The Content of Their Coursework: Understanding Course-Taking Patterns at Community Colleges by Clustering Student Transcripts

Matthew Zeidenberg
Community College Research Center

Marc Scott
New York University

October 2011

CCRC Working Paper No. 35

Address correspondence to:

Matthew Zeidenberg
Senior Research Associate, Community College Research Center
Teachers College, Columbia University
525 West 120th Street, Box 174
New York, NY 10027
212-678-3091
Email: zeidenberg@tc.edu

Funding for this study was provided by the Bill & Melinda Gates Foundation. The authors thank Sung-Woo Cho, Madeline Joy Weiss, Michelle Van Noy, Shanna Smith Jaggars, and Davis Jenkins of CCRC and David Prince and Tina Bloomer of the Washington State Board for Community and Technical Colleges (SBCTC) for their useful feedback. We also thank the SBCTC for providing the data used in this paper and Betsy Yoon for her excellent editorial work.

Abstract

Community college students typically have access to a large selection of courses and programs, and therefore the student transcripts at any one college or college system tend to be very diverse. As a result, it is difficult for faculty, administrators, and researchers to understand the course-taking patterns of students in order to determine what programs of study they appear to be pursuing. Attempting to examine these patterns and then comparing them with listed program requirements would be a very time-consuming activity. The most common way of assigning a program of study to a student—picking the subject in which she has taken the most courses—is overly simple, because many programs require courses across several subjects. However, because students who have similar patterns of course-taking in terms of subjects and particular courses taken are likely to be in similar programs, clustering can be a useful way to make sense of the relevant data.

Clustering allows researchers to group similar items into clusters, relying only on a measure of the similarity of those items. In this paper, we apply a clustering algorithm to the problem of understanding college transcripts, which serve as the items to be clustered. To our knowledge, this is the first effort to organize transcripts based on their course content using clustering. We base the measure of similarity on the proportion of curricular subjects that each transcript has in common with every other one.

Our data are community and technical college transcripts for a cohort of students who first entered the Washington State system during the fall of the 2005–06 academic year and who had no prior postsecondary experience. We used our clustering algorithm to separately cluster liberal arts and career-technical students. We found that the algorithm did a good job of separately clustering each of these groups. The clusters roughly corresponded to programs of study, so we were able to estimate how many students were undertaking each program and what subjects students were studying within each cluster. We were also able to examine the demographics and the completion and transfer rates of the students within each cluster, in order to get an idea of what types of students were in each program of study and how successful they seemed to be in college. We found substantial variation on these dimensions as well as on the extent to which students' programs were either concentrated in a single subject or spread across several subjects.

Clustering is a powerful way to understand the course-taking patterns of students and assign programs of study. It makes few prior assumptions about the data; rather, it allows the data to organize itself based on a similarity measure. It relieves the analyst of determining what the program categories should be. It has the ability to detect patterns of activity across subjects within student transcripts. Note that although we have applied this method to community college students, it is applicable at all levels of postsecondary education. We conclude that this method would be useful to researchers throughout education who are trying to understand student course-taking patterns and programs of study, and who need to organize large amounts of transcript data.

Table of Contents

1. Introduction

It is often quite difficult to make sense of the course-taking patterns of community college students, but such patterns are central to the life of any college. The typical community college course catalog lists many programs, and as students are often given a great deal of freedom in selecting the courses that they take, they often take courses in largely unrelated subjects, as opposed to taking a focused set of courses corresponding to a program. This is in part due to the fact that many students do not settle on a program early in their college career.

Since most community colleges offer numerous programs, it would be a very laborious task to compile college program requirements and then compare them with the transcripts of individual students. In community college systems with numerous colleges, each with its own set of program offerings, such a task would be compounded. In this paper, we describe a technique that can substantially reduce the time and complexity required to understand student transcripts.

For this study, we examine student transcripts from two-year public colleges in Washington State for students who first enrolled in the state's community and technical college system during the 2005–06 school year and who had no prior postsecondary experience. The aim is to understand student pathways empirically, in terms of actual student behavior, rather than normatively, in terms of prescribed programs. Presumably, there will be substantial overlap between student course-taking and program requirements. This work has the goal of developing a useful tool that will allow researchers and practitioners to better understand the course-taking activity of community college students, either at the college system level or at the level of the individual institution.

