Via: Illuminating Undergraduate Academic Pathways at Scale

Geoffrey Angus
Stanford University
Stanford, USA
gangus@stanford.edu

Richard Diehl Martinez
Stanford University
Stanford, USA
rdm@stanford.edu

Mitchell Stevens
Stanford University
Stanford, USA
stevens4@stanford.edu

Andreas Paepcke
Stanford University
Stanford, USA
paepcke@cs.stanford.edu

ABSTRACT

The processes through which course selections accumulate into college pathways in US higher education is poorly instrumented for observation at scale. We offer an analytic toolkit, called *Via*, which transforms commonly available enrollment data into formal graphs that are amenable to interactive visualizations and computational exploration. The tool is intended for a variety of stakeholders: college students, instructors, and administrators. We explain the procedures required to project the enrollment records onto graphs, and then demonstrate the toolkit utilizing eighteen years of enrollment data at a large private research university. We show that resulting findings complement prior research on higher education.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

Course Sequences; Academic Pathways; Graph Visualization

INTRODUCTION

US higher education is unique in the world in the extent to which schools expect undergraduates to explore a variety of courses before committing to a field of study. In contrast with virtually all other national postsecondary systems, in which students enter schools and programs with relatively structured curricula, US undergraduates are encouraged to explore a variety of academic options through an iterative course search and selection process. We call a student's eventual sequence of course elections a *pathway*. Pathways are notoriously poorly instrumented for observation by students, educators, administrators or researchers [10].

L@S'19, June 24–25, 2019, Chicago, CA, USA

ISBN 123-4567-24-567/08/06…$15.00

DOI: http://dx.doi.org/10.475/123_4

While enriching when successful, the contingencies of course elections that accumulate into pathways is poorly understood and are almost always fateful. Ethnographic research suggests that students tend to select courses based on partial, poorly integrated information, in light of logistical constraints and personal preferences that have little directly to do with academics [30, 36, 37]. Absent conscientious design and signposting, students can easily spend time, credit hours, and tuition accumulating courses that do not lead efficiently to majors and completion [2]. In addition to students, many other academic stakeholders could benefit from better information about college pathways. The work and priorities of instructors, department chairs, deans, and budget officers are implicated in the relative clarity of pathways and the efficiency with which courses can be sequenced to degree completion.

Fortunately the information necessary to observe pathways systematically and at scale is present in all colleges and universities in the form of academic transcripts. Transcripts are the official records documenting courses accumulated by each student as he or she makes academic progress. Yet transcripts typically are housed in tables of databases to which few have access, and in their "raw" form exceptionally opaque to interpretation. Most schools have specialized staff who generate specifically requested reports for individuals in high level academic positions[1], but these personnel typically cannot interact directly with all the parties in need of insight on pathways at varying levels of detail.

We report here on *Via*, an analytic toolkit we have built to observe and understand undergraduate pathways utilizing deidentified transcript data held by a large private research university. Our approach is to provide a zoomable, investigative instrument for large-scale, qualitative and quantitative investigations of pathways. Built on graph theory, *Via* provides a visual interface for observing the course sequences embedded in tens of thousands of transcripts. In addition, *Via*'s grounding

[1]We acknowledge here our own version of such a unit, which has provided us with numerous insights into both the information needs, and data semantics at our institution.

in graphs allows us to bring associated mathematical computations to bear on the problem of pathway evolution.

Graph approaches have been applied to a wide variety of other tasks, such as detecting communities [18], collecting materials for survey articles [26], the augmentation of collaborative recommendation records [23], predicting future collaborations between scholars [29], and suggesting drug interactions [46]. Our primary contribution in this paper is to apply graph approaches to the sequencing of academic coursework. Because our analytic strategy relies on data of a sort held by every legally recognized US college and university, it is amenable to application throughout the US postsecondary sector.

In our first approximation here, we convert transcript information describing course enrollments at our case school to directed graphs, with nodes modeling courses, links modeling sequential enrollments, and link weights modeling conditional probabilities of enrolling in a particular course before another one. We partition the graphs by the departments to which each course is officially associated by the university registrar. An existing graphing tool [40] exposes more or less information, depending on chosen zoom levels.

After related work and a brief introduction to our dataset, Section 4 introduces how we construct the node and edge relationships in the *Via* network. Section 5 then demonstrates the visual component of the model. Section 6 highlights diverse use-cases of our model to address questions of different academic stakeholders: students, instructors, and administrators. This final section demonstrates the wider applicability of our model, by comparing our results against well-studied phenomena in education research.

RELATED WORK

The dangers of choice overload are well documented. Research in marketing [8] and psychology [38, 24] has demonstrated both the stated preference for choice and improved factual success when choice is limited. In higher education, Bake et al. [4] find that simply reducing choice is not an acceptable answer. While students do not have strong reactions to increased guidance, they react negatively to a reduction in choice. Thus enriching guidance is one promising approach. The author of [43] shows that simulations and predictions based on student models can successfully identify points where advisors could intervene to improve graduation outcomes.

Community colleges and other broad-access schools are at least as affected by the tension between the benefits of choice and student confusion. Educational outcomes in community colleges are negatively affected by "chaotic enrollment patterns" [12, 3, 39]. A number of these institutions have explored options for improved support by providing "guided pathways" [25], but without the benefit of graph visualization and other computational-techniques. A few comprehensive universities have built carefully instrumented "early alert" and other advising tools utilizing institutional data [16], but the computational techniques underlie these tools remain opaque.

Throughout both the social and physical sciences, graphs are used to visualize, simplify and facilitate computational analysis of complex dynamic systems. Graphs have been applied to model a diverse set of networks such as food chains [21], the human genome [35] and ecological systems [17]. Within the social sciences, graphing methods have a long history of application in the study of systems-level phenomena [7]. The use of graphs within the social sciences was particularly spurred on by the insight that human societies could be structured like biological systems. The 19^{th} century French philosopher, Emile Durkheim, argued for instance that social regularities could be found in the structure of social environments in which they were embedded [14]. By studying systemic regularities it is possible to derive macro-level insights about the structure of many micro-level interactions.

