Finding Topics in Enrollment Data

Benjamin Motz, Thomas Busey, Martin Rickert, and David Landy
Department of Psychological and Brain Sciences
Indiana University, Bloomington, Indiana
bmotz@indiana.edu, busey@indiana.edu, rickertm@indiana.edu, dlandy@indiana.edu

ABSTRACT

Analyses of student data in post-secondary education should be sensitive to the fact that there are many different topics of study. These different areas will interest different kinds of students, and entail different experiences and learning activities. However, it can be challenging to identify the distinct academic themes that students might pursue in higher education, where students commonly have the freedom to sample from thousands of courses in dozens of degree programs. In this paper, we describe the use of topic modeling to identify distinct themes of study and classify students according their observed course enrollments, and present possible applications of this technique for the broader field of educational data mining.

Keywords

Higher education, Learning analytics, Topic modeling, Student data, Course enrollment, Transcripts, Educational data mining

1. INTRODUCTION

At any large educational institution, the student population is likely to be rather heterogeneous. One prominent source of variability is the range of academic topics available to students, reflected in the breadth of courses available to them, and the diverse requirements of numerous degree and pre-professional programs. This variability can make it challenging to analytically characterize the behaviors of students (e.g., graduation rates, engagement, grades), because students with different academic interests will have different experiences of higher education.

But nevertheless, some students will share similar experiences. There may be ways of parceling the diverse population to identify distinct groups of students whose academic interests are relatively homogenous within-groups, but differ between-groups. A simplistic strategy would divide students by major; but it may be desirable to identify groups before students have explicitly declared a degree program. In addition, these data may be unreliable as students often switch majors, and majors may artificially segregate students with generally similar interests and behaviors (e.g., students majoring in Chemistry, Biochemistry, or Biotechnology are probably quite similar). And dividing students by major may also be circular: Do majors describe student interests or merely describe the administrative landscape of degree programs? Instead of segmenting by major, the analytical challenge is to identify distinct areas of study directly from student course enrollments, where students assigned to each area have similar academic interests, experiences, and behaviors.

In educational data mining, clustering is the most commonly-applied method for classifying students [3, 19]. Vellido et al. [21] recently summarized a range of cluster analysis techniques, reviewed their applications to educational data mining, and compiled a bibliography of published studies that pursued such applications, particularly in e-learning environments.

It is conceivable that one might perform cluster analysis with course enrollments, as recorded on student transcripts. Each individual course might be treated as a single dimension in a high-dimensional space (e.g., one dimension for every course), and a transcript would be a single point in this space (with enrollments, 0 or 1, along each dimension). But there are major problems with this approach, particularly the "curse of dimensionality." In high-dimensional space, the data become sparse, and distances between individual points become almost equal, often yielding meaningless clustering results [1]. Recently-developed algorithms and distance metrics may improve the performance of high-dimensional clustering [13], but in this paper, we propose an entirely different approach that is better-suited to this particular analytical case.

We propose the use of topic modeling to address the challenge of classifying student transcripts. Topic modeling is commonly used for natural language processing applications (e.g., [10]) to identify abstract themes, or "topics," that exist in a collection of documents by analyzing the statistical distribution of words across these documents (for a review, see [5]). For our purposes, each document is a student transcript and each word is a course enrollment.

2. METHOD AND CONSIDERATIONS

Topic modeling is an umbrella term for a handful of methods that accomplish similar goals. The most popular method, and the one that we recommend for this application, is Latent Dirichlet allocation (LDA; [6]). Intuitively, in its simplest form, the approach initializes by assigning every token (each word in each document, or in the present analysis, each course in each transcript) to a random topic, and then repetitively iterates through the tokens, updating topic assignments in order to reduce the occurrence of individual words across multiple topics, while still preserving the contexts of words that tend to appear together within individual documents. Ultimately the method will produce a model of topics, a description of the words that tend to occur together.

Topic modeling has many advantages for the purposes of classifying academic topics:

- Rather than unequivocally classifying documents to topics, LDA assigns each word to a topic, producing a distribution of topic assignments for each document, and a probabilistic distribution of words for each topic (similar to soft-clustering approaches [18]). For educational data mining purposes, this is advantageous because a single course might occur in several topics with different probabilities, depending on the course's context in different students' transcripts (e.g., the course "Elementary Calculus" might be

differentially-predictive for students interested in Biology vs. Computer Science). In order to produce a coarse classification of students, we can simply assign students to the topic that appears most frequently in their transcripts.

