideals, structure, economic policy, and world outlook.
23.1301	General Writing	11	Expository writing with emphasis on effective communication and critical thinking. Emphasizing the writing process writing topics are based on selected readings and on student experiences.
45.1101	Sociology	11	Fundamental concepts of sociology and introduction to the analysis of social problems and interactions (e.g., wealth, gender, race, inequality, family, crime) using sociological theories.
38.0101	Philosophy	11	Inquiry into the meaning and justification of fundamental ideas and beliefs concerning reality, knowledge, and values; application to relevant topics in ethics, religion, and politics.
52.0501	Business Communications	10	Principles of business communication through letters, memos, email, text messages, group leadership and participation and presentations. Clear, accurate, and focused communication; practical psychology with attention to communication ethics and diversity.
42.0101	General Psychology	11	A prerequisite to advanced courses; a broad survey of psychological science. Application of the scientific method to the empirical study of behavior with emphasis on individual and cultural differences.
38.0201	Religious Studies	11	Introduction to the academic study of religion through comparison among major traditions (Judaism, Christianity, Islam, Hinduism, Buddhism, etc.) and smaller communities.
52.0301	Accounting	11	Introduction to financial accounting and accounting information systems (AIS), including basic concepts, limitations, tools and methods. Use of AIS-generated information, including financial statements in decision making by investors, creditors, and other users external to the organization.
27.9995	Calculus	11	Vectors, functions, limits, derivatives, Mean Value Theorem, applications of derivatives, integrals, Fundamental Theorem of Calculus.
5.0209	Folklore Studies	8	A general study of the field of folklore including basic approaches and a survey of primary folk materials: folktales, legends, folksongs, ballads, and folk beliefs.

Administrative Data

In addition to text data, I use administrative data provided through the University of Michigan's Learning Analytics Research Architecture. This includes two data tables that include student level data including demographics and major, and course level attributes including student course-taking. For my analytical cohort, I include only data from 41,010 students who enter the University of Michigan as first-term freshman[2] between Fall 2010 and Fall 2016.

Course Similarity Data

Following Pardos and Nam (2019b), to validate the course2vec similarity measures I use a list of 916 course equivalencies from the College of Literature, Science, and the Arts at the University of Michigan. Pairs of courses in this list have content that has been deemed too similar to receive credit for both courses by the institution.

In Table 2, I provide a snapshot of a simulated transcript and associated course descriptions to provide greater context on my data. My objective in what remains in this paper is to develop methods that capture many of the notions of similarity that can be seen in Table 2, capturing the idea that ECON 370 and EARTH SCIENCE 380 have meaningful similarities, both because they have similar language use of semantic content and because they are more likely to co-occur on student transcripts because they are likely to be taken by students with a particular interest in natural resources and the environment.

Methods

Foundations of Neural Embeddings for Representation Learning

Embedding approaches learn a low-dimensional representation that preserves much of the variation in high dimensional spaces like language or transcripts. While the numeric values of learned embeddings can be difficult to interpret, their ability to reduce high-dimensional information like text into lower dimensions while preserving the meaningful variation makes them particularly appealing as measures of similarity (Grimmer, et al. 2022). While embedding approaches have historically been employed since at least the 1980s, embedding methods have become increasingly popular since Mikolav et al. (2013) developed word2vec to efficiently learn representations of words in a low-dimensional space. The intuitive idea of word2vec is captured by Firth's (1957) distributional hypothesis, pithily summarized by "you shall know a word by the company it keeps." word2vec seeks to predict a target word based on the that word's surrounding context. Mikolav et al. formalize Firth's intuition as follows. Given a sequence of context words, $w_1, w_2, \ldots w_t$ the objective of word2vec is to maximize the log probability of a target word:

[2] While measuring the depth and breadth of study to transfer students would be desirable, I am limited in my ability to acquire data that would allow my similarity measures to be meaningfully applied to students outside of the University of Michigan. The doc2vec requires course descriptions for courses and collecting course descriptions from all transfer institutions would be prohibitively time intensive. The course2vec approach requires transcripts from all institutions students are transferring from and has yet to be extended beyond two comparison institutions (Pardos et al., 2019a).

Table 2:
Simulated Student Transcript

Term	Subject	Catalog Number	Course Title	Course Description
Fall 2015	Environmental Sciences	317	Conservation of Biological Diversity	Overview of historic and present-day causes of species extinction, and of biological principles central to species conservation and sustainable management of ecosystems. Topics covered include episodes of extinction and diversification over earth history; geographic distribution strategies; and sustainable use of ecosystems. Weekly recitation sections discuss material from lectures, assigned readings and films, and perform computer and gaming simulations.
Fall 2015	Earth Sciences	331	Climate Change	This course examines the physical and chemical processes influencing Earth's climate and the methods of quantifying past and present climate change. Emphasis is placed on understanding the mechanisms of climate change from ice ages through the near future. The evidence of human-caused changes in climate is also discussed. Students with interests in global change and the environment are encouraged to enroll. A background in college science is not required.
Winter 2016	Environmental Sciences	490	War and the Environment: A Lethal Reciprocity	This seminar examines war and environmental degradation. We begin with the recognition that: a) war and the preparation for war typically lead to depletion and degradation of the biosphere; and b) resource mail-distributions, depletion, and degradation can frequently lead to armed conflict within and between territorial states.
Winter 2016	Earth Sciences	380	Natural Resources, Economics, and the Environment	This course deals with natural resource-related challenges in a complex society. The course discusses the origin, distribution, and remaining supplies of natural resources, including fertilizers, metals and fossil fuels, in terms of the economic, engineering, political, and environmental factors that govern their recovery, processing, and use. Topics covered in the course include nuclear waste disposal, strip mining, continent-scale water transfers, mineral profits and taxation, and estimation of remaining mineral reserves.
Winter 2016	Economics	370	Environmental and Resource Economics	This course focuses on the contribution economics has made in understanding and managing environmental and natural resource problems. The course will analyze the sources of environmental and natural resource problems using economic tools. Given this knowledge students will learn how these economic tools, through market-based incentives, may resolve these problems. Finally, we'll take a look at real policies and discuss the problems of transitioning policies from theory into the practical realm.

$$\frac{1}{T}\sum_{t=k}^{T-ik} \log p(w_t|w_{t-k}, \dots, w_{t+k})$$

where w_t is the target word and $w_{t-k}, \dots w_{t-1}$ and $w_{t+1}, \dots w_{t+k}$ are the context words in a k token window preceding and following the target word. For instance, given the sentence "the cat sat on the mat," we might seek to maximize the probability of predicting the target word "on" by providing the 2 context words preceding and following it, "cat", "sat", "the", "mat". Typically, the prediction task is done via a multiclass classifier like softmax or some variation like hierarchical softmax that captures the intuition of softmax while providing a more efficient training algorithm. In the case of a softmax multiclass classifier, we have:

$$p(w_t|w_{t-k}, \dots, w_{t+k}) = \frac{e^{y_{wt}}}{\sum_i e^{yi}}$$

Where each of y_i is an un-normalized log-probability for each output word *i*, computed as

$$(1)\ y = b + Uh(w_{t-k}, \dots, w_{t+k}; W)$$

where *U, b* are the softmax parameters and *h* is constructed by averaging the word vectors extracted from *W* that represent the context words. In *W*, each word is associated with an N-dimensional vector, where N is a tunable hyperparameter.