Since the mid-1950s, graphs have been applied to model the flow of information in social and professional networks. As part of the MIT Group Networks Laboratory, Leavitt et al. observed how the structure of interpersonal relationships between groups of coworkers facilitated the spread of information throughout a team of colleagues [28]. More recently, similar methods have been applied to study how workplace professional networks influence the spread of information through company email chains [15]. The modeling of knowledge transmission parallels closely how graph-theoretic methods have been applied to understand social dynamics in the field of education. Studies of citation networks, for instance, describe how intellectual advances spread through academic space [5, 19].

A number of the challenges posed in representing citation networks, such as learning optimal edge weights, are directly applicable to the problem of modeling course sequences. Within citation networks, it is useful to modulate the edge strength between two papers in order to represent the relative influence of a cited paper. Batagelj proposes solutions to this problem by introducing SPC weights on each edge of the network to capture the incoming and outgoing "flow of information" for a given paper [5]. Hajra et al. also observe aging phenomena among papers in which the probability that a paper is cited decays with time at an exponential rate proportional to $t^{0.9}$ up to ten years after its publication date [20]. Course selection may display a similar exponential time-dependent probability. Moreover, just as citation networks provide a framework for modeling the flow of knowledge within academia, course sequence networks imply shared and prerequisite knowledge between courses.

Also analogous to our particular framework is recent work done in the field of representing social connections within massive open online courses (MOOCs). Large online education providers, such as *Coursera*, use thread messaging boards to facilitate collaboration between students. Within these forums, any student in a particular course can post a question or remark to start a thread of conversation. The data gathered from these educational messaging boards enables researchers to study the exchange of information between students [9], the influence of students on others' participation in the course [42] and more general social dynamics of a class.

Sinha et al., for instance, model each participant within a particular *Coursera* class as a node, and draw a directed edge

from a thread or sub-thread initiator to any of the people who engaged in that discussion [41]. By doing so, the authors are able to represent the flow of topics and engagement within the social dynamics of a certain course. Sinha et al. were able to analyze measures of degree and betweenness centrality on this model of student engagement, in order to understand the influence of discussion initiators on the overall collaboration of students in the class. The parallels are made clear if we interpret discussion topics as self-contained units of knowledge that are part of a course.

Similarly to our network construction, Sinha et al. use a gradient coloring of nodes to indicate its relative betweenness centrality. Zhu et al. also seek to model the engagement of students in discussion forums on a week-by-week basis using Exponential Graph Models (EGM) [45]. The use of EGMs allows the authors to model a student's participation for a given week by incorporating a student's performance for both the previous and subsequent weeks. While inspired by EGMs, our model does not directly include aggregate network statistics, in order to compute node and edge properties. Instead, we model the probability of two courses being taken sequentially given data in which the courses may have been taken several academic terms apart. Finally, *NetworkSeer* uses a similar modeling framework as in the previous papers but additionally models individual students' demographic information within the course discussion thread [44]. Although we do not have direct access to students' demographic data, our model uses a student's final major to study the differences in course-taking behavior between students.

The social dynamics of student participation in MOOCs and the spread of knowledge through citation networks are the closest parallel to modeling students' course-taking behavior in US universities. Courses can be observed as distinct units of information that share overlap and prerequisite knowledge with other courses. Academic publications and forum discussions are similarly self-contained units of knowledge that build upon and interact with other publications and discussion threads, respectively. Given a dearth of direct research in course sequence networks, our *Via* toolkit builds upon the modeling frameworks of citation graphs and MOOC social dynamic networks. In contrast, course embeddings [33] represent courses and enrollments as vectors, which are then clustered and otherwise manipulated.

DATA

The dataset we use to build our graph for mapping sequential relationships between courses was obtained directly from a large US private research university. This dataset contains the anonymized enrollment data of over 52,000 students who were enrolled at the university during any time between Fall 2000 and Fall 2018. Each of the roughly two million rows in this table consists of a unique student identifier, a course in which they enrolled, and the academic term for which they enrolled in the class. Our dataset also contains supplementary information, such as each student's major during time of enrollment in a course and each student's major upon graduation. Depending on the class of problems of interest, we filter along these additional values.

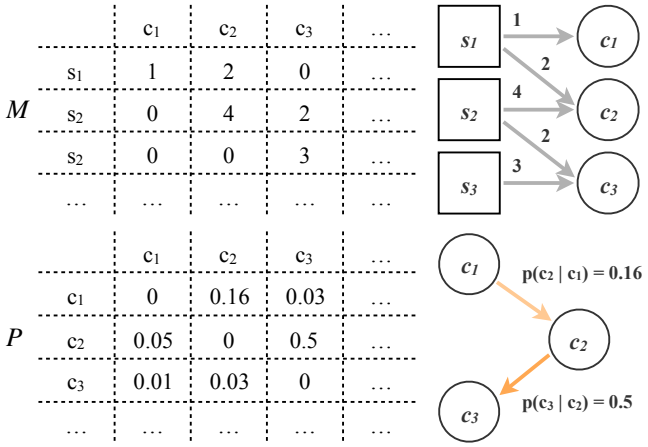

Figure 1. An overview of the projection created from enrollment data. Enrollment sequences are tallied and normalized to define a conditional probability of taking one course after another. Rows s_i describe students, while columns c_i correspond to course offerings.

Data Preprocessing

Data preprocessing involved the parsing of the requisite information made available through the dataset; for each student, all courses are paired with their terms of enrollment. In order to preserve the normal, sequential nature of courses across academic years, we omit the Summer Period from all experiments. Student metadata is not incorporated in the generation of sequence projection models; however, we do filter by student attributes in order to create subsets of the original dataset – a strategy that might also be deployed with demographic or other student characteristics.

METHODOLOGY

Our first challenge is to build an expressive, yet interpretable model of course-to-course interactions that results in the graph based representation of academic pathways. The representation must support the desired ability to discover student enrollment patterns over time and enrollment interactions between departments.

Towards this end, we compile student-to-course enrollment information into what we will call a course-to-course projection model, a network that represents courses as nodes and course relationships as weighted edges. Figure 1 illustrates the process. We begin with a matrix M representing student enrollment in courses over time. Indeed, we can interpret this matrix M as the adjacency matrix of a bipartite, labeled graph documenting the interaction of student nodes with course nodes. We end with a matrix P that defines a projection of the courses in M, where edge weights signify how strongly two courses are linked. This projection can be visualized as a network such as the one found in Figure 2, which maps all course interactions in a way that describes the aggregate behavior of student enrollment patterns.