- LDA is insensitive to the order of words (although it needn't be; [22]), which therefore allows the analysis of courses that are not taken in a strict sequence
- LDA is also generally insensitive to the length of documents (for documents with a small number of words, the prior probability of the topic distribution across all documents has a larger effect), allowing the analysis of incomplete transcripts.
- LDA can be parameterized in a way that tends to yield similarly-sized topics, minimizing the possibility of disproportionately large or small groups (which tends to occur with clustering).

Many software implementations of topic modeling methods are available[1]; we selected MALLET [15], which is open-source and has a large community of active users.

2.1 Case Data and Preprocessing

We identified all full-time students enrolled in a baccalaureate program at Indiana University Bloomington who initially became new or transfer students between 1995 and 2009. Students who did not complete any courses at our local institution were excluded. We then constructed course identifiers (which served as "words") by concatenating the academic program code, the course inventory, and the course number for every enrolled course appearing on the students' transcripts[2], irrespective of earned grade. There were 9,566 unique courses, 86,808 unique students, and students had an average of 29.3 courses listed on each transcript.

2.2 Modeling Topics

In traditional lexical analyses, documents contain words, and the topic model probabilistically associates the latent topics to each document through the words that it contains. In our analysis, courses were treated as words, and each student was represented by a document, the student transcript, that contained a collection of all courses taken by the student as part of their undergraduate education. Thus we are able to associate both students and courses with the discovered latent topics.

In its most basic form, the only parameter that needs to be supplied when modeling topics is the desired number of topics (see Section 2.3, below).

While there are various ways to visualize extracted topics (e.g., [8]), perhaps the easiest way to summarize a topic model is to present the words that are most probable in a particular topic for a set of representative topics, sometimes called "topic keys." A summary of a topic model on our transcript dataset, describing 6 of 24 topics, is shown in Figure 1. An interpretive gloss (in quotations) is provided above the ten most probable courses for that topic, listed in descending order (labels for the full set of 24 topics are shown on the right side of Figure 3, which is described later in this article). Students were assigned to the topic that appeared most frequently on their transcript, and the percent of the full student cohort that was assigned to each topic is also provided next to the topic label (if students had been evenly-allocated to the 24 topics, there would be 4.2% of the cohort in each topic). At face value, the algorithm did an impressive job of allocating the nearly 10,000 courses into distinct academic topics, particularly when considering that the model is entirely unsupervised. These topics were identified simply by analyzing the contextual trends in students' transcripts.

Importantly, one should not assume that these topics would emerge if the same analysis were performed on a different dataset. Different institutions have different academic programs and requirements, and different enrollment patterns. The current results are presented as a methodological case study, not as results that should be expected to generalize.

Some predictable patterns emerge in the current dataset, such as topics that clearly reflect the curriculum of popular majors, including "Business" and "Psychology." Other topics seem to slice across traditional academic silos, such as "Language Education," or the "Government" topic, which features courses from the Department of History and also from the Department of Political Science, even though neither department's undergraduate degree program explicitly requires courses from the other. Yet other topics seem to identify subgroups within a field, such as "Health Science" and "Basic Science," which segregates premedical interests from more basic science coursework, even though many of the students assigned to these topics are pursuing the same undergraduate degrees (e.g., Biology).

An essential caveat with topic modeling is that the algorithm yields a description of latent topics (in this case, themes of undergraduate study), but does not describe the behavior of any individual student. The topics can be used to partition students into distinct groups (e.g., by assigning a student to the most frequent topic in their transcript), but the topics themselves do not characterize individual students with any specificity. Rather, they describe statistically separable academic themes. When interpreting topic models, it is important to remember that the topics characterize themes of study, but individual student behaviors may be more complex, as any student's transcript would be expected to contain courses from multiple topics with different frequencies.

2.3 The Number of Topics

Topic modeling requires that the analyst specify the appropriate number of topics (T) in the dataset. For some applications, T may be a known quantity; perhaps there are predetermined academic tracks that any student might pursue, and the goal is simply to characterize the enrollments that co-occur with these known topics. However, for most analyses, the number of topics is unknown, and the analyst must determine the appropriate number of topics to account for information in the dataset, according to the desired granularity of the analysis. There are methods for automatically inferring optimal values for T according to model performance measures [2, 16], but we preferred a more exploratory approach. Specifically, we extracted topics for a range of desirable values for T, evaluated these models using hold-out data, and then selected a value T to maximize likelihood while

[1] David Mimno maintains a reference list including software tools for topic modeling: http://mimno.infosci.cornell.edu/topics.html

[2] When dealing with this type of codified data, with course identifiers that may include numbers and punctuation symbols, it is important to specify the structure of the words as a regular expression in the analysis software, so that the documents are parsed appropriately.