Typically, neural word embeddings are trained using stochastic gradient descent until training converges. Remarkably, after this training task, word2vec models tend to learn vector representations that capture semantically meaningful information about words. For instance, Mikolav et al. (2013) train a model that correctly identifies vectors that approximate the following equation: king - man + woman = queen. Representation learning models like word2vec tend to outperform simpler representations of language like bag-of-words in downstream natural language processing tasks and have been applied to a variety of social science concerning the similarity of words (for instance, Garg et al, 2018; Lucy et al., 2020; Rodman, 2020). However, because my unit of analysis is the course, not individual words, I consider two modifications of the word2vec approach.

doc2vec

Building on word embedding models, Le and Mikolov (2014) develop doc2vec to learn distributed representations of paragraphs and short documents. While word2vec learns neural representations that embed words in a low-dimensional space, doc2vec learns representations that embed the entire content of a short document. The only modification to training doc2vec rather than word2vec algorithms is a modification to equation (1). In the word2vec case, we train a model to predict the target word from an average of the context words vectors. In the doc2vec case, we assign each document a document id in a matrix *D* and include the vector associated with this id in constructing *h* from *W* and *D*. In word2vec, we use the average of vectors from the context words to predict our target; in doc2vec, we use the average of vectors from the context words and the vector associated with the document id. Like word2vec, in doc2vec words and document ids are associated with an N-dimensional vector, where N is a tunable hyperparameter. This process maps both documents and words to N dimensional vectors in the same vector space. For instance, given the sentences "the cat sat on the mat" with document id 1 and "a cat sat on the mat near me" with document id 2, we could predict the context word "on". With regards the

first document, we predict “on” with the average of vectors associated with “cat”, “sat”, “the”, “mat”, and the vector associated with document id 1. With regards to the second document, we would predict “on” with the average of vectors associated with “cat”, “sat”, “the”, “mat”, and the vector associated with document id 2. In this toy example, we would anticipate that the paragraph vectors associated with document ids 1 and 2 would be highly similar, because the text of the documents is very similar.

doc2vec, like word2vec, has been applied to social science questions. For instance, Nay (2016) use doc2vec to learn representations of governmental institutions, where the document id for a given speech is the institution that gave that speech. Analogously, in my case, I learn representations of courses, where the document id is the unique course identifier and the text used to train the model to represent a course is a course description. When measuring the similarity of two courses using doc2vec, I assume that each course generates a course description based on its content and that this content proxies the knowledge, values, and skills learned a student would learn in a course. As a result, the similarity of two courses’ vectors learned from course descriptions captures the similarity of those courses' content.

Preprocessing Text for doc2vec

doc2vec models operate on sequences of tokens. I pre-process course descriptions by lower casing all text, remove punctuation, split documents into tokens on white spaces, remove all words that occur fewer than five times, and remove stopwords.[3] I then remove administrative information such as notes about prerequisites, credit numbers, course equivalencies, and former course names. After this preprocessing, I remove course descriptions with fewer than three tokens remaining because the language used in these descriptions are unlikely to meaningfully capture the content of course. Largely, this means removing special topics and independent study courses that lack substantive information about their content in the descriptions provided. This results in a dataset of 5,700 course descriptions that covers approximately 91% of all courses among freshmen in my sample.

Training and validating doc2vec embeddings of courses

As noted above, the dimensionality of a word2vec embedding model is a tunable hyperparameter. I empirically tune the number of vectors in my doc2vec model by training 92 models for a range of plausible values from 8 to 100 vectors for 300 epochs then evaluating the performance of each model on three evaluation tasks. Embedding methods minimize an objective function but there is no guarantee that the results capture a statistical relationship of theoretical interest to social scientists. As such, the results of learned embeddings must be validated to be useful in developing theory in the social sciences.

Embedding methods can be evaluated with both intrinsic and extrinsic approaches (Grimmer et al., 2022). In intrinsic approaches to validation, we explore whether the similarity metrics captured by the model track similarity determined by experts. In extrinsic methods, we explore whether the learned embeddings are useful in downstream prediction tasks that we would anticipate they should be if we were accurately capturing the meaning of text. As my primary means of doc2vec model validation, I use intrinsic evaluation methods and provide supplementary extrinsic validation performance as further evidence that the learned doc2vec embeddings are meaningfully capturing course content.

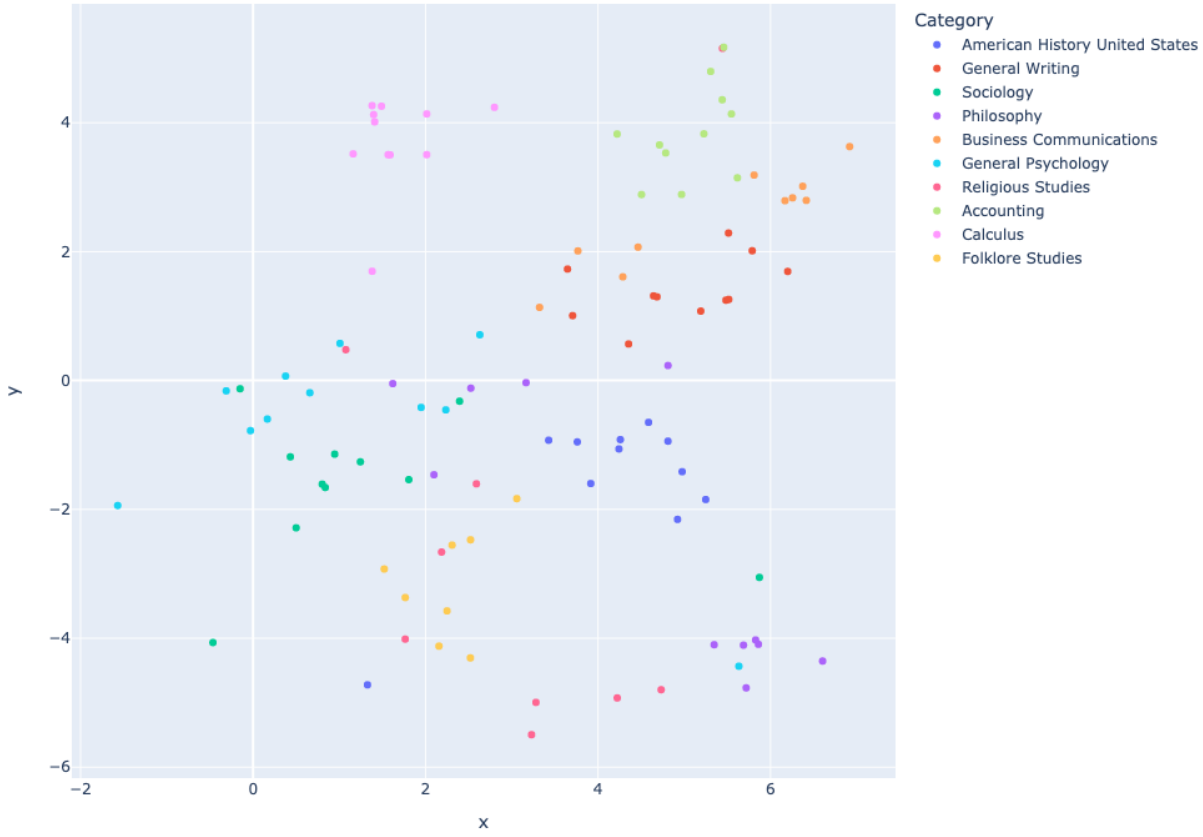
---

[3] All training of doc2vec and course2vec models is done in python using nltk (Bird et al., 2009), pandas (McKinney, 2010) and gensim (Rehurek & Sojka, 2011). To remove stopwords, I use nltk’s corpus of stopwords for English.

For my primary intrinsic evaluation methods, I evaluate whether the similarity of two courses with similar content is higher than the similarity of one of these courses and a randomly sampled course from my full corpus. As noted above, I collect a set of 988 validation pairs of similar courses across 11 institutions using the College Course Map typology to determine similar courses. For each validation pair of courses, I randomly sample 10 courses from my corpus and compare the similarity of the validation pair of courses against that of one of the courses and the randomly sampled courses. This yields 9,880 similarity scores. I choose the model with vector size that maximizes the number of courses correctly assigning higher similarity to the similar pair relative to the sampled course. Based on this validation method, I choose a model with vector size 13 that correctly assigns a higher similarity to validation pairs relatively to one course from the validation pair and a random course from my corpus approximately 91% of the time.

Figure 1 presents a t-SNE plot of this model's embeddings of my 105 validation courses, suggesting the chosen models' embeddings capture intuitive senses of these courses. The t-SNE algorithm allows us to visualize high-dimensional space by finding a two-dimensional representation of the higher dimensional space that preserves the distance between a point and its closest neighbors (Maaten & Hinton, 2008). In this figure, each point represents a course, and the location of each point (or course) is derived from a two-dimensional approximation of the thirteen-dimensional doc2vec embedding. In Figure 1, two points (or courses) are closer to one another if their associated embeddings are more similar, and courses are colored according to the type of course. This provides qualitative evidence that the learned embeddings are capturing intuitive ideas about courses. Examining the top right quadrant, for instance, we see that 1). general writing courses (red) are typically closer to business communication courses (orange), 2). business communication courses are typically closer to accounting (light green) courses, and 3). business communication courses are largely in the space between general writing and accounting courses. In sum, this suggests the chosen model's vector embeddings have successfully captured content of business courses like accounting, content of composition courses like writing, and the intuitive idea that business communication courses have a combination of both business and writing content.