Combined with an existing graph manipulation kit—in our case [40]—the graph described by P serves as our target toolkit for university stakeholders. We now show how we endow the construction process of P with flexibility that can tune the

resulting graph for visualizations that help answer a variety of questions.

As per the overview above, the first step is to prepare a sequence matrix. The second involves the calculation of the projection. The following sections merely formalize the overview.

Sequence Matrix Generation

Via receives data in the form of tabular student enrollment records. The input to the projection algorithm is a matrix M of shape $(|S|,|C|)$, where S is the set of all students and C is the set of all courses. A given entry M_{ij} represents the time when a student i enrolls in some course j. For example, an entry $M_{ij} = 1$ signifies that student i enrolled in course j during their first academic term, generally quarter or semester. An entry $M_{ij} = 0$ implies that i never enrolled in j. This thus implies that each row m_i represents the entire course enrollment history of some student i. We generate various forms of matrix M from the raw enrollment data by filtering on different student attributes. The nature of the filter depends on the questions being asked. For example, if we are only interested in those who majored in the Humanities, we would limit M to this student subset. Another filter one might apply at this step is year of enrollment, if only that time slice is of interest. The more filtering is applied to M the faster the subsequent processing, and visual interaction response.

Our data provides information on each enrollment's time, the enrolling student's major at the time of enrollment, and the student's final major. Depending on data availability, filters based on gender, status as underrepresented minority, or college entrance scores could be applied as well.

Graph Projection

Next, given the sequence matrix M, we generate the matrix P of shape $(|C|,|C|)$. This matrix represents the adjacency matrix for a one-mode projection of the bipartite network specified by matrix M. We weight each edge in P using conditional probabilities that describe how likely one is to take one course given another course. Thus, entry P_{ij} represents the conditional probability of taking course j given that one has taken course i. P aggregates students, losing some information, but gaining the probability of moving from one course to another.

Matrix P is thus a representation of how even non-adjacent course nodes in M interact, based on the nature of student enrollment. We use this matrix P as the basis of all calculations and visualizations. The parameters for the conditional probabilities are fit based on counts determined in the calculation of intermediary matrix $\tilde{P}$ of shape $(|C|,|C|)$. We calculate each entry $\tilde{P}_{ij}$ by accumulating the occurrences of course i taken at some point before course j across all students in set S:

$$\tilde{P}_{ij} = \sum_{s=1}^{|S|} \mathbb{1}\{M_{si} - M_{sj} \geq 0\} * d(M_{si} - M_{sj}) \tag{1}$$

where $M_{si} - M_{sj}$ is the academic timestep delta (commonly measured in semesters, or quarters) between a student s's taking course i and course j. Function d is a second point in the visualization construction where flexibility is provided. The function may be chosen to de-emphasize the connection between subsequent courses.

For example, in an analysis of degree completion we may be interested in cases when course offerings were taken as closely together as possible. In contrast, when planning curricula we may wish simply to learn the probabilities of two courses taken in sequence, even if several academic terms apart.

Function d may be chosen to be continuous or discrete. For example, the following choice attenuates the relationship between two courses through an exponential decay over enrollment term distance:

$$\tilde{P}_{ij} = \sum_{s=1}^{|S|} \mathbb{1}\{M_{si} - M_{sj} \geq 0\} * \lambda^{M_{si} - M_{sj}} \tag{2}$$

where, λ is a constant that controls decay rate. Intuitively, this type of function gives more weight to close course-pair enrollments than temporally distant course-pair enrollments.

Alternatively, we may choose a function d that tallies a relationship between two courses only when course j immediately follows course i. In the following example, elapsed time between two courses beyond one term would sever the relationship:

$$d = \begin{cases} 1, & M_{si} - M_{sj} = 1 \\ 0, & M_{si} - M_{sj} > 1 \end{cases} \tag{3}$$

More elaborate discrete functions could amplify course relationships of particular time distances.

Using $\tilde{P}$, we calculate the final projection matrix P. In some experiments, we set $P = \tilde{P}$, in order to gain access to a raw count projection matrix. In other experiments, we set P to be a matrix of conditional probabilities between courses. In this case, each entry P_{ij} is a conditional probability $p(j|i)$ whose parameters are computed using the following closed form expression:

$$P_{ij} = p(j|i) = \frac{\tilde{P}_{ij}}{\sum_{s=1}^{|S|} \mathbb{1}\{M_{si} > 0\}} \tag{4}$$

P_{ij} thus represents the proportion of students who take the course sequence: course i followed by j out of the total number of students who take course i at any point. One may note that this bears a resemblance to a Bayesian Network; however, we make the assumption that the process of transitioning from course node to course node is Markovian in nature. This eases implementation but precludes our model from being a true Bayesian Network due to the emergence of cycles.

In the following section we describe how we combine the projection matrix with an existing graph visualization tool. For simplicity we choose the timestep delta function d to be discrete, although we will alter the nature of the function for different examples.

GRAPH VISUALIZATION

We use *Cytoscape* to visualize the generated network [40]. Cytoscape was developed to model biological systems and

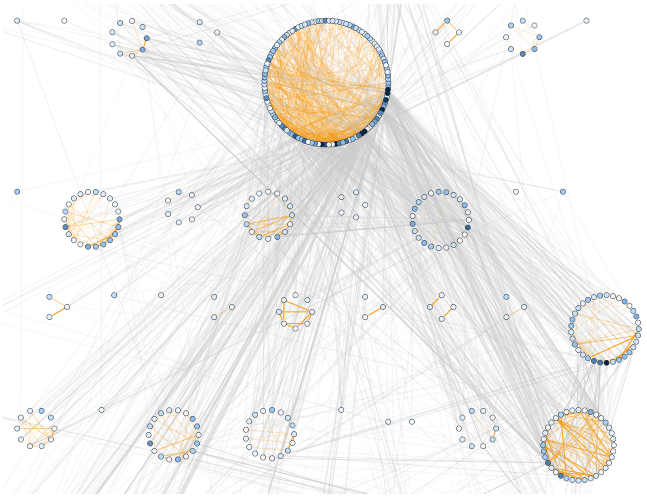

Figure 2. An example graph generated using our projection model, showcasing enrollment patterns within different departments. Each ring is comprised of courses offered by different departments.

interactions between cellular organisms. Yet the tool is well suited for representing our academic pathway data.