"GOVERNMENT" (3.2% of cohort)	*"BUSINESS"* (12.5% of cohort)	*"PSYCHOLOGY"* (4.6% of cohort)
POLSY200: Contemporary Political Topics	**BUSX201**: Technology & Business Analysis	**PSYP324**: Abnormal Psychology
POLSY103: Introduction to American Politics	**BUSX420**: Business Career Planning	**PSYK300**: Statist cal Techniques
HISTH105: American History I	**BUSA202**: Intro to Managerial Accounting	**PSYP102**: Introductory Psychology II
HISTA300: Issues in United States History	**BUSX220**: Career Perspectives	**PSYP199**: Career Planning for Psychology
COASW333: Intensive Writing	**BUSZ302**: Managing & Behavior in Organizations	**PSYP151**: Introductory Psychology I for Majors
POLSY109: Introduction to International Relations	**BUSF370**: Integrated Business - Finance	**PSYP335**: Cognitive Psychology
HISTB300: Issues in Western European History	**ECONE370**: Statist cal Analysis for Business	**PSYP320**: Social Psychology
HISTH106: American History II	**BUSX204**: Business Communication	**PSYP211**: Methods in Experimental Psychology
POLSY100: American Political Controversies	**BUSP370**: Integrated Business - Operations	**PSYP315**: Developmental Psychology
HISTJ300: Seminar in History	**BUSJ370**: Integrated Business - Strategy	**PSYP152**: Introductory Psychology II for Majors

"LANGUAGE EDUCATION" (4.1% of cohort)	*"HEALTH SCIENCE"* (4% of cohort)	*"BASIC SCIENCE"* (7.8% of cohort)
HISPS275: Intro to Hispanic Culture	**ANATA215**: Basic Human Anatomy	**CHEMC117**: Principles of Chemistry II
HISPS310: Intro to Hispanic Linguistics	**MSCIM131**: Disease and the Human Body	**BIOLL112**: Biological Mechanisms
COASW333: Intensive Writing	**SOCS100**: Introduction to Sociology	**BIOLL113**: Biology Laboratory
ENGL202: Literary Interpretation	**PSYP101**: Introductory Psychology I	**BIOLL111**: Evolution & Diversity
EDUCM300: Teaching in Pluralistic Society	**PHSLP215**: Basic Human Physiology	**PHYSP201**: General Physics I
ENGW203: Creative Writing	**CHEMC101**: Elementary Chemistry	**CHEMC341**: Organic Chemistry I
HISPS331: The Hispanic World	**ENGW131**: Elementary Composition	**BIOLL211**: Molecular Biology
ENGW103: Introductory Creative Writing	**PSYP102**: Introductory Psychology II	**PHYSP202**: General Physics II
HISPS317: Spanish Conversation & Diction	**CLASC209**: Medical Terms from Greek & Latin	**CHEMC105**: Principles of Chemistry I
EDUCH340: Education & American Culture	**HPERH160**: First Aid and Emergency Care	**CHEMC342**: Organic Chemistry II

Figure 1: Top 10 most probable courses for 6 representative topics (of 24 total). These results are provided for illustration purposes, and topics will likely vary between institutions.

balancing the risk of overfitting with too many topics for the intended analysis.

2.3.1 Evaluating using hold-out data

Current guidelines for topic model evaluation were proposed by Wallach et al. [23]. This approach seeks to estimate the probability of hold-out data, given a particular topic model. The first step was to segregate the documents into a training set (random 90% of documents) and a hold-out set (remaining 10%). Topic models were then developed on the training set for a range of reasonable values for T (we used 2 to 100, in steps of 2). For each of these models, we estimated a log likelihood (LL) value for every document in the hold-out set, given that particular model. LL is a negative number, and intuitively, it provides an estimate of how unexpected the hold-out document's collection of words would be, considering the model's configuration of topics; LL values closer to zero indicate that the model was better, as any given document was less unexpected. Because the evaluation process is non-deterministic, we repeated the evaluation process 10 times, averaged the LL for each document across these 10 runs to obtain a more stable probability estimate, and then summed the averaged LL across all hold-out documents. This summed, averaged LL was finally divided by the total number of tokens in the hold-out set to produce a normalized-LL estimate; in many studies, this normalized value ranges between -10 and -6. The solid black line in Figure 2A illustrates the averaged LL/token for topic models on student transcripts, for a range of T.