Figure 1: t-SNE Plot of doc2vec Validation Courses



As supplementary evaluation metrics, I use the learned doc2vec embeddings of the best performing model on my intrinsic evaluation task as features in gradient boosted decision trees predicting structural aspects of courses. First, I predict whether courses satisfy general education requirements within the College of Literature, Science, and the Arts at the University of Michigan. For each requirement, I train a gradient boosted decision tree on all courses that meet that requirement and a random sample of courses that do not meet that requirement of equivalent size. This model uses the vector embeddings of those courses as features. I split the data into a 90 percent training set and 10 percent test set and select the number of trees based on five-fold cross-validation within the training set. Table 3 summarizes the model accuracy on unseen test data for each requirement. In all cases, the predictive accuracy of these models is higher than chance, ranging from moderate to strong predictive accuracy, even when provided with limited training data. This suggests that the learned embeddings of the chosen doc2vec model meaningfully capture the content of courses that would qualify courses to meet general education requirements.

Table 3

Accuracy of Models Predicting General Education Requirements Using doc2vec Embeddings

General Education Requirement	N	Training Accuracy	Test Accuracy
Natural Science	602	0.994	0.869
Social Science	928	0.966	0.785
Quantitative Reasoning	236	1	0.875
Writing Requirements	506	0.905	0.647
Race and Ethnicity	608	0.991	0.82
Humanities	2116	0.915	0.788
Math and Symbolic Analysis	78	1	0.875
Creative Expressions	174	1	0.944

As a second form of extrinsic validation of my chosen model's doc2vec embeddings, I divide departments that courses are offered in into a set of seven broad disciplines based on the Classification of Instructional Programs code associated with the department and predict the discipline associated with each course using gradient boosted decision trees. Again, I split the data into a 90-10 train and test set and tune the number of trees in the gradient boosted decision tree using five-fold cross validation within the training set. The mapping of departments into disciplines is available in Appendix A. This model achieves a training accuracy of 82% on the training set and 70% on the unseen test set, suggesting the learned embeddings meaningfully capture aspects of a course's disciplinary context.

Figures 2 and 3 display t-SNE plots of the learned embeddings for courses from the University of Michigan. In Figure 2, I include all courses with each course colored according to the broad discipline it is associated with. Qualitatively, moving from left to right we see the doc2vec embeddings capture intuitive notions of the "hardness" of the disciplines. Broadly, humanities courses (dark blue) are on the left, social sciences courses (light blue) are near the middle, and the hard sciences (red) and engineering (orange) are on the right with the social sciences largely dividing the humanities from the natural sciences and engineering. In Figure 3, I highlight only courses from three departments - math (green), statistics (blue), and biology (green) - in the same t-SNE locations as Figure 2. This suggests that the doc2vec embeddings also capture finer grained departmental distinctions. For instance, while the majority of courses from these three departments occupy the space associated with the natural sciences in Figure 2, in general, math and statistics courses are closer to one another than biology. However, math and statistics courses for computational biology and neuroscience are closer to biology courses, while courses in math and statistics education occupy a space closer to that of the social sciences, near the middle of the space. Again, this suggests that the doc2vec embeddings have captured intuitive ideas of course content.