For one, Cytoscape uses a spring-embedder system for optimally spacing and aggregating nodes, and additionally allows the tool operator to cluster nodes by attributes [6]. For our purposes, we display courses within departments as ring structures. For example, the large ring in Figure 2 represents the Computer Science department in a graph generated from enrollment data filtered to include only CS majors. Each of the dots that comprise the circumference of the ring is one course in that department. The interior connecting arrows display the flow of students from one CS course to another.

Cytoscape allows us to assign unique colorings for both nodes and edges to visualize node and edge-level attributes. In all of the figures below, we color edges between nodes in the same department orange and all edges between courses in different departments grey. We further modulate the opacity of an edge from course *i* to course *j* to represent the conditional probability of taking course j after course i. Finally, each node is colored in a blue-gradient, representing the prior probability of a student enrolling in the respective course.

GRAPH ANALYSIS & MODEL APPLICATIONS

We have elaborated upon and presented a mapping of student flow through the university's curricular landscape. We will now showcase the mapping's utility from the perspective of several stakeholder groups. The following sections will exemplify two dimensions of use. We will explain how the mathematical tools associated with formal graphs can be deployed to answer already well defined questions. In addition, we will showcase visual representations of the enrollment graph as a means to support interactive exploration and discovery.

Evolution of Department Accessibility

We begin by showing how *Via* recovers high-level, university-wide trends that have been independently observed in the literature.

In particular, we determine the overall *accessibility* of departments at varying historic time periods. By accessibility we mean the breadth of courses that feed into particular departments. Alternatively, the term describes the degree to which courses in one department act as resources for students who take courses beyond their immediate core interest.

Our analysis employs the *PageRank* algorithm devised by Page *et al.* [32]. We compute PageRank on a series of networks generated from three different temporal slices of the dataset. The conclusions we derive from our analysis are directly applicable only to the university from which we acquire our data. However, the manner in which we employ the *Via* toolkit holds generally, and can be used on any dataset similar in structure to our own.

In particular, a temporal analysis of changing course enrollment patterns can be achieved by leveraging the course-level attribute filtering mechanism of our model. Recall that the *Via* toolkit allows users to specify particular attributes to filter by when creating a final projection graph. Our dataset contains entries for the date at which a particular course was completed from 2000 to 2018. Using this information we create three separate course sequence graphs.

Each graph represents course sequences for courses enrolled in between each of the following time windows 1) 2000-2008, 2) 2005-2013, and 3) 2010-2018. We use raw counts of consecutive course enrollments from term to term in order to define edge weights between courses. The PageRank execution then turns those raw counts into transition probabilities.

On these separate networks, we run PageRank with random restarts in order to determine a general distribution over the courses in a particular network. Observe that the final state distribution of the PageRank scores can be interpreted as the probability of ending up at a certain node given a set of random walks. We then sum the PageRank scores of courses to represent and analyze the overall PageRank distribution of the departments within the filtered time-windows. We do the same with a network generated on students who graduated with Bachelors of Arts degrees. The results of running PageRank on these two sets of three graphs are presented in Figure 3.

We can interpret the final results as follows: for a given department D with PageRank score X, starting at a random course, and selecting successors according to successor probability will land students in D $(X * 100)\%$ of the time. Thus, the percentages in Figure 3 signal the ease with which a department's courses are found from the other courses in the graph.

More generally, we can interpret the relative size of the departments in the Figure 3 as the accessibility of departments across the different disciplines at the university throughout any time period in an undergraduate's career. One could interpret this quantity as having an inverse relationship with the activation energy required to begin taking courses in a given department. In this case a department's high PageRank score would indicate a relative ease of entering and taking a course within a particular department. While closely linked, this ease differs from metrics regarding course composition and univer-

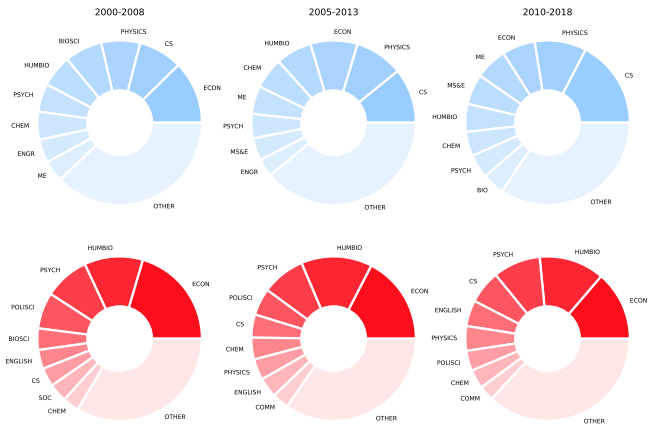

Figure 3. A comparison over time of PageRank values by department in both the entire undergraduate network (top row) and the undergraduate network filtered on students who earned a Bachelor's of Arts (BA) degree (bottom row). We argue that PageRank offers us a highly flexible method of determining course accessibility based on aggregate student behavior.

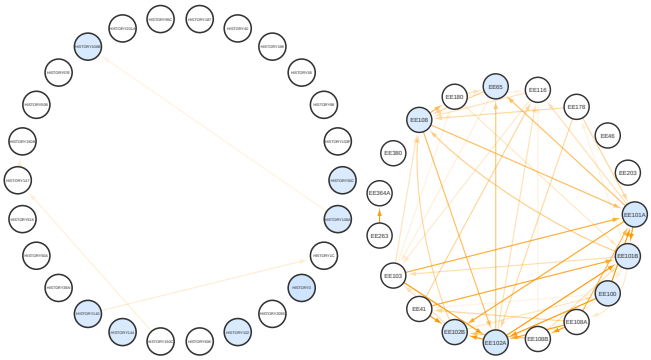

Figure 4. A comparison between the student behavioral patterns in two departments— History (left) and Electrical Engineering (right)—from the years 2012-2016. Labels in the small circles are course names, such as "EE41" and "HISTORY198".

sity enrollment due to the emphasis on the flow of students from course to course. For example, we see that missing from the 2010-2018 undergraduate results is the mathematics department, despite it ranking 4th in raw number of enrollments during the given timeframe. This is likely because students become much less likely to enroll in math courses after fulfilling the requirement (commonly during their freshman year), whereas many students enroll in many of the popular courses in the CS, ECON, and ME departments throughout their undergraduate career, perhaps because these subjects are less dependent on knowledge acquired—or not acquired— during the final years of high school.