2.3.2 Finding the inflection point

Ultimately, the desired number of topics should be determined through a combination of statistical analysis, general insights into the structure of the data, and consideration of the purpose of the model. Increasing the number of topics will generally improve the LL/token estimates, but above a certain point, these incremental improvements are trivial. For the current application, we sought to determine the fewest number of topics, such that additional topics would yield minimal improvements to the quality of the model. To find this inflection point, we fit a piecewise linear regression model on the LL/token estimates, seeking the value T* that minimized the root mean square error of the linear trends, $T<T^*$ and $T>T^*$. As illustrated by the dashed lines in Figure 2A, this point was $T^*=24$ topics; importantly, the topic keys (six are shown in Figure 1) made intuitive sense, and yielded an insightful model for this analysis.

2.3.3 Stop lists and frequent courses

There is a convention in topic modeling to remove high-frequency tokens from the training dataset. When modeling topics in linguistic corpora, this pre-processing step is intended to filter words that do not contain meaning (such as "the", "a", "of", etc.) and would add unnecessary noise to the identification of topics. These excluded tokens are called a stop list.

Along these lines, it may be useful to exclude high frequency courses when modeling academic topics. Courses that appear on a large proportion of student transcripts (general education courses, high-enrollment prerequisites, etc.) may be practically meaningless for the purposes of classifying student interests. However, there has been relatively little empirical work evaluating the use of stop lists when modeling topics. In information retrieval algorithms more generally, Manning et al. [14] note that the cost of including high-frequency tokens (in computational time) is minimal, and that the recent trend is to use smaller stop lists, if any at all.

We approached this issue as an empirical question (i.e., a sensitivity analysis): Will the use of a stop list affect model

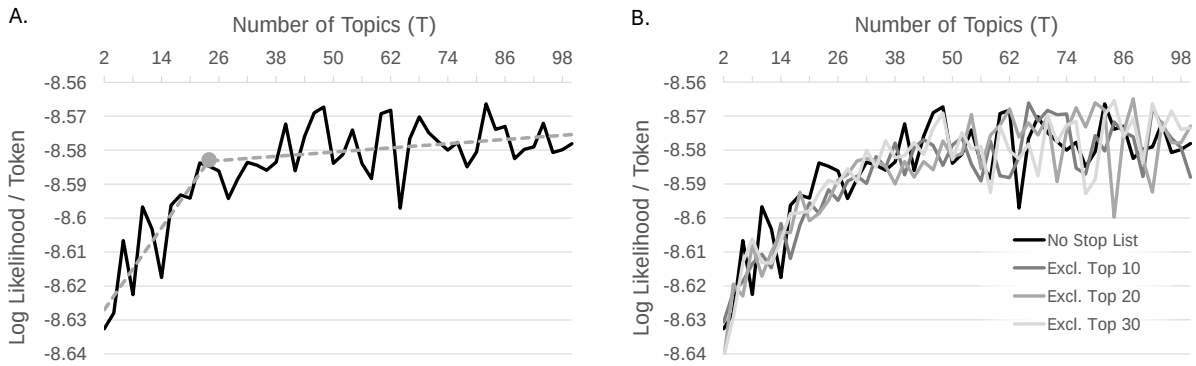

Figure 2: Model performance for a range of values for T (A), and comparing stop lists (B).

performance? The previously-described modeling analysis was repeated three additional times, filtering the top-10, top-20, and top-30 most-frequent courses (the 10th-most-frequent course appears in 19.7% of transcripts, the 20th- appears in 13.0%, and the 30th- appears in 10.5%). For reference, the highest-frequency course at our institution (Elementary Composition) appeared on 46.3% of student transcripts. The log-likelihood of these models, with T ranging from 2-100, is illustrated in Figure 2B.

There was no clear effect of including a stop list on the model's performance, irrespective of the number of topics or the number of words on the stop list. It may be that, at our institution, the highest-enrollment courses impart a minor amount of information about the thematic structure of a student's enrollments, but not enough to substantively improve or impair the model's performance. For most natural language processing applications, stop lists typically aim to filter words that occur in a very large proportion (e.g. 85%) of documents, so even though they're relatively popular at an institutional level, the highest-enrollment courses (appearing in less than 50% of transcripts) simply might not rise to the level of frequency that would merit their exclusion. These effects may vary across institutions, but without a clear separation in model performance, we suggest including all relevant courses in the analysis, and do not advocate the use of a stop list.