1.1 The Problem: Understanding Student Course-Taking Patterns

Because they are pressed for time doing other necessary research, it is often the case that researchers for individual educational institutions or for a community college system do not have a good overall sense of the programmatic pathways their students are pursuing or how students pursue these pathways. While they may be able to provide a general sense of things, they often do not have good information on how many students are pursuing each pathway, or exactly what courses these students are enrolled in. The

pathways that community college students take are diverse. Some pathways are pursued by many students; others by just a few. But there is generally a set of pathways that correspond more or less to program requirements that students take on the way to earning a credential.

Each college typically has a wide variety of programs, each with different requirements. It is not feasible in most cases to enter each of these programs into an IT system and then have the system determine what students are following what program. It is not at all clear how such a system would function, in any case, given that many students may not be following any single program, but rather may be taking a smattering of courses in a variety of programs. This may be particularly true of liberal arts students, who tend to "shop around" before choosing a major. Often, they do not choose a major until they are ready to transfer to a four-year institution.

In addition, in many if not most community colleges, there is no reliable way to measure student intent and major. Measures of student intent can change and are not reliable (Bailey, Jenkins, & Leinbach, 2006). Many community colleges allow credential-seeking students to pursue a program without declaring a major, and even when students do declare a major, it may not correspond very well to their actual course-taking behavior.

Community college career-technical education (CTE) programs, particularly those at the associate degree level, often have liberal arts course pre- or co-requisites, so it is not possible to fully characterize students' programs if we rely on the field of study of individual courses. Some community college CTE programs are designed in a sequenced or "stacked" format in which students can earn first a certificate and then an associate degree. It is sometimes hard to tell whether a student who is taking courses for a certificate is also pursuing an associate degree.[1]

1.2 Prior Research on Assigning Students to a Major

Since community college students often do not declare a major, and because even if they do, the actual courses taken may not correspond closely (or at all) to their declared major, researchers have devised methods of determining the majors of students in a

[1] We do not address the issue of determining this in this paper, although we hope to address it in further research.

college or system who have taken more than a few courses. There are several reasons for determining students' majors. For instance, one may want to compare graduation rates across majors, and the only way to do this is to have a way to identify when a student has entered a major. Or, similarly, one may want to examine the labor market outcomes of students who do coursework in particular fields without completing a degree. Or, one may want to compare the gender, ethnic and racial, and socioeconomic status (SES) composition of students in particular program areas, because labor market returns to different fields of study vary (Jacobson & Mohker, 2010).

As far as we know, all approaches to date to identifying a student's major that do not rely on a student's declared major, other than that described herein, rely on the single subject in which that student has taken or completed the largest number of courses. For instance, Jenkins and Weiss (2011), using Washington State community college data, defined a student as having *entered a concentration* if they take at least three courses or 12 quarter credits in a single subject. They defined *being in a concentration* as having completed three courses or 12 quarter credits, again in a single subject. They defined *subjects* in terms of sets of codes from the federal Classification of Instructional Programs (CIP) system. CIP codes are set by the U.S. Department of Education to classify curricula into particular subjects.

Jacobson and Mohker (2010) took a similar approach in their study of the effect of concentrating in a career and technical education (CTE) field on subsequent earnings. They used high school and college data from Florida. For those students completing at least 12 credits in college, the authors identified the major as one of 11 CTE fields or seven academic fields, based on the field in which they took the most courses.

Stuart (2009) classified students into one of seven categories based, again, on the field in which they took the most credits overall. However, Stuart added a bit of nuance to this, because, for all of the terms prior to the term in which the student took her first course in her main major, there is also a *term major*, which is the subject in which the student took the most courses in that term.

These studies take this "single subject" approach, despite the fact that most majors generally require a pattern of coursework across multiple subjects, including, typically, the major subject, which may be relatively specific (e.g., accounting) or general

(e.g., business), related subjects (in this case, mathematics, computers, economics, and English), as well as distributional requirements (e.g., science, history, and so on.) Only some programs of study, typically those for certificates, allow study in only a single subject. The identification of major on the basis of a single subject probably works best for those students pursuing relatively short certificates in focused subjects, such as auto repair or cosmetology, and worst for transfer students in the liberal arts.