We observe that between 2000 and 2018, the PageRank score of the Computer Science department has only increased. This applies in general to all degree-seeking students and to the BA-seeking students. Similarly we note a rise in the amount of Physics courses taken during this time window. It is unclear whether this increase is due to growing interest in Physics courses, or whether it is a consequence of the popularity increase in Computer Science: the CS major requires several Physics courses.

Another point to note is a general decrease in enrollment in economics courses, but a noticeable uptick in the amount of classes taken in Management Sciences and Engineering (MS&E). These two trends are strongly correlated, as the material of MS&E seeks to combine economics and finance with computer science and statistics.

Our results are in harmony with observations made by education researchers. Over the past 20 years the proportion of students majoring in STEM, and particularly computer science has varied greatly [11]. Introductory, mid-level and upper-level computer science courses have all shown a demonstrable increase upwards of 150% of student enrollment between 2000 and 2017. Driving student's decisions to enroll in more technical courses is primarily the greater monetary compensation that these majors promise at the workplace [13]. Interestingly, however, while the total amount of classes taken within Computer Science has risen noticeably, we do not observe that the total amount of student majors in Computer Science has increased at the same rate. Rather between 2000 and 2010 the number of Computer Science majors showed an uptick of 33%. However, between 2010 and 2018 the percent increase of Computer Science majors drops to 5.8%. This finding is more generally in line with the observed trend that the Computer Science major has not experienced a constant exponential increase in popularity. Rather, 2005 was one of the years with the lowest rates of students who self-reported interest in majoring in Computer Science when entering college [34]. Between 2000 and 2008 it was, in fact, economics and business-related courses that remained one of the most popular majors of study nation wide [31]. Overall, our quantitative analysis is able to corroborate much of the education research regarding course enrollment patterns over the past 20 years. We stress that the methodology used to generate these results is not dependent on the source of data, but is rather facilitated by the graphical modeling of student course enrollments that *Via* enables. In addition, the incorporation of filtering mechanisms beyond time, such as student demographics, would allow for more fine grained understanding of course accessibility based on enrollment data alone.

Student Stakeholders

Among university stakeholders, students are one of the principal groups that would benefit from the visual insights that can be gleaned from the *Via* toolkit. In this section we will highlight two particular use cases. First, our visual modeling of course sequences allows a student to identify which majors require students to take courses from within only a certain department, and which allow for more academic exploration across departments. We can report this result both visually, by directly looking at the intra-department orange arrows in a graph, and by calculating the modularity of certain departments. Modularity is a metric that represents the connectivity of clusters within a graph, by calculating the over-representation of edges among groups of nodes.

In Figure 4, we compare the interconnectivity of the History and Electric Engineering majors. Observe that the rings come

from a conditional probability projection model using a discrete discount function d that only captured proximate course enrollments (one academic time-step apart). Here we see that the visualizations align with the modularity scores associated with History (0.003) and Electrical Engineering (0.028). The higher the modularity score the more a department is intra-connected. This effect is caused by highly prescriptive requirement structures.

As a second use case, we illustrate how *Via* can be leveraged to discover the "try-me" courses. Departments offer such service courses for students majoring in unrelated fields, but who are interested in exploring other areas of study. *Via* enables quick discovery of such courses.

We find the "try-me" courses by filtering on students of a particular major during sequence matrix generation. We then identify the most popular courses within this generated graph that are not in the major used as a filter. For instance, by filtering on the History major, we observe that of the History majors, nearly 28% take CS105 and 21% take CS106A. These trends are visualized through variations in node coloration in Figure 5.

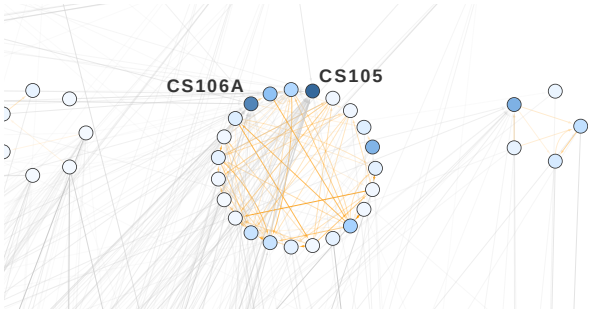

Figure 5. A visualization of the most popular courses in the Computer Science department for History majors.

Babad et al. have discovered that the primary reasons students decide to enroll in a particular course are the learning value of the class followed closely by the lecture style [1]. Particularly difficult courses are avoided by students unless otherwise required. Our assessment of the recommended Computer Science "try-me" courses for History majors corroborates these results. Although CS106A is branded as the most popular introductory Computer Science course at the university from which we have acquired our dataset, CS105 presents itself as the Computer Science course for non-majors and is also known to be less rigorous. Our *Via* toolkit is thus able to offer a more personalized course discovery system for students of a particular major.

Instructor Stakeholders

Instructors are a second group of stakeholders that can benefit from the easy interpretability of our course sequence visualizations. It is of particular interest to instructors to gain an overview of the sorts of students that are enrolling in their courses, in order to customize and improve lesson plans [27].

We show that it is possible to use the course probabilities associated with course-to-course interactions to study the background of students who take higher-level Spanish language classes. We accomplish this analysis by examining the conditional probability graph generated with all enrollment data. We set our function d as sensitive only to enrollments in course pairs within one year of each other in order to capture delayed enrollment behaviors. We can thus interpret edge weights as probability $p(i|j$ within the past year$)$. In Figure 7, we examine the courses associated with the V-structure colored in red— SPANLANG3 (bottom), SPANLANG2A (left), and SPANLANG11C (top). SPANLANG3 is the final term of the non-accelerated offering of the first-year Spanish language sequence, SPANLANG2A is the final term of the accelerated first-year sequence (often taken by students with prior exposure to the language through high school programs), and SPANLANG11C is the first academic term of the second-year sequence.