3. EXAMPLE APPLICATIONS

In a rapidly growing field such as educational data mining, it is difficult to anticipate the full range of uses of a relatively new method, such as topic modeling, or any analytical technique. The following three examples are only intended to help illustrate, at a very high level, the general value of identifying academic topics, and the wide range of potential applications.

3.1 Sandwich Estimator

In educational data mining, researchers commonly try to predict the effect of one variable on another variable, such as the effect of an automated flagging system on graduation rates. Common modeling approaches (such as ordinary least squares regression) typically carry the assumption that each observation is independent from the others. But in higher education, this is a weak assumption. Different students are jointly exposed to the same classes, instructors, student groups, and graduation requirements, and moreover, they might be expected to communicate with each other about these experiences, and influence each other's behaviors. Although violating the independence assumption will not affect the point estimate (i.e., magnitude) of a regression parameter, it can significantly change the interval estimate (i.e., precision) of the parameter, which in turn, changes the probability of making a Type I or Type II error.

One solution to this issue would be to fit multilevel random effects models to account for the non-independence of observations and the cross-classified data structure (with students not strictly nested within grouping variables). However, this would be an absurdly complex model, with every course, semester, instructor, etc. included as a crossed random effect; we feel that such an effort is impractical.

But considering that topic models are derived from patterns in course enrollments, the topic classifications can be used as a grouping variable that will account for the non-independence of student experiences and produce corrected (i.e., sandwich) estimates of the standard errors for the model parameters [24]. By classifying students according to the most frequent topic in their transcript, we are able to identify subgroups of students such that their coursework and learning activities are correlated within-groups, and are independent between-groups. In our enrollment data, using T=24 and a binary response variable indexing graduation within four years of initial enrollment, we obtained an estimated intraclass correlation of 0.254. This suggests that about a quarter of the variance of within-class 4-year graduation rates are explained by topic assignment, heteroscedasticity that can be easily corrected in regression models.

3.2 The Alignment of Programs and Topics

There are latent interests held by students that influence the courses they select. Sometimes these enrollment choices are codified in degree requirements or prerequisites, or even by external forces (such as medical school requirements). However, as mentioned in the Introduction, the nuanced boundaries that delineate different degrees do not necessarily provide a fair representation of the different topical interests that might motivate students' course selections. This relative alignment of degree programs with students' interests can be investigated using topic modeling.

For example, in discussions of such academic restructuring, it is often suggested that departments with similar interests should

merge or combine resources [11]. The current topic modeling approach may reveal different degree programs that are jointly represented by a single topic, and these might be candidates for this type of restructuring. At our institution, our analysis reveals notable overlap between History and Political Science, and there may be administrative synergies between these programs.

In contrast, there may be topics that integrate courses from different departments in stable ways that are unaccommodated by any degree. Beyond mere overlap between programs, students may be sampling courses from multiple programs to construct "hidden majors," academic chimeras that may not exist as formalized degree programs, but that integrate diverse coursework to create stable topics of interest. For example, course enrollments at our institution revealed a "Media Studies" topic of study that was not accommodated by any single major; it blended courses from Communications, Comparative Literature, Sociology, and more. Our discovery of this topic provided support for our institution's recent initiative to create a new Media School.

And topic modeling might also be used to reveal separable sub-disciplines within a single degree program. Even within an individual major (such as Psychology) there is ample opportunity for students to focus on subdisciplines (such as counseling, human factors, child development, behavioral research, etc.). Just as topic modeling can reveal latent academic themes in an entire university's course catalogue, it can also be applied to a single academic division or program, to evaluate the thematic structure within a single unit. At our institution, by modeling topics from the enrollments of recent graduates in Psychology, we identified themes related to law, medicine, and social psychology. These have enabled us to tailor career planning events, course offerings, and advising materials to the specific interests of our students.

3.3 Transitions and Outcomes

At colleges with flexible degree requirements, undergraduate students typically undergo an academic metamorphosis, enrolling in first-year general survey courses to eventually enrolling in specialized advanced courses [4]. This transition, from the nonspecific enrollment behaviors of freshmen to the niche upper-division coursework of soon-to-be graduates, is an area that has begun to receive increasing attention in higher education research, particularly in efforts to improve retention and eliminate boundaries and bottlenecks to STEM fields. By characterizing the various transitional paths from first year study to subsequent disciplinary specialization (and the success rates associated with these paths), institutions would be better-equipped to test hypotheses about pipeline issues, and to develop effective advising strategies and interventions for beginning students [12].