In what follows, we propose and demonstrate a method of assigning students into groups that is both more nuanced than the single subject approach and can empirically assign students into meaningful groups without applying top-down categories to the data. In the proposed method, we use characteristics of the data to organize the data themselves. By grouping together students who have similar patterns of study, we are able to "see where the action is," and to see whether these patterns are very specific, very broad, very focused, or very unfocused.

Thus, for instance, if there are a number of students who are taking something very specific, they will form a group that will show up just as significantly as a group of students, approximately equal in size, who are taking a broader program of study. For instance, real estate is a specific field of study within business; some students could be studying only real estate, while others might be studying business as a whole. This is exactly what we found with our data, and it is difficult to see how such a finding could be accomplished by assigning students to fields of study that are predefined by the researcher or by a coding scheme alone.

1.3 Clustering: A Way to Organize Student Transcripts into Meaningful Groups

Clustering, a data analysis technique, solves the problem of assigning students into meaningful groups by simply grouping students who have similar course transcripts. Each such set of students is placed into a cluster by the clustering algorithm, which relies on having a similarity measure (the precise similarity measure we used is described below), or a measure of how alike two transcripts are with each other. The exact sets that one obtains is dependent on the choice of clustering algorithm (there are many; we describe the specific one we have chosen below), the choice of the similarity measure, and the number of clusters that are specified.

The hope is that the clusters will contain meaningful content; that is, that each cluster will contain a set of students who are easily identified as being in the same or at least related programs of study, such as accounting or medical assisting. If this is the case, the clusters can also be described in terms of their demographics and their outcomes. For instance, a business assisting cluster might be predominantly female, and an auto repair cluster, predominantly male. This type of demographic information may be of interest to institutional researchers who want to know in which types of programs students with different characteristics are enrolling.

Hypothetically, both the business assisting and the auto repair cluster might award certificates, but the students in each cluster will earn these certificates at different rates, and one cluster might award primarily short certificates, others, predominantly long ones. Thus the clustering can provide information about content of the programs undertaken by students at the college, as well as the characteristics of the students in the programs and the outcomes of each program. Note that the clustering will put together both completers and non-completers, since it is done based on course-taking patterns alone and not on whether students complete a program. Often, we know something about what students who have completed a program have done because we have information about the program requirements, but often we know little about the activity of non-completers, who usually represent most of the students in a community college. Since each cluster contains both completers and non-completers, and each cluster contains corresponds (roughly) to a program of study, examination of each cluster makes it possible to examine the rate at which students in a program of study actually complete it and the impact of college efforts to improve completion rates. Note also that the clusters may not be precisely aligned with particular programs at a college, but that there should be overlap.

2. Data and Methods

The data that we used for the analysis described herein were drawn from the Washington State public, two-year college system. This is a system of 34 colleges, of which 29 are comprehensive colleges offering both baccalaureate transfer and career-technical programs, and of which five are technical colleges that emphasize technical

programs. Each college offers a distinct set of programs and courses; there is no common course numbering system across the colleges. Each course, however, is assigned a federal Classification of Instructional Programs (CIP) code. We used the CIP code as a proxy for the actual course department and number when conducting the clustering analysis.

The students whose transcripts we studied are a cohort of all students for whom 2005–06 was their first year in college, whose studies were at least partially funded by the state, and who were not international students. We restricted our scope to this cohort in order to have a relatively homogeneous set of students; otherwise, our data would include various other types of students, such as full-time employees who are taking a work-related course or two under a contract between their employer and a college. We have their transcripts as far forward as the fall of 2009. If students had enrolled in summer 2005 (the beginning of that academic year according to Washington State's record-keeping) they could have been enrolled for 18 quarters or for four-and-a-half academic years, although most students in our sample enrolled for many fewer quarters.

2.1 The Clustering Method

We now review the approach we used to cluster students based on their transcripts. Some details, particularly concerning the alternatives considered and choices made at different stages of the research, are omitted here but can be found in Appendix A.