From the visualization alone, it is evident that the relationship between SPANLANG3 and SPANLANG11C is stronger than the relationship between SPANLANG2A and SPANLANG11C. Indeed, the probability that a student enrolls in SPANLANG11C within a year of completing SPANLANG3 is almost three times that of a student who enrolls in SPANLANG11C after completing SPANLANG2A, with $p(\text{SPANLANG11C}|\text{SPANLANG3}) = 0.038$ and $p(\text{SPANLANG11C}|\text{SPANLANG2A}) = 0.013$. It is possible that this effect stems from a difference in the instruction of grammatical foundations in the university's introductory courses, and high school language programs. If so, then this observation is in accordance with the research of Leaver *et al.*, which details the significance in varying teaching philosophies commonly seen in foreign language classrooms at different language acquisition levels [27].

While one may interpret this observation in many ways, we argue that the simplicity of the process through which this information can be found greatly decreases the time and effort required to observe enrollment patterns from the perspective of university administrations.

Administrator Stakeholders

Finally, our model provides an intuitive medium for understanding department-wide dynamics across varying student demographics. Administrators in charge of course resource allocation may be interested in the enrollment patterns of students as they move through a department's course offerings. We provide two use cases to illustrate how *Via* can be leveraged to derive insights into the structure of courses within departments.

First, we study which classes tend to guide a student into a certain department. We report the top 10 classes within a department that, once taken by a student, have the highest average number of subsequent enrollments within the same department. In Figure 6 we show the student persistence within a department using a raw count projection model with discrete discount function d that captured all non-co-enrollment relationships between courses from 2014-2018. This "persistence within a department" metric can be computed for a given class i in a department x by finding the out-degree of the class node and then dividing by the total number of students who enrolled in class i. This method is equivalent to finding the expected number of courses taken within department x after

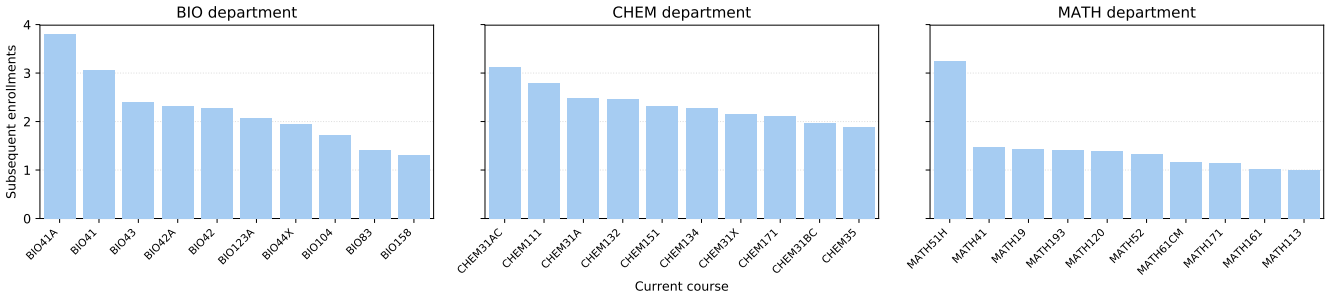

Figure 6. Student persistence metrics in Biology, Chemistry and Mathematics. The support courses for the biology and chemistry departments' introductory courses, BIO41A and CHEM31AC, respectively, lead to higher subsequent enrollment rates within the respectivedepartments.

enrollment in class *i*. Observe that this is robust to a course's discontinuation across academic years due to the conditions under which the ratio is calculated. Here we measure the average number of courses taken within a department after having enrolled in a given course. Within the Biology and Chemistry departments, we recover the insight that the 1-unit companion courses to the introductory courses (BIO41A and CHEM31AC for Biology and Chemistry, respectively) increased student persistence. Intuitively, this observation indicates that enrollment in these courses led to an overall increase in the number of courses taken within the department, on average. We contrast this with the Math department. Students who enrolled in the most courses offered within the Math department were those who enrolled in the first course of the honors multivariable mathematics sequence, MATH51H. This observation can be explained by the fact that the classes in the primary introductory course sequence in the Math department (MATH51, MATH52, MATH53) often serve as prerequisite courses for classes in other departments in engineering and social sciences. Enrolling in the honors-level mathematics sequence perhaps indicates a given student's higher intrinsic inclination to the subject matter. **Course Roles:** courses often play particular roles in their department, or across a university. We already mentioned service courses that provide non majors with a taste of the department's discipline. Personnel within a department generally understand these roles. But department outsiders do not possess such knowledge. Researchers such as education or sociology scholars studying universities and colleges other than their own have even less access to role information.

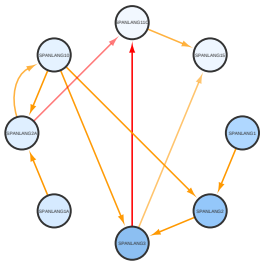

Figure 7. Relationships within the language department's Spanish course offerings from 2012-2016 (Best viewed in color). We note that the probability that a student transitions from first to second year Spanish significantly increases if initially enrolled in the non-accelerated track.

We illustrate here how *Via* can recover course "roles." The application of more sophisticated algorithms rooted in graph theory allows *Via* users to detect both roles and communities within course networks. To this end we applied a RolX [22] analysis. The analysis requires enrollment data alone. Intuitively, the analysis attempts to identify sets of nodes across a graph that are similar in the subgraph structure that surrounds them. The algorithm then attempts to identify families of such nodes, analogously to how *k-means* defines clusters. As with clustering, RolX does not provide semantics for discovered clusters. Those must be provided by human interpretation.

Figure 8, presents the usage of RolX to discover the roles of courses as they pertain to students. We posit the successful recovery of introductory courses (Role 1, red), intermediate courses (Role 2, blue), and senior project / enrichment courses (Role 3, green). This assignment of roles is supported by the NodeSense metrics that are returned by the RolX algorithm. We will examine these metrics by case. According to the NodeSense metrics, Role 1 is a classification whose member nodes tend to have high closeness to other nodes, as well as a very low in-degree. These properties, coupled with high betweenness, signify that these nodes are likely gateway courses that enable students who enroll in these courses to access new parts of the network, perhaps through an acquisition of foundational skills.

It may seem counter intuitive that introductory courses have low in-degree. Recall, however, that inflow represents students arriving from other courses. Yet, introductory courses are often consumed by students just entering higher education. They will not yet have taken other courses. In our graph model there will thus not be a 'feeder course,' because we do not model high school-terminating courses.