Figure 2: t-SNE Plot of all Michigan doc2vec Embeddings

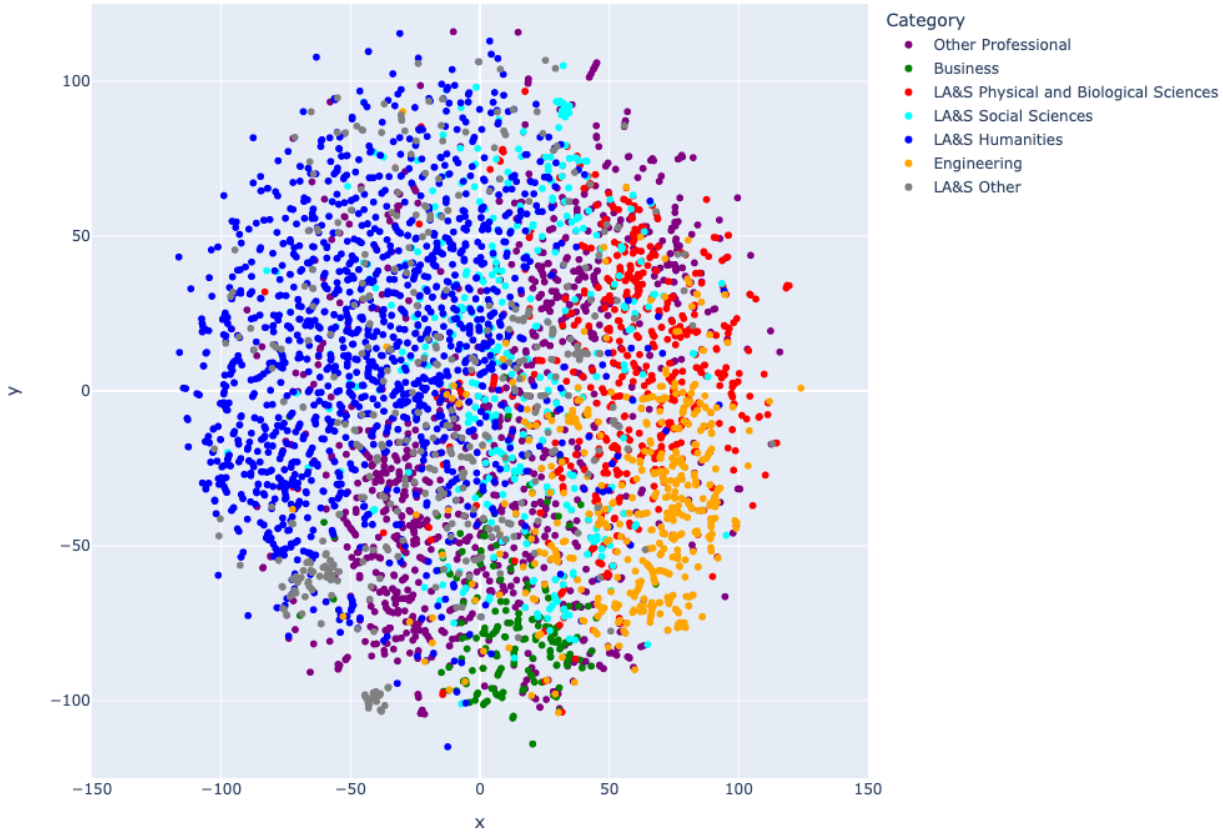


Figure 3: t-SNE Plot of Mathematics and Statistics doc2vec Embeddings

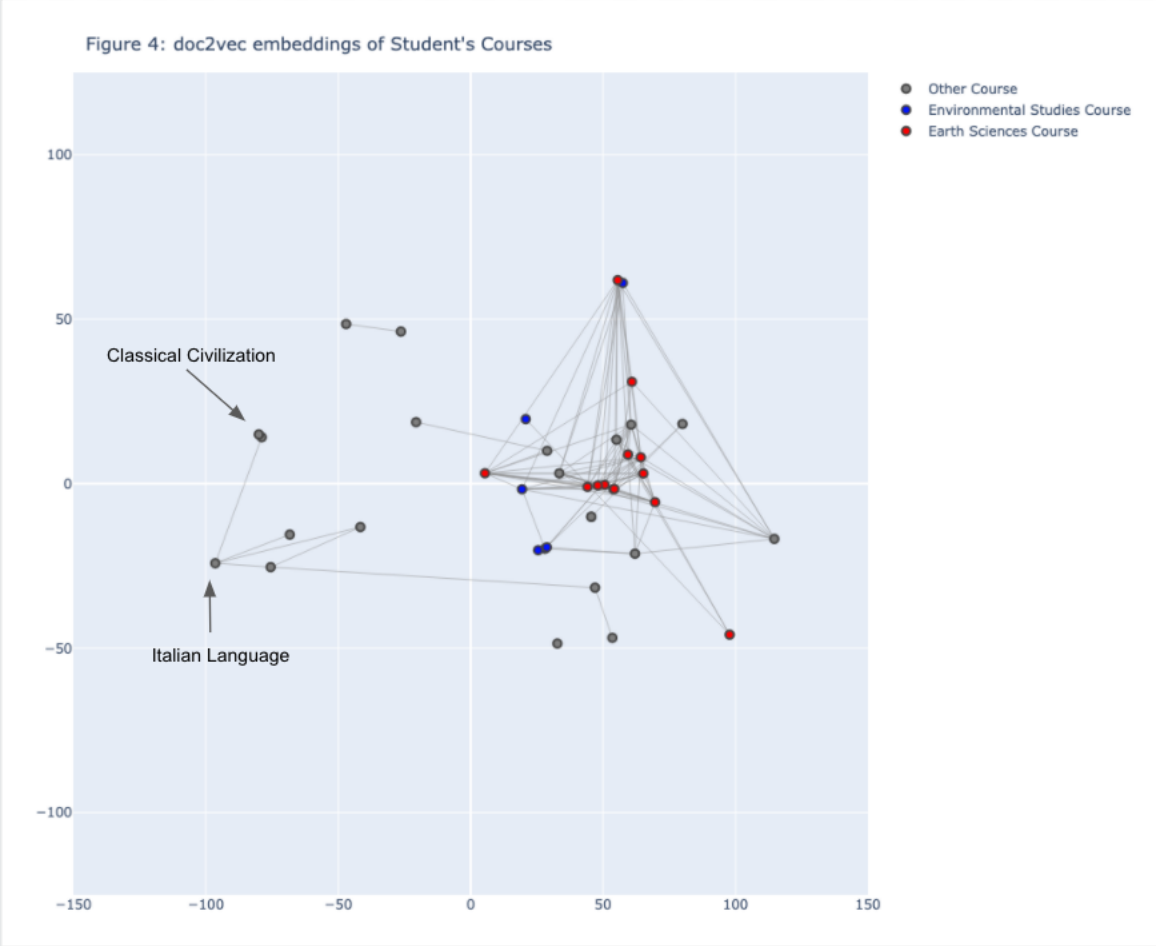


Figure 4 displays only the courses a single student that majored in environmental studies and earth sciences enrolled with courses in the same locations as the previous figures determined by the t-SNE algorithm. Because the two-dimensional approximation of the 13-dimensional vectors may not fully capture similarity of two courses in the higher dimensional embedding, I connect pairs of courses whose embeddings have a cosine similarity of 0.6 or higher with an edge. Capturing the intuitive notion of depth discussed by advisors in Brady (n.d.), we see that courses in environmental studies (red) and earth sciences (blue) are part of a large densely connected component of coursework across two departments and other courses in the natural sciences. While these courses are from two separate departments and are associated with two majors, they have very similar content. However, we also see the possible role of general education courses in creating breadth of study, as the loosely connected courses in the lower left corner satisfy the student's language requirements through coursework in Italian and the student's humanities requirements through coursework in the history of classical civilizations. The connections between these courses indicate that perhaps this student has a supplementary interest in Italian language and Roman culture.

In sum, I have presented a series of quantitative evaluation methods and qualitative figures, that suggest that the learned doc2vec embeddings capture the intuitive notions of course content I set out to measure. Further, in Figure 4, I present qualitative evidence that the learned embeddings capture intuitive notions of depth and breadth of study we sought to capture. Highly similar courses across departments are

placed closer together and have a high cosine similarity while the role of highly dissimilar courses outside the student's major can be seen as providing depth.

course2vec

In course2vec, we treat each transcript as a sentence of "words" where each word is a unique course id and use word2vec to learn a neural embedding representation of courses based on the context courses that are enrolled in. Intuitively, if two courses occur in similar contexts in a transcript, we may believe that they have similar content. Just as in word2vec, the training objective of course2vec is:

$$\frac{1}{T}\sum_{t=k}^{T-ik} \log p(w_t | w_{t-k}, \dots, w_{t+k})$$

However, rather than predict the target word from surrounding context words in a sentence, we predict a target course, from the surrounding context courses on a student transcript. In word2vec, when given the sentence "the cat sat on the mat", we predict "on" from the vectors representing "cat", "sat", "the", and "mat." In course2vec, when given the student transcript "MATH 105, CHEM 112, BIOL 130, ENGL 125, SPAN 115", we predict "BIOL 130" from the vectors representing "MATH 105", "CHEM 112", "ENGL 125", and "SPAN 115." In a series of studies, Pardos and collaborators have shown that course2vec embeddings capture disciplinary divides, predict structural features of courses including the language used in course descriptions, and can be useful in downstream tasks including course recommendation and course articulation (Pardos et al., 2019a, 2019b; Pardos and Nam, 2020; Jiang and Pardos, 2020).

When measuring similarity of two courses using course2vec, I assume that each course occurs in a location on a student transcript based on its content and, as a result, the similarity of two courses' locations on student transcripts capture the similarity of those courses' content. Courses that are more similar are more likely to be found in similar places on student transcripts and courses that are less similar are less likely to be found in similar places on student transcripts.

Training and validating course2vec embeddings of courses

To convert a transcript into a sentence, following Pardos and Nam (2020), I associate each course id with a token randomly shuffle the order of courses within a semester, then append the courses together into a single "sentence." At the University of Michigan, each course is associated with a unique id that is stable over time (in cases where the catalog number or subject description change) and across cross-listed courses. I tune both the vector size and context window of course2vec models and again evaluate the model intrinsically and extrinsically. I explore models with between five and 300 vectors and context windows of lengths between two and thirty. Because course2vec is learned from student transcripts, the coverage is much higher than that of doc2vec. Prior to training, I remove all unique course ids that have fewer than five instances across transcripts, to avoid noisy vectors. After removing these courses, I am able to associate course2vec vectors with more than 99.9% of courses in my sample.

Because course2vec learns neural embeddings of courses at a single institution, I cannot use the same intrinsic validation set as doc2vec. Instead, following Pardos et al. (2019b), I evaluate the similarity of 916 pairs of course equivalencies. These are courses that have content that has been determined to be so similar a student can only receive credit for one course from the pair. Again, I compare the similarity of a pair of similar courses against that of one of the courses and a random selection of 10 other courses, resulting in 9160 validation pairs. I choose the model that maximizes the number of times a pair of similar courses is more similar than one of that pair of courses and sampled courses. Results of this intrinsic evaluation task suggest a model with 140 vectors and a context size of 23 that correctly assigns higher similarity scores to approximately 91% of evaluation pairs.

Following model selection, I extrinsically evaluate the model's learned embeddings by again predicting whether the course meets general education requirements and the broad discipline the course is associated with gradient boosted decision trees, tuning the number of trees using cross-validation on 90 percent training set and evaluating performance on a 10 percent unseen test set. Table 4 summarizes the results of the extrinsic evaluation task of predicting general education requirements, showing that, in line with the work of Pardos and colleagues, the course2vec embeddings are reasonably effective in predicting course features. However, the course2vec embeddings appear to be less effective at predicting general education requirements than the best doc2vec embeddings in my context.

Similarly, the performance of gradient boosted decisions trees trained to predict a course's discipline using course2vec embeddings suggests the learned embeddings have some predictive power. After tuning, the trained model achieved 91% accuracy on the training set and 76% accuracy on the unseen test, outperforming the doc2vec embeddings on this extrinsic evaluation task. In sum, these intrinsic and extrinsic quantitative evaluation tasks suggest that the learned course2vec embeddings are meaningfully capturing similarity of courses and can predict structural aspects of courses.