Each student is represented by his or her transcript; these transcripts are clustered. The clustering algorithm requires a method of computing the similarity between transcripts. Considering two transcripts, we first determine the number of courses, in terms of CIP codes, that they have in common (the overlap for each transcript pair). Note that since we are clustering system-wide, two courses will be considered the same if they have the same CIP code, because each college has different course numbers. We cannot use the overlap directly, because it will tend to be higher, all things being equal, if the two transcripts are longer. So, instead, we consider the share of each transcript accounted for by the overlap. For instance, if we have two transcripts, one of length six and one of length eight, and four courses overlap, then that is two thirds of the first transcript and one half of the second transcript. We then average these shares to get the "similarity." In

this case, the average of one half and two thirds is seven-twelfths, or about 0.5833. Note the similarity ranges between 0, when the transcripts are completely disjointed or non-overlapping, and 1, where they are identical. The clustering algorithm we used actually uses the dissimilarity, which is defined as 1 minus the similarity.

Note that if a CIP code appears more than once in a transcript representing a distinct course each time, it has the opportunity to match more than once. For instance, if a given CIP code appears two times in one transcript and three times in the second, it would contribute twice to the overlap. If there were only one instance of that CIP code in the first transcript, then it would only contribute once to the overlap, and there would be a higher dissimilarity value between the transcripts (and a lower similarity). Note that in computing the similarity between two students' transcripts, we ignored the temporal component of the transcript, that is, the order in which students took their courses; in future research, we may include it.

The particular clustering algorithm that we used is called partitioning around medoids (PAM) (Kaufman & Rousseeuw, 2008). This algorithm is often used when clustering categorical data. College transcripts are a kind of categorical data known as nominal data. Nominal categorical data are not continuous measurements, such as height or weight, but are composed of discrete, dissimilar, unordered categories. A simple example of categorical data is a set of flowers consisting of the rose, tulip, and lily. There is no obvious way to put these on a scale, that is, into the familiar *xyz* space of analytic geometry. With data that can be mapped into ordinary space, rather than categorical data, clustering algorithms analogous to PAM are often used, and the data item that is in the physical center of each cluster is referred to as the *centroid.*

Here, with categorical data and using PAM, the analogous center is called the *medoid*. It is the data item that is closest, on average, to all of the other items in the cluster. Here, *closest* is defined by the dissimilarity metric we have defined. Note that this dissimilarity metric is not drawn from ordinary space; all we have, numerically, is the pair-wise dissimilarity between students' transcripts; we have not placed the transcripts into a space, nor would they easily fit in such a space, because the dissimilarities, which are analogous to physical distances, do not follow the laws of ordinary geometry (such as the Pythagorean theorem).

The way that PAM works, in this context, as follows:

1. The user specifies a number of clusters (note that there is some art to this; we will discuss how we did it in our particular case below).
2. A set of transcripts is randomly selected to act as the medoid of each of these clusters.
3. Each of the remaining transcripts is assigned to the cluster of the medoid it is closest to.
4. For each cluster in turn, the student transcript that minimizes the sum of all dissimilarities to it is promoted to medoid (note that the current medoid may retain its position).
5. All of the transcripts are reassigned to their closest medoid, globally.

Steps 4 and 5 above are then repeated until there is no change in the selected medoids and thus in the clustering.

2.2 Tuning the Clustering

Note that this type of clustering is as much of an art as a science, requiring some human judgment to get good results. The following describes key decisions that we made in this case.

First of all, we decided to restrict the set of students to students who had completed at least four college-level courses. This cut down the number of students substantially, because many students never get out of developmental courses that are below college-level, and many who do reach college-level courses do not get very far in terms of completed college courses. But we believe that it is not meaningful to assess what students are doing in college unless they have taken some number of college-level courses, so we selected four courses as a minimum. Below this number, we do not have much indication of a student's interests.

Community colleges mainly serve two groups of students: transfer students, who typically take a liberal arts program, and students who enroll in a career-technical educational (CTE) program. In our initial experiments with clustering, we clustered all the students together. This created a number of liberal arts clusters, which were not very

distinct from one another, as well as CTE clusters that were more clearly distinguishable. As a result, we decided to cluster liberal arts students and CTE students separately. We defined a *liberal arts student*, conservatively, as one for whom 75% of his or her transcript was comprised of courses defined as liberal arts in the CIP system; the remainder we called *CTE students*, although some of them had substantial liberal arts content in their transcripts (up to 75%); as a result, we found some liberal arts clusters for these students.