Role 2, which includes courses with high out-degree and high pagerank score, can be interpreted as intermediate level courses, which are less accessible than Role 1 nodes yet lead to a more diverse set of courses. A low betweenness score in this case likely points to the fact that there are many ways to navigate the mid-level requirements of one's undergraduate career.

Finally there are the Role 3 courses, which show themselves to have very low closeness, betweenness and degree. We posit that these are enrichment or senior project courses, which lead

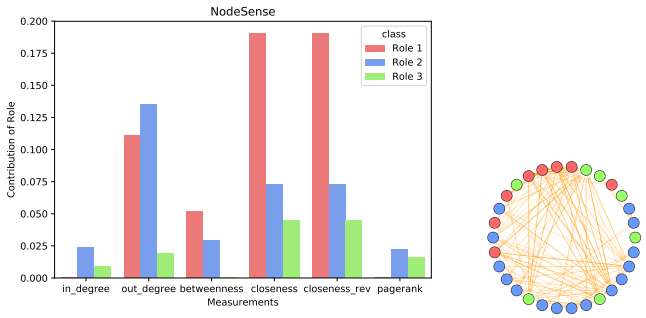

Figure 8. RolX analysis of course sequences (Best viewed in color). We posit that Roles 1, 2, and 3 describe introductory courses, intermediate level courses, and senior project / enrichment courses.

to very few follow up courses, either due to completion of the degree or due to students who only enroll to try out a new subject.

Although RolX role assignments were developed using conditional probability edge weights, we find that membership in Roles 1,2, and 3 align well with high, medium, and low persistence scores of Figure 6. This consistency further supports the above role interpretations.

CONCLUSION AND FUTURE WORK

We have described and demonstrated our *Via* toolkit, which is constructed atop a new approach for visualizing and interpreting transcript enrollment data. Such data are available at every major US university. *Via* allows stakeholders as diverse as students, instructors, and administrators to understand aggregate student behavior over time. Using graph visualization and graph theoretic computation in concert we have presented several *Via* use cases.

We explained the process through which standard enrollment data can be transformed into graph structures that may be tuned to particular investigative goals. This flexibility arises both during graph construction, and during interactive manipulation of the graphs. We deploy the existing tool Cytoscape for such manipulations, but other tools may be just as appropriate, once graphs are constructed through the algorithm we have presented.

We will continue our exploration by investigating networks with multiple node types: one to represent students, another to model courses. The answers to other types of questions will be found through this different family of graphs.

We further plan to extend *Via* to include support for simulations and what-if analyses. Further effort will also need to be invested into making the graph construction process easy to use.

Many questions may be answered by skillful SQL queries over university datasets. But these approaches often fall short when questions are not yet clearly defined, and relatively large amounts of data need to be shaped and reshaped to discover patterns. As postsecondary education is increasingly held accountable for performance, a deep understanding of such patterns and trends over time will be required. We have presented a fresh step toward such capability.

REFERENCES

1. Elisha Babad and Arik Tayeb. 2003. Experimental analysis of students' course selection. (sep 2003). DOI: http://dx.doi.org/10.1348/000709903322275894
2. Thomas R Bailey, Shanna Smith Jaggars, and Davis Jenkins. 2015a. *Redesigning America's community colleges*. Harvard University Press.
3. Thomas R Bailey, Shanna Smith Jaggars, and Davis Jenkins. 2015b. *Redesigning America's community colleges*. Harvard University Press.
4. R. Baker and N. Huntington-Klein. 2018. *Student Preference for Guidance and Complexity in College Major Requirements*. Technical Report 18-06. Stanford Center for Education Policy Analysis, http://cepa.stanford.edu/wp18-06.
5. Vladimir Batagelj. 2003. Efficient Algorithms for Citation Network Analysis. (2003), 1–27. DOI: http://dx.doi.org/arXiv%20:%20cs%20/%200309023v1
6. Giuseppe Di Battista, Peter Eades, Roberto Tamassia, and Ioannis G. Tollis. 1994. Algorithms for drawing graphs: an annotated bibliography. *Computational Geometry: Theory and Applications* 4, 5 (1994), 235–282. DOI: http://dx.doi.org/10.1016/0925-7721(94)00014-X
7. Stephen P. Borgatti, Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca. 2009. Network analysis in the social sciences. *Science* 323, 5916 (2009), 892–895. DOI: http://dx.doi.org/10.1126/science.1165821
8. Simona Botti and Sheena S Iyengar. 2006. The dark side of choice: When choice impairs social welfare. *Journal of Public Policy & Marketing* 25, 1 (2006), 24–38.
9. Christopher G. Brinton, Swapna Buccapatnam, Felix Ming Fai Wong, Mung Chiang, and H. Vincent Poor. 2016. Social learning networks: Efficiency optimization for MOOC forums. *IEEE INFOCOM* 2016-July (2016), 1–9. DOI: http://dx.doi.org/10.1109/INFOCOM.2016.7524579
10. Daniel F Chambliss. 2014. *How college works*. Harvard University Press.
11. Computing Research Association. 2017. Generation CS: Computer Science Undergraduate Enrollments Surge Since 2006. (2017), 1–51. http://cra.org/data/Generation-CS/
12. Peter M Crosta. 2014. Intensity and attachment: How the chaotic enrollment patterns of community college students relate to educational outcomes. *Community College Review* 42, 2 (2014), 118–142.
13. James P Downey and David Roach. 2007. MIS versus Computer Science : An Empirical Comparison of the Influences on the Students ' Choice of Major. *Journal of Information Systems* 20, 3 (2007), 357–369. http://search.proquest.com/openview/9ddf48c93e42103802b16117f8a6675a/1?pq-origsite=gscholar
14. Emile Durkheim. 1951. *Suicide : a study in sociology*. Free Press. 405 pages. https://www.google.com/search?q=E.Durkheim+suicide
15. Danyel Fisher and Paul Dourish. 2004. Social and temporal structures in everyday collaboration. In *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*. 551–558. DOI: http://dx.doi.org/10.1145/985692.985762
16. Jeffrey Fletcher, Markeisha Grant, Marisol Ramos, and Melinda Mechur Karp. 2016. Integrated Planning and Advising for Student Success (iPASS): State of the Literature. CCRC Working Paper No. 90. *Community College Research Center, Teachers College, Columbia University* (2016).
17. Marie Josée Fortin, Patrick M.A. James, Alistair MacKenzie, Stephanie J. Melles, and Bronwyn Rayfield. 2012. Spatial statistics, spatial regression, and graph theory in ecology. *Spatial Statistics* 1, February (2012), 100–109. DOI: http://dx.doi.org/10.1016/j.spasta.2012.02.004