Table 4:
Accuracy of Models Predicting General Education Requirements Using course2vec Embeddings

General Education Requirement	N	Training Accuracy	Test Accuracy
Natural Science	526	1	0.755
Social Science	710	0.958	0.789
Quantitative Reasoning	202	1	0.667
Writing Requirements	442	1	0.644
Race and Ethnicity	342	1	0.743
Humanities	1398	0.967	0.7
Math and Symbolic Analysis	78	1	0.625
Creative Expressions	288	0.996	0.862

In Figures 5, 6, and 7, I repeat many of the same visualizations previously shown with doc2vec embeddings using course2vec embeddings. In Figure 5 I show courses colored by discipline, in Figure 6 I show just mathematics, statistics, and biology courses, and in Figure 7, I show just the courses from the same student majoring in earth sciences and environmental science examined previously and add edges between pairs of courses if their cosine similarity is higher than 0.4. At a high level, these figures reflect similar trends to those in Figures 2, 3, and 4. In Figure 5, for instance, natural science and engineering courses generally cluster closer to each other and the social sciences while humanities courses are more likely to be near the social sciences than the natural sciences and engineering. In Figure 6, we see that math, statistics, and biology courses tend to cluster near each other in the portion of the space we attribute to the natural sciences and that there is a fairly distinct cluster of math and statistics courses near each other and farther away from biology courses, although the differences are less defined than when using the doc2vec embeddings. Finally, in Figure 7, we again see a connected component of courses across environmental sciences (red), earth sciences (blue), and other natural sciences in the top left corner.

Figure 5: t-SNE Plot of all Michigan course2vec Embeddings

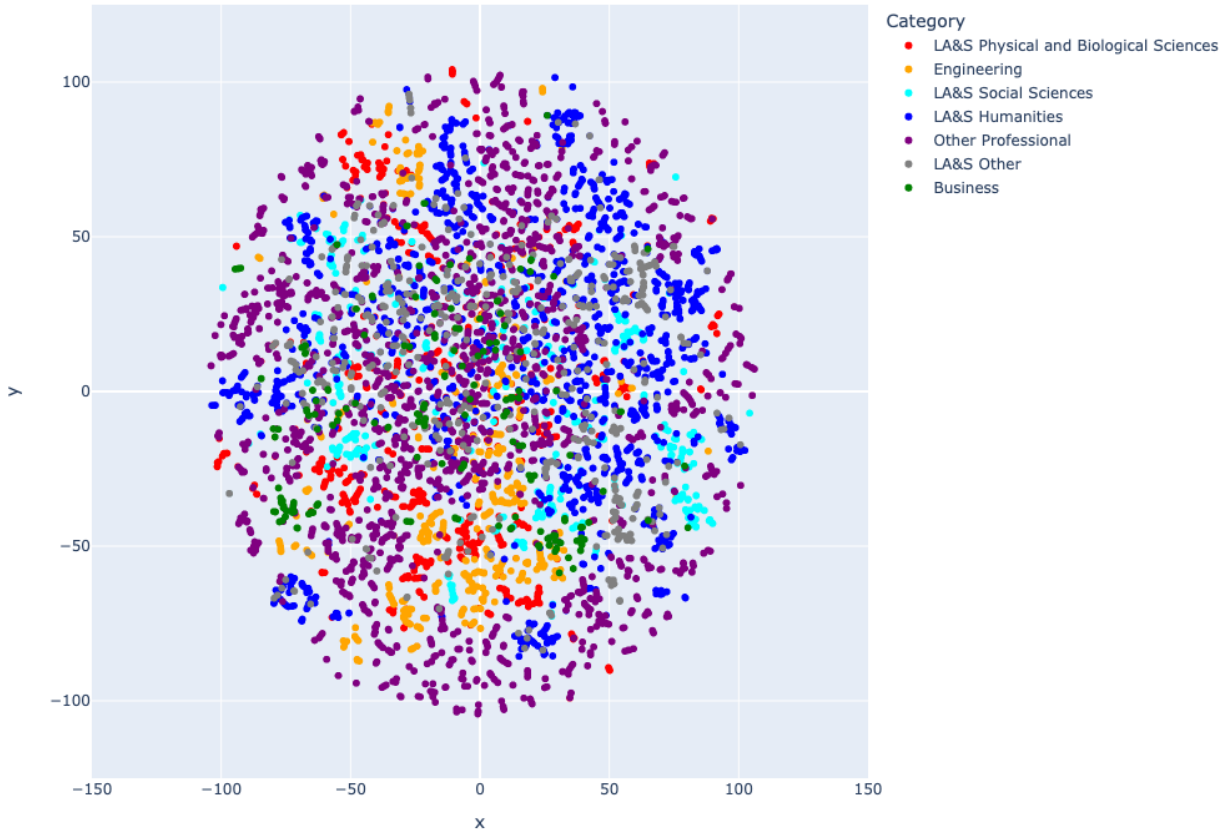


Figure 6: t-SNE Plot of Math, Statistics, and Biology course2vec Embeddings

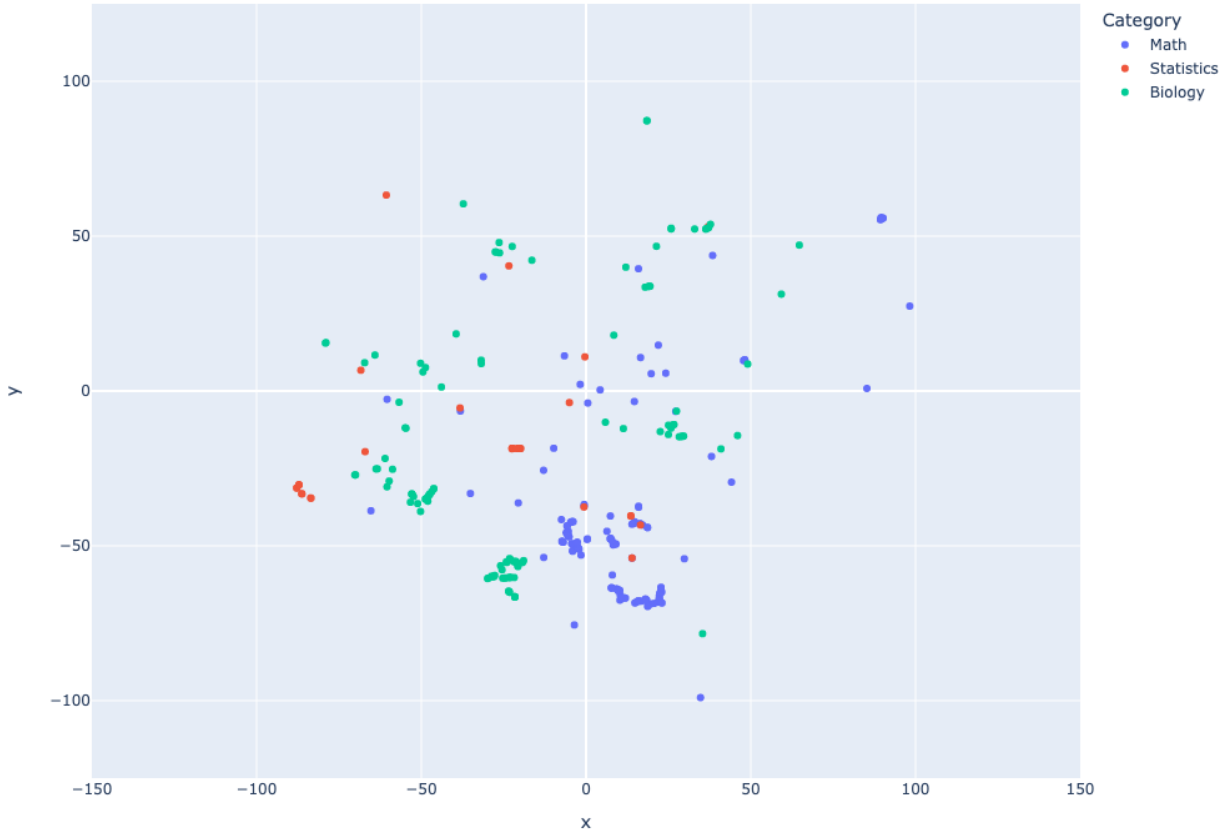
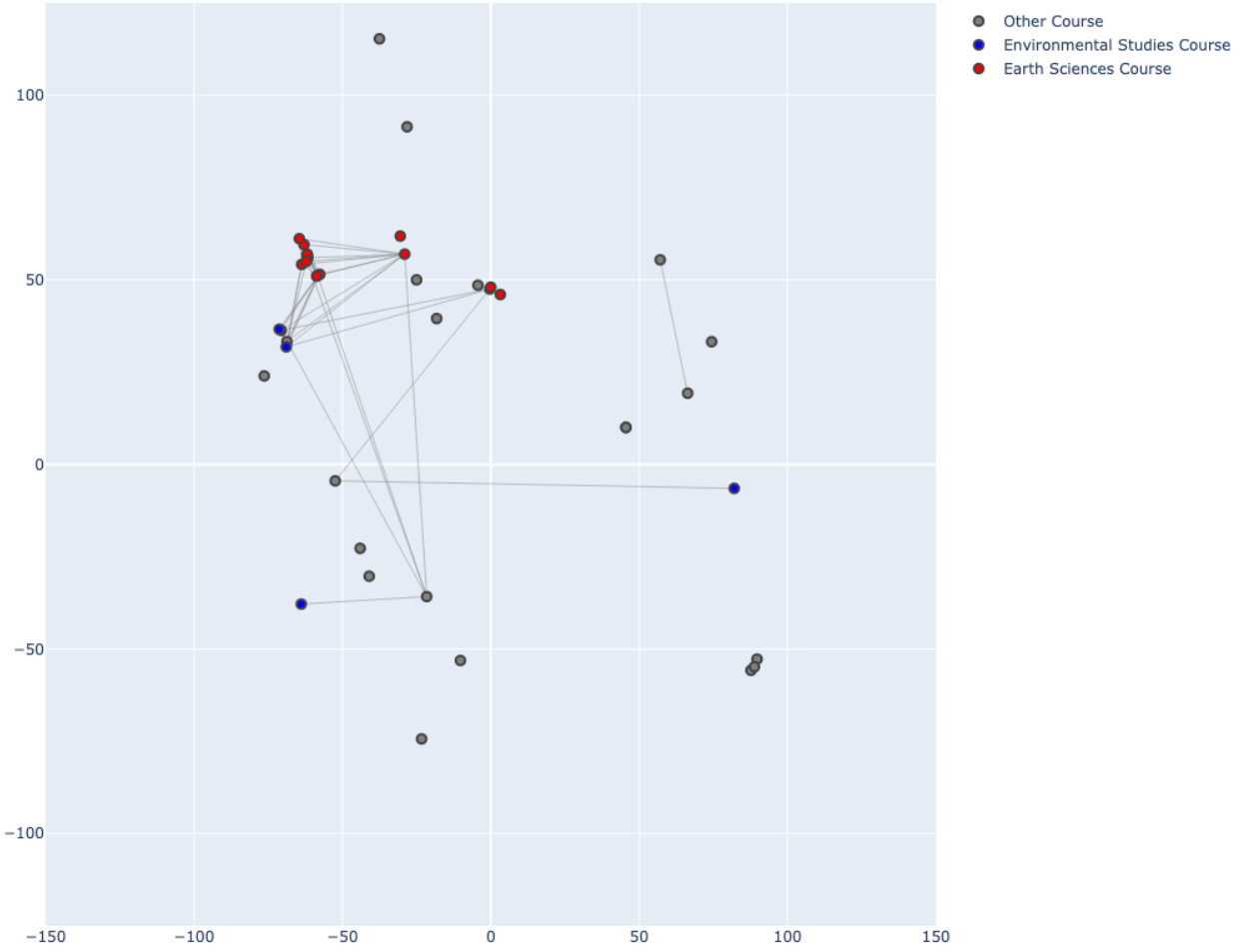