Along these lines, topic modeling might be applied to first year enrollments, in order to identify the broad thematic enrollment trends of beginning students. And then we might draw the paths from first year topics to the topics derived from full transcripts, to illustrate how students transition from initial coursework to eventual specialization in an established topic. This analysis is described below, and illustrated in Figure 3.

3.3.1 Visualizing Paths from First Year Topics

The previously-described topic modeling approach was performed on the same set of students, but we limited their transcripts to only include courses that were credited during the student's first year of study. There were 3,330 distinct courses on these truncated transcripts, and on average, there were 9.0 courses per student during this first year. After evaluating models for a range of values for T we found an inflection point at 5 topics, and determined that this provided the appropriate balance of model

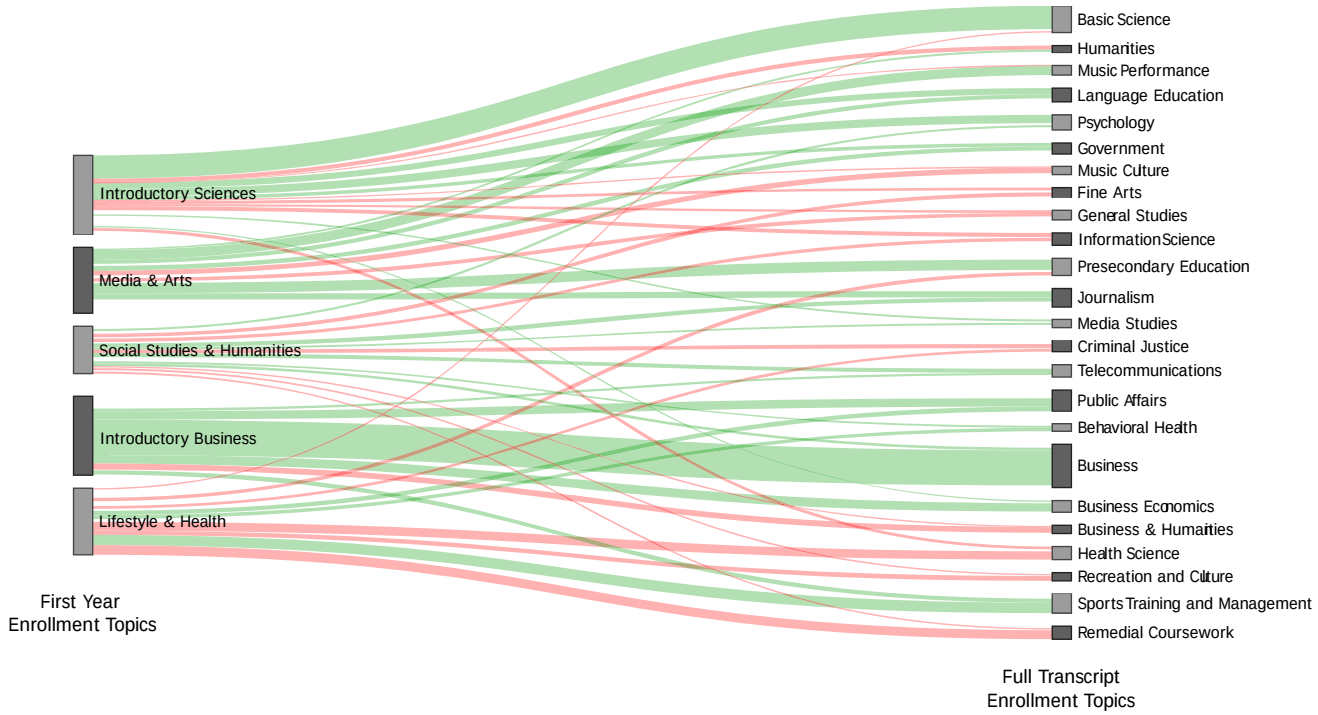

Figure 3: Diagram of transitions between first year and full transcript enrollment topics.

performance and face validity. And, as previously discussed, students were assigned to the topic that appeared most frequently on their first-year transcripts. For every student in our sample, we now had one topic assignment (1-5) for their first year enrollments, and another topic assignment (1-24) for their complete transcript.