18. Santo Fortunato, Vito Latora, and Massimo Marchiori. 2004. Method to find community structures based on information centrality. *Phys. Rev. E* 70 (Nov 2004), 056104. Issue 5. DOI: http://dx.doi.org/10.1103/PhysRevE.70.056104
19. Eugene Garfield and Eugene Garfield. 1964. Science Citation Index. *Science* 144, 3619 (1964), 649–654.
20. Kamalika Basu Hajra and Parongama Sen. 2005. Aging in citation networks. *Physica A: Statistical Mechanics and its Applications* 346, 1-2 (feb 2005), 44–48. DOI: http://dx.doi.org/10.1016/J.PHYSA.2004.08.048
21. S. J. Hall and D. G. Raffaelli. 1993. Food Webs: Theory and Reality. *Advances in Ecological Research* 24, C (1993), 187–239. DOI: http://dx.doi.org/10.1016/S0065-2504(08)60043-4
22. Keith Henderson, Brian Gallagher, Tina Eliassi-Rad, Hanghang Tong, Sugato Basu, Leman Akoglu, Danai Koutra, Christos Faloutsos, and Lei Li. 2012. RolX. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*. ACM Press, New York, New York, USA, 1231. DOI: http://dx.doi.org/10.1145/2339530.2339723
23. Zan Huang, Xin Li, and Hsinchun Chen. 2005. Link Prediction Approach to Collaborative Filtering. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '05)*. ACM, New York, NY, USA, 141–142. DOI: http://dx.doi.org/10.1145/1065385.1065415
24. Sheena S Iyengar and Mark R Lepper. 2000. When choice is demotivating: Can one desire too much of a good thing? *Journal of personality and social psychology* 79, 6 (2000), 995.
25. Davis Jenkins and Sung-Woo Cho. 2013. Get with the program… and finish it: Building guided pathways to accelerate student completion. *New directions for community colleges* 2013, 164 (2013), 27–35.
26. Xiaonan Ji, Raghu Machiraju, Alan Ritter, and Po-Yin Yen. 2015. Examining the Distribution, Modularity, and Community Structure in Article Networks for Systematic Reviews. In *AMIA Annual Symposium Proceedings*, Vol. 1927. American Medical Informatics Association.
27. Betty Lou. Leaver and Boris. Shekhtman. 2002. *Developing professional-level language proficiency*. Cambridge University Press. 308 pages.
28. Harold J. Leavitt and Ronald A. H. Mueller. 1951. Some Effects of Feedback on Communication. *Human Relations* 4, 4 (nov 1951), 401–410. DOI: http://dx.doi.org/10.1177/001872675100400406
29. David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American society for information science and technology* 58, 7 (2007), 1019–1031.
30. Rebekah Nathan. 2006. *My freshman year: What a professor learned by becoming a student*. Penguin.
31. National Center for Education. 2018. Undergraduate Degree Programs. *Learning* (2018), 1–6. https://education.uky.edu/academics/degree-programs/
32. L Page, S Brin, R Motwani, and T Winograd. 1999. The PageRank citation ranking: Bringing order to the web. (1999). http://ilpubs.stanford.edu:8090/422
33. Zachary A. Pardos and Andrew Joo Hun Nam. 2018. A Map of Knowledge. *CoRR* abs/1811.07974 (2018). http://arxiv.org/abs/1811.07974
34. David a. Patterson. 2005. Restoring the popularity of computer science. *Commun. ACM* 48, 9 (2005), 25. DOI: http://dx.doi.org/10.1145/1081992.1082011
35. P A Pevzner. 1989. 1-Tuple DNA sequencing: computer analysis. *Journal of biomolecular structure & dynamics* 7, 1 (1989), 63–73. DOI: http://dx.doi.org/10.1080/07391102.1989.10507752
36. James E Rosenbaum. 2011. The complexities of college for all: Beyond fairy-tale dreams. *Sociology of Education* 84, 2 (2011), 113–117.
37. James E Rosenbaum, Regina Deil-Amen, and Ann E Person. 2007. *After admission: From college access to college success*. Russell Sage Foundation.
38. Barry Schwartz. 2004. The paradox of choice: Why more is less. Ecco New York.
39. Judith Scott-Clayton. 2015. *The Shapless River*. Routledge, New York and London, Chapter 6, 102–124.
40. Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* 13, 11 (2003), 2498–2504.
41. Tanmay Sinha. 2014a. Supporting MOOC Instruction with Social Network Analysis. (2014). http://arxiv.org/abs/1401.5175
42. Tanmay Sinha. 2014b. Who negatively influences me? Formalizing diffusion dynamics of negative exposure leading to student attrition in MOOCs. (2014). http://arxiv.org/abs/1407.7133
43. Ahmad Slim. 2016. *Curricular Analytics in Higher Education*. Ph.D. Dissertation. The University of New Mexico.
44. Tongshuang Wu, Yuan Yao, Yuqing Duan, Xinzhi Fan, and Huamin Qu. 2016. NetworkSeer: Visual analysis for social network in MOOCs. *IEEE Pacific Visualization Symposium* 2016-May (2016), 194–198. DOI: http://dx.doi.org/10.1109/PACIFICVIS.2016.7465269
45. Mengxiao Zhu, Yoav Bergner, Yan Zhang, Ryan Baker, Yuan Wang, and Luc Paquette. 2016. Longitudinal Engagement, Performance, and Social Connectivity: A MOOC Case Study Using Exponential Random Graph Models. *Conference on Learning Analytics & Knowledge* (2016), 223–230. DOI:http://dx.doi.org/10.1145/2883851.2883934
46. Marinka Zitnik, Rok Sosic, and Jure Leskovec. 2018. Prioritizing network communities. In *Nature communications*, Vol. 9.