Figure 7: course2vec embeddings of Student's Courses



Networks, Quantitative Measures of Depth and Breadth, and Descriptive Results

To convert course similarity measures of pairs of courses on a student's transcripts into aggregated measures of depth and breadth across a student's course-taking, I take two approaches. First, I take a simple mean of the cosine similarity of all pairs of courses a student took using both embedding approaches. Second, I convert similarity measures into a course similarity network. For each pair of courses a student enrolls in, I calculate the cosine similarity between those courses using both the doc2vec and course2vec embeddings. I then create a network where nodes represent each course a student takes during their enrollment and an edge between two nodes indicates that the courses have a high similarity according to the respective embedding. Figures 4 and 7, demonstrate this network for a single student's courses. Based on a qualitative examination of course similarity among courses in transcripts, I say that a course has high doc2vec similarity if the cosine similarity between two courses' doc2vec embeddings is 0.6 or higher and a course has a high course2vec similarity if the cosine similarity between two courses' course2vec embeddings is 0.4 or higher.

Once converted into a network, we could examine the structure of student course-taking using any number of traditional network analysis measures including network density, betweenness, centrality

of nodes, and community detection. To capture a simple intuition regarding breadth and depth, I focus on network density. Intuitively, a student with very deep learning in a topic will take very similar pairs of courses while a simple notion of breadth involves taking very dissimilar pairs of courses. To calculate the share of similar course pairs, we take the number of edges present in the network - that is the number of course pairs that have a high similarity - and divide by the number of possible edges in the network or the total combinations of course pairs a student enrolls in. For each student, this yields a number between zero and one that captures the density of similar course pairs they enroll in. A 0.01 increase in this score represents a one-percent increase in the number of course pairs that are highly similar. Using this approach, a student with a network density of zero experiences maximal breadth while a student with a network density of one experiences maximal depth.

I calculate the mean similarity and the network density of all student's course-taking and link these measures with student level characteristics from the university's administrative data. In Table 5, I describe the mean similarity score and network density of graduates across race and ethnicity, sex as recorded in university administrative records[4], estimated family income, and major discipline.

Recall that course2vec measures structural similarity of courses: two courses are more similar if they occur in similar places on student transcripts. In contrast, doc2vec measures semantic similarity: two courses are more similar if the language they use in their course descriptions is similar. This difference between the doc2vec similarity and coruse2vec similarity can help to explain variation in these measures. For example, as measured by course2vec, the mean similarity and network density of humanities majors is higher than that of engineering majors. In contrast, similarity derived from doc2vec is higher for engineering majors than humanities majors. This suggests that, relative to engineering majors, the language used in courses that humanities majors enroll in is less similar but the course-taking patterns they follow are more similar. If all humanities majors were to enroll in chemistry courses in their first semester alongside introduction to composition, the course2vec similarity of these two courses would be high, although the doc2vec representations of the text used in their descriptions is unlikely to be.

We can see similar dynamics with regards to particular social identities. Relative to male students, female students tend to enroll in more similar course-taking patterns but less semantically similar courses while, among race and ethnicity groups, Asian students enroll in the least structurally similar pairs of courses and the most semantically similar pairs of courses.

[4] The university documentation is relatively unclear whether this variable measures sex or gender and the timing of when this variable is measured. Over the timeframe I accessed this data, the documentation changed the name of the variable in question from gender to sex but does not indicate how this variable is measured (and if truly measuring sex, what determines a student's sex in the data). In my view, social identities like sex, gender, and race are meaningfully socially constructed and I do not mean to endorse the view that these are fixed attributes. However, I am limited by the administrative data available and believe that it is helpful to examine the extent to which student's social identities shape course-taking, breadth, and depth, although the measures are flawed.