For each student, we also identified whether they received a baccalaureate degree up to 4 years after their initial enrollment. For this sample of students during this time frame, the overall 4-year graduation rate was 51.2%. Of course, many more students will eventually receive a degree after their 4th year, but this 4-year rate is relevant for institutional benchmarking purposes.

Figure 3 was produced using the Sankey diagram plugin for D3 [7]. Paths are green or red if the students within that path have a 4-year graduation rate above or below 51.2%, respectively. The relative size of the gray boxes and colored paths represent the number of students assigned to the topic or transition. To make this diagram more readable, we have only included the two largest entry paths for each of the full-transcript enrollment topics.

3.3.2 Interpreting Topic Transitions

One of the immediate observations from this analysis is that a student's first year enrollments tend to be reasonably predictive of the themes of study where the student may ultimately arrive at the end of their career—the flat paths tend to be thicker than the sloped cross-cutting paths. Initially one might attribute this to the fact that first-year enrollments are included in the full-enrollment transcripts. This artifact may play a role, however, looking back to Figure 1, an important observation is that the most probable courses for full-enrollment topics (those at the top of the list) are commonly 200-level courses, typically beyond the first year (100-level) introductory sequence. We observe that students' first year enrollments are not dissociated from their future enrollment tendencies.

This observation might suggest that students who transition to a relatively unrelated topic after their first year would be at a disadvantage to graduate in 4 years. But the data seem to suggest otherwise: that some full-transcript topics simply have lower 4-year graduation rates than others, regardless of whether the students followed a straight thematic trajectory, or seemed to originate from an untraditional first-year topic. For example, students who ultimately study "Recreation and Culture" have lower graduation rates, regardless of whether they began college by studying "Lifestyle and Health" (a structurally similar theme) or "Social Studies and Humanities" (a relatively distant theme).

These exploratory analyses and interpretations have their limitations, and the hypotheses derived from a visualization like this should receive further scrutiny on the local level. As discussed previously, our topic models describe abstract themes of study, and do not characterize students per se. The students whom we've identified as being members of a theme (because the theme appears most commonly on their transcript) may have other similarities, besides their course enrollments (e.g., third variables such as family expectations, cultural values), that contribute to their graduation rates or enrollment behaviors more directly than their coursework. Nevertheless, being able to easily visualize the flow of the entire student body (albeit indirectly) across the academic landscape can serve useful purposes toward understanding the inflow into a particular area, and ultimately developing better-informed advising strategies.

4. CONCLUSION

Blanket generalizations that treat an institution's "students" as a single group are likely to be either ineffectively vague, or not applicable to all members of the student population [20]. In the classroom, post-secondary instructors find value in knowing the differentiating characteristics of the students in their classes, and tailoring instruction to accommodate their unique attributes [17]. Data-driven interventions and analytical characterizations of student behaviors should also be sensitive to the differences between students. In this paper, we've described an effective method for identifying one prominent source of variability: students' academic interests. By applying topic modeling to student transcripts, we are able to identify separable topics of study at our institution, and these topics can be further used to roughly classify students into distinct groups that feature similar enrollment behaviors.

Considering that it was originally developed as a natural language processing tool, topic modeling has well-documented applications to educational data mining in the analysis of student discourse (e.g., in a discussion forum; [9]) or written coursework, but it could also be applied to any form of unstructured categorical data at the university, such as LMS web traffic, library checkouts, or even meal point expenditures. Similarly, we believe that topic modeling is a straightforward and uniquely suitable method for identifying patterns in raw enrollment data.

5. ACKNOWLEDGMENTS

This work was supported by a grant from the Bay View Alliance. Special thanks to Mark Steyvers for his expertise and guidance, as well as Linda Shepard, Stefano Fiorini, and Mike Sauer for access to and assistance with the institutional data used in this analysis.

6. REFERENCES

[1] Aggarwal, C. C., Hinneburg, A., and Keim, D. A. 2001. On the surprising behavior of distance metrics in high dimensional spaces. *ICDT, Lect. Notes Comput. Sc.*, 1973 (Oct. 2001), 420-434. DOI= http://dx.doi.org/10.1007/3-540-44503-X_27

[2] Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. 2010. On finding the natural number of topics with Latent Dirichlet allocation: Some observations. In *Advances in Knowledge Discovery and Data Mining*, Springer, Berlin. 391-402.

[3] Baker, R. S. J. d. 2010. Data mining for education. In *International Encyclopedia of Education*, B. McGaw, P. Peterson, E. Baker, Eds. Elsevier, Oxford. 112-118.