Table 5:
Means Similarity and Network Density by Demographics and Majors

	N	Proportion of Sample	course2vec Similarity > 0.4	course2vec Mean Similarity	doc2vec Similarity > 0.6	doc2vec Mean Similarity
Overall Mean	41010	1.000	0.116	0.103	0.141	0.319
Female	20907	0.510	0.130	0.110	0.131	0.307
Male	20103	0.490	0.102	0.096	0.152	0.331
2 or More Races	1467	0.036	0.118	0.103	0.136	0.316
Asian	7008	0.171	0.091	0.088	0.153	0.332
Black	1537	0.037	0.120	0.104	0.114	0.288
Hispanic	1939	0.047	0.124	0.107	0.137	0.316
Race Not Indicated/Other	2280	0.056	0.111	0.100	0.143	0.321
White	26779	0.653	0.123	0.107	0.140	0.318
Don't Know Family Income	9119	0.222	0.120	0.105	0.140	0.318
Less than $100,000	10700	0.261	0.116	0.102	0.138	0.316
More than $100,000	20955	0.511	0.114	0.102	0.144	0.321
Business	3781	0.092	0.050	0.055	0.167	0.326
Engineering	6754	0.165	0.102	0.102	0.198	0.383
LA&S Humanities	4571	0.111	0.144	0.116	0.095	0.271
LA&S Other	1806	0.044	0.118	0.105	0.094	0.267
LA&S Physical and Biological Sciences	5484	0.134	0.065	0.074	0.146	0.329
LA&S Social Sciences	8385	0.204	0.095	0.096	0.103	0.281
Other Professional	10229	0.249	0.183	0.136	0.153	0.332

In Table 6, I describe the mean similarity score and network density of one class of entering freshman over eight semesters and four years. For each term, I calculate the similarity of all pairs of courses each student took within that term. Unsurprisingly, mean scores are lowest in the first fall and

winter of this cohort's time in college, suggesting breadth of study is most intense in the first year. Depth of study as measured by doc2vec embeddings peaks in the Winter of junior year, but depth of study as measured by course2vec embeddings increases monotonically over the course of a student's enrollment. While students are enrolling in semantically distinct courses in their final year, they also seem to be following somewhat predictable course-taking patterns.

Table 6:
Course-taking Similarity by Term

Term	course2vec Similarity > 0.4	course2vec Mean Similarity	doc2vec Similarity > 0.6	doc2vec Mean Similarity
Freshman Fall	0.026	0.048	0.197	0.386
Freshman Winter	0.021	0.029	0.187	0.371
Sophomore Fall	0.037	0.072	0.219	0.392
Sophomore Winter	0.062	0.08	0.223	0.393
Junior Fall	0.144	0.152	0.242	0.429
Junior Winter	0.18	0.177	0.276	0.455
Senior Fall	0.195	0.195	0.245	0.429
Senior Winter	0.225	0.218	0.243	0.419

Exploratory Regression Analysis

In Table 7, I present exploratory regressions of associations with depth and breadth, regressing measures of course similarity on demographics, field of study, and high school GPA using OLS. My analytical sample for this portion of the analysis contains 39,832 freshman who entered the University of Michigan between Fall 2010 and Fall 2016, received a bachelor's degree, and have a non-missing high school GPA. In specification (1), I regress on the mean cosine similarity of all pairs of courses a student takes using course2vec embeddings. In specification (2), I regress on the proportion of pairs of courses that have a cosine similarity above 0.4 using course2vec. In specification (3) I regress on the mean cosine similarity of course pairs using doc2vec embeddings and in (4), I regress on the proportion of courses with cosine similarity greater than 0.6 using doc2vec embeddings. My demographics variables include categorical variables of students reported estimated family income (with reference category of estimated family income greater than $100,000), race and ethnicity (with reference category of White), and sex (with reference category of Female). The field of study variable is a categorical variable measuring the broad discipline the student received their first degree in (with reference category of majoring in the Natural Sciences), and high school GPA is a continuous variable. I stress that these regressions should be interpreted solely as associations and should not be read as claiming causal relationships between variables.

Table 7:
OLS Regression of Course Similarity on Demographics and Field of Study

	Dependent variable:			
	course2vec Sim.	course2vec Sim. > .4	doc2vec Sim.	doc2vec Sim. > .6
	(1)	(2)	(3)	(4)
Don't Know Family Income	0.002***	0.002***	-0.001*	-0.002***
	(0.001)	(0.001)	(0.001)	(0.001)
Family Income < $100,000	-0.0001	0.002***	-0.003***	-0.003***
	(0.001)	(0.001)	(0.001)	(0.001)
2 or More Races	-0.007***	-0.006***	-0.003*	-0.004***
	(0.001)	(0.002)	(0.001)	(0.002)
Asian	-0.015***	-0.019***	0.005***	0.004***
	(0.001)	(0.001)	(0.001)	(0.001)
Black	-0.008***	-0.009***	-0.012***	-0.008***
	(0.001)	(0.002)	(0.001)	(0.002)
Hispanic	-0.003***	-0.002	-0.0002	-0.001
	(0.001)	(0.002)	(0.001)	(0.001)
Race Not Indicated/Other	-0.007***	-0.010***	-0.001	-0.002
	(0.001)	(0.001)	(0.001)	(0.001)
Male	-0.012***	-0.018***	0.008***	0.007***
	(0.0005)	(0.001)	(0.001)	(0.001)
Business Major	-0.018***	-0.011***	-0.003***	0.020***
	(0.001)	(0.001)	(0.001)	(0.001)
Engineering Major	0.030***	0.030***	0.053***	0.051***
	(0.001)	(0.001)	(0.001)	(0.001)
Humanities Major	0.036***	0.040***	-0.053***	-0.046***
	(0.001)	(0.001)	(0.001)	(0.001)
Other Liberal Arts Major	0.025***	0.020***	-0.058***	-0.046***
	(0.001)	(0.002)	(0.001)	(0.002)
Social Science Major	0.019***	0.009***	-0.045***	-0.040***
	(0.001)	(0.001)	(0.001)	(0.001)
Professional Major	0.061***	0.082***	0.004***	0.010***
	(0.001)	(0.001)	(0.001)	(0.001)
High School GPA	-0.021***	-0.021***	0.021***	0.032***
	(0.001)	(0.002)	(0.001)	(0.001)
Constant	0.164***	0.136***	0.244***	0.020***
	(0.005)	(0.007)	(0.005)	(0.006)
Observations	39,823	39,823	39,823	39,823
R^2	0.263	0.232	0.350	0.299
F Statistic (df = 15; 39807)	948.981***	799.999***	1,426.295***	1,132.697***

Note: * p < .10 **p <.05***p<0.01

Unsurprisingly, we can see that major field of study is a statistically significant and relatively large predictor of a student's course similarity, all else held equal. For instance, looking at column (2), relative to majoring in natural science, majoring in engineering is associated with a 3-percentage point increase in the share of highly similar course pairs a student enrolls in while majoring in social science is associated with a .9 percentage point increase, all else being equal. However, in some cases, student's social identities have as large an or larger magnitude of association with measures of course similarity. Relative to female students, male students are associated with a 1.8 decrease in the share of highly similar pairs of courses, as measured by course2vec. In contrast, some social identities show little meaningful variation. If we take estimate family income as a proxy for socio-economic class, for instance, we see statistically significant but very small estimates across income brackets, suggesting there may not be as large a difference in depth and breadth of study across class.

Limitations

When using both course2vec and doc2vec approaches we make assumptions about what counts as similarity. However, these assumptions may not perfectly align with intuitive ideas about similarity operating in discussions of depth and breadth. For instance, in the doc2vec case, the model may assign high similarity to two courses if they use similar language about assignments or structure. For instance, an introduction to chemistry course that uses words like "survey", "introduction", and "quiz" in its description and an introduction to Greek course that uses similar words may have high similarity according to doc2vec despite seeming to represent drastically different modes of teaching and course content. Similarly, course2vec defines similarity based on the structure of student transcripts. Given this, courses that students tend to enroll in early in their career may have strong similarity; since students often take introductory courses in a variety of subjects in their first few terms, these courses may have high similarity despite have different content. Pardos et al. (2020) find that the large majority of lower division courses tend to cluster near one another and have high similarity while more specialized coursework in specific disciplines is relatively distinct from lower division courses. This means that these approaches may not perfectly capture notions of similarity implicit in discussions of breadth and depth of study.

Future Work

In this study, I explored two approaches to course similarity that allow us to derive measures of breadth and depth of study and link these measures with student characteristics. However, much of the motivation for depth and breadth of study concerns how they affect longer term life outcomes such as effective democratic participation, labor market outcomes, and life satisfaction (Bok, 2006; Goldhaber et al., 2015; Seah et al., 2020). As such, a necessary next step is to analyze the relationship between measures of breadth and depth and these outcomes. For instance, using a human capital framework from economics, we might associate breadth of study with general skills and depth of study with specific skills and make a set of predictions about the relationship between depth/breadth of study and labor market outcomes. Using this framework, we might anticipate that students with greater depth have higher wages immediately after graduation due to greater specific skills, while students with greater breadth of study wages will grow more quickly over time and their labor market outcomes may be more robust to labor market shocks.

Methodologically, there is also potential to improve similarity measures. For instance, one could combine the doc2vec and course2vec approaches in a multi-task learning approach. Such an approach would simultaneously learn an embedding for each course that maximize its ability to predict courses on a transcript and the contents of course descriptions. Finally, I only calculated depth/breadth of study at a single institution. Finding ways to align measures across a wide variety of institutions is an important area for future research.

Conclusion

Institutions devote substantial effort to ensuring students balance both breadth and depth. However, we have little empirical evidence that these curricular structures improve student outcomes and are worth requiring for all students. Drawing on approaches to neural embeddings from learning analytics and natural language processing techniques and novel sources of data like course description text, I developed tools that make legible the complexity of depth and breadth in course-taking at one post-secondary institution through measures of course similarity. I provided evidence that my measures of similarity capture notions of course content through both intrinsic and extrinsic evaluation methods. Aggregating these measures to the student level, I then explored how these measures of course-taking breadth and depth correlate with student demographics and field of study. While exploratory, this analysis suggests that some social identities may play as large or larger a role as field of study in shaping breadth and depth. Developing meaningful measures of breadth and depth of study is a necessary prerequisite for future research that explores how these constructs may affect longer term outcomes and whether institutions should continue to require participation in curricular structures that were developed for students of the early 20th century.

References

Altonji, J. G., Blom, E., & Meghir, C. (2012). Heterogeneity in human capital investments: High school curriculum, college major, and careers. *Annu. Rev. Econ.*, *4*(1), 185–223.

Altonji, J. G., Kahn, L. B., & Speer, J. D. (2014). Trends in Earnings Differentials across College Majors and the Changing Task Composition of Jobs. *American Economic Review*, *104*(5), 387–393. https://doi.org/10.1257/aer.104.5.387

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Bok, D. (2006). *Our Underachieving Colleges: A Candid Look at How Much Students Learn and Why They Should Be Learning More*. Princeton University Press.

Bleemer, Z., & Mehta, A. (2022). Will studying economics make you rich? A regression discontinuity analysis of the returns to college major. *American Economic Journal: Applied Economics*, *14*(2), 1–22.

Brady, J. (n.d.). *College and Beyond II Phase 1 Validation Study*. Unpublished Manuscript.

Brint, S., Proctor, K., Murphy, S. P., Turk-Bicakci, L., & Hanneman, R. A. (2009). General Education Models: Continuity and Change in the U.S. Undergraduate Curriculum, 1975–2000. *The Journal of Higher Education*, *80*(6), 605–642. https://doi.org/10.1080/00221546.2009.11779037

Bryan, M. & Simone, S. (2012). 2010 College Course Map (NCES 2012-162rev). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Firth, John Robert. 1957. (1957). *Studies in Linguistic Analysis*. Wiley-Blackwell.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Goldhaber, D., Cowan, J., Long, M., & Huntington-Klein, N. (2015). College Curricular Dispersion: More Well-Rounded or Less Well Trained? CEDR Working Paper. WP# 2015-6. *Center for Education Data & Research*.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Hart Research Associates. (2016). *Recent trends in general education design, learning outcomes, and teaching approaches: Key findings from a survey among administrators at AAC&U member institutions*. Washington, DC: Association of American Colleges and Universities.

Jiang, W., & Pardos, Z. A. (2020). Evaluating Sources of Course Information and Models of Representation on a Variety of Institutional Prediction Tasks. *International Educational Data Mining Society*.

Kinsler, J., & Pavan, R. (2015). The specificity of general human capital: Evidence from college major choice. *Journal of Labor Economics*, *33*(4), 933–972.

Lattuca, L. R., & Stark, J. S. (2009). *Shaping the college curriculum: Academic plans in context*. John Wiley & Sons.

Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In E. P. Xing & T. Jebara (Eds.), *Proceedings of the 31st International Conference on Machine Learning* (Vol. 32, Issue 2, pp. 1188–1196). PMLR. https://proceedings.mlr.press/v32/le14.html

Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2579–2605.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [Cs]*. http://arxiv.org/abs/1301.3781

Nay, J. J. (2016). Gov2Vec: Learning Distributed Representations of Institutions and Their Legal Text. *Proceedings of the First Workshop on NLP and Computational Social Science*, 49–54. https://doi.org/10.18653/v1/W16-5607

Pardos, Z. A., Chau, H., & Zhao, H. (2019a). Data-Assistive Course-to-Course Articulation Using Machine Translation. *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale*, 1–10. https://doi.org/10.1145/3330430.3333622

Pardos, Z. A., Fan, Z., & Jiang, W. (2019b). Connectionist recommendation in the wild: On the utility and scrutability of neural networks for personalized course guidance. *User Modeling and User-Adapted Interaction*, *29*(2), 487–525. https://doi.org/10.1007/s11257-019-09218-7

Pardos, Z. A., & Nam, A. J. H. (2020). A university map of course knowledge. *PLOS ONE*, *15*(9), e0233207. https://doi.org/10.1371/journal.pone.0233207

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

Rodman, E. (2020). A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors. *Political Analysis*, *28*(1), 87–111. https://doi.org/10.1017/pan.2019.23

Seah, K. K. C., Pan, J., & Tan, P. L. (2020). Breadth of university curriculum and labor market outcomes. *Labour Economics*, *65*, 101873. https://doi.org/10.1016/j.labeco.2020.101873

Stange, K. (2015). Differential Pricing in Undergraduate Education: Effects on Degree Production by Field. *Journal of Policy Analysis and Management*, *34*(1), 107–135. https://doi.org/10.1002/pam.21803

Undergraduate Education Advisory Committee. (2011). *Revising the State Core Curriculum: A focus on 21st century competencies*. Texas Higher Education Coordinating Board. https://reportcenter.highered.texas.gov/reports/data/revising-the-state-core-curriculum-a-focus-on-21st-century-competencies/

Wells, C. A. (2016). Realizing General Education: Reconsidering Conceptions and Renewing Practice. *ASHE Higher Education Report*, *42*(2), 1–85. https://doi.org/10.1002/aehe.20068

Appendix A: Classification of Instructional Programs to Disciplines

Discipline	Two Digit CIP Code	Two Digit CIP Description
Business	52	Business, management, marketing, and related support services
Engineering	14	Engineering
Engineering	15	Engineering technologies/technicians
LA&S Humanities	5	Area, ethnic, cultural, and gender studies
LA&S Humanities	9	Communication, journalism, and related programs
LA&S Humanities	16	Foreign languages, literatures, and linguistics
LA&S Humanities	23	English language and literature/letters
LA&S Humanities	38	Philosophy and religious studies
LA&S Humanities	39	Theology and religious vocations
LA&S Humanities	54	History
LA&S Other	24	Liberal arts and sciences, general studies, and humanities
LA&S Other	30	Multi/interdisciplinary studies
LA&S Physical and Biological Sciences	26	Biological and biomedical sciences
LA&S Physical and Biological Sciences	27	Mathematics and statistics
LA&S Physical and Biological Sciences	40	Physical sciences
LA&S Physical and Biological Sciences	41	Science technologies/technicians
LA&S Social Sciences	19	Family and consumer sciences/health sciences
LA&S Social Sciences	42	Psychology
LA&S Social Sciences	45	Social sciences

Other Professional	1	Agriculture, agriculture operations, and related services
Other Professional	3	Natural resources and conservation
Other Professional	4	Architecture and related services
Other Professional	10	Communications technologies/technicians and support services
Other Professional	11	Computer and information sciences and support services
Other Professional	12	Personal and culinary services
Other Professional	13	Education
Other Professional	22	Legal professions and studies
Other Professional	25	Library science
Other Professional	29	Military technologies
Other Professional	31	Parks, recreation, leisure, and fitness studies
Other Professional	43	Security and protective services
Other Professional	44	Public administration and social service professions
Other Professional	46	Construction trades
Other Professional	47	Mechanic and repair technologies/technicians
Other Professional	48	Precision production
Other Professional	49	Transportation and materials moving
Other Professional	50	Visual and performing arts
Other Professional	51	Health professions and related clinical sciences
Other Professional	28	Military Science, leadership and operational art
Other Professional	60	Health professions residency/fellowship programs