[4] Babad, E., Darley, J. M., Kaplowitz, H. 1999. Developmental aspects in students' course selection. *J Educ. Psychol.*, 91, 1 (Mar. 1999), 157-168. DOI= http://dx.doi.org/10.1037/0022-0663.91.1.157

[5] Blei, D. M. 2012. Probabilistic topic models. *Commun. ACM*, 55, 4, 77-84. DOI= http://doi.acm.org/10.1145/2133806.2133826

[6] Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3 (Jan. 2003), 993-1022.

[7] Bostock, M., Ogievetsky, V., and Heer, J. 2011. D^3 Data-Driven Documents. *IEEE T. Vis. Comput. Gr.*, 17, 12 (Dec. 2011), 2301-2309. DOI= http://dx.doi.org/10.1109/TVCG.2011.185

[8] Chaney, A. J. B. and Blei, D. M. 2012. Visualizing topic models. In *International AAAI Conference on Web and Social Media* (ICWSM '12). AAAI Press.

[9] Ezen-Can, A., Boyer, K. E., Kellogg, S., and Booth, S. 2015. 2015. Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (LAK '15). ACM, New York, NY, USA, 146-150. DOI= http://dx.doi.org/10.1145/2723576.2723589

[10] Griffiths, T. L. and Steyvers, M. 2004. Finding scientific topics. *P. Natl. Acad. Sci. USA*, 101, suppl. 1 (Apr. 2004), 5228-5235. DOI= http://dx.doi.org/ 10.1073/pnas.0307752101

[11] Gumport, P. J. 2000. Academic restructuring: Organizational change and institutional imperatives. *High. Educ.*, 39, 1 (Jan. 2000), 67-91. DOI= http://dx.doi.org/ 10.1023/A:1003859026301

[12] Heileman, G. L., Babbitt, T. H., and Abdallah, C. T. 2015. Visualizing student flows: Busting myths about student movement and success. *Change: The Mag. of High. Educ.*, 47, 3 (Jun. 2015), 30-39. DOI= http://dx.doi.org/10.1080/00091383.2015.1031620

[13] Kriegel, H. P., Kröger, P., and Zimek, A. 2009. Clustering high-dimensional data. *ACM Trans. Knowl. Discov. Data*, 3, 1, Article 1 (Mar. 2009). DOI= http://dx.doi.org/10.1145/1497577.1497578

[14] Manning, C. D., Raghavan, P., and Schütze, H. 2008. *Introduction to information retrieval*. Cambridge University Press, Cambridge.

[15] McCallum, A. K. 2002. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu

[16] Mimno, D. and Blei, D. 2011. Bayesian checking for topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP '11). Association for Computational Linguistics, Stroudsburg, PA, USA, 227-237.

[17] Motz, B. A., Teague, J. A., Shepard, L. L. 2015. Know thy students: Providing aggregate student data to instructors. *EDUCAUSE Review Online* (Mar. 2015).

[18] Peters, G., Crespo, F., Lingras, P., and Weber, R. 2013. Soft clustering – Fuzzy and rough approaches and their extensions and derivatives. *Int. J. Approx. Reason.*, 24, 2, 307-322. DOI=https://doi.org/10.1016/j.ijar.2012.10.003

[19] Romero, C. and Ventura, S. 2010. Educational data mining: A review of the state-of-the-art. *IEEE T. Syst. Man Cy. C*, 40, 6 (Nov. 2010), 601-618. DOI= http://dx.doi.org/10.1109/TSMCC.2010.2053532

[20] Quaye, S. J. and Harper, S. R. 2014. *Student engagement in higher education: Theoretical perspectives and practical approaches for diverse populations*. Routledge, New York.

[21] Vellido, A., Castro, F., and Nebot, A. 2010. Clustering educational data. In *Handbook of Educational Data Mining*, C. Romero, S. Ventura, M. Pechenizkiy, R. S. J. d. Baker, Eds. CRC Press, Boca Raton, FL. 75-92.

[22] Wallach, H. M. 2006. Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning* (ICML '06). ACM, New York, NY, 977-984. DOI=http://dx.doi.org/10.1145/1143844.1143967

[23] Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. 2009. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning* (ICML '09). ACM, New York, NY, 1105-1112. DOI= http://dx.doi.org/10.1145/1553374.1553515

[24] Williams, R. L. 2000. A note on robust variance estimation for cluster-correlated data. *Biometrics*, 56, 2 (Jun. 2000), 645-646. DOI= http://dx.doi.org/10.1111/j.0006-341X.2000.00645