For liberal arts students, we did three clusterings, into 5, 10, and 15 clusters each. For CTE students, we did four clusterings, into 10, 15, 20, and 25 clusters each. These ranges of numbers of clusters were an attempt to find the appropriate number in each group, based on our knowledge of how many significant programs there were likely to be in the system (see Appendix A for further details about this choice). Below, when we examine the results, we attempt to find a good balance between too much detail and too little.

In our sample of first-time college students, there were 13,337 CTE students and 5,610 liberal arts students (based on the 75% criterion). For each group of clusterings, we looked for a clustering result that seemed to give the best balance between differentiating students who were studying different subjects and not separating students who were similar into two different clusters. The technique is not perfect, but it does give informative results. We first describe the results of clustering the CTE students; then we look at the results of clustering the liberal arts students.

3. Results and Analysis

3.1 Results of Clustering the CTE Student Transcripts

The results of the CTE clusterings are summarized in Table 1. The descriptions are based loosely on the official descriptions of the CIP codes and only include the top courses in each cluster. Note that the mappings are rough because the clusters are not the same across clusterings based on differing choices for the number of clusters. We have matched each clustering as best as possible to the most differentiated set of clusters

present at the 25-cluster level. The rest of the courses are in the order provided by the clustering algorithm at the 25-cluster level, which is arbitrary.

Table 1
CTE Cluster Contents and Cluster Presence by Number of Clusters

Description of Cluster Contents	10	15	20	25
Physical education emphasis with liberal arts[a]	x	x	x	x
Physical education, but with more liberal arts				x
Physical education, liberal arts		x	x	x
Liberal arts with some physical education		x	x	x
Accounting, math, economics	x	x	x	x
Accounting, economics, math				x
Liberal arts, personal awareness, mental health	x	x	x	x
Business assisting, computer operations	x	x	x	x
Vehicle maintenance, precision auto repair	x	x	x	x
Liberal arts, personal awareness, physical education				x
Microcomputer applications/Engineering technologies	x	x	x	x
Computer networking/Data processing/Computer programming			x	x
Precision metal working		x	x	x
Industrial production	x	x	x	x
Dental support/Medical administration				x
Early childhood education			x	x
Criminal justice				x
Nursing		x	x	x
Business administration			x	x
Design, fine arts, computer software			x	x
Parenting education		x	x	x
Real estate	x	x	x	x
Cosmetology			x	x
Culinary arts	x	x	x	x
Allied heath/Medical assisting	x	x	x	x

[a]The rows in italics indicate pairs of clusters that are near duplicates (courses with essentially the same content).

Table 1, which is based only on the CTE subsample, reveals the fruitfulness of this technique. Unlike techniques that select students' majors based on the single CIP in which they have taken the most courses, this method is able to detect patterns of course-taking across subject areas. For instance, the accounting students took courses in CIP 5203 (accounting), as well as in 4506 (Economics) and 2701 (Mathematics). The latter two CIPs are considered to be liberal arts while the former one is not. But only by

looking at the overall course-taking pattern can the courses included in a program of study be understood.

The same observation holds even for more purely vocational programs. For example, the 434 students in business assisting, in the 25-cluster version, took 31% of their coursework in business assisting (CIP 5204), but they also took 10% of their coursework in microcomputer applications (CIP 1106) and 8% of their coursework in health and medical administrative services (CIP 5107).

Clustering also allows us to disaggregate programs of study that might otherwise be lumped together. For instance, under the general category of business, accounting, business assisting, business administration, and real estate are distinct programs that the clustering method is able to pick out based solely on student activity; without knowing student activity, it would be difficult to recognize that these distinct programs are of interest to researchers and administrators. In other words, all of these are business programs, but short of a manual examination of transcripts or some prior knowledge of program requirements, it would be difficult to determine that these were the particular subjects that this cohort of students were studying within business. The clustering identified them for us.

Clustering also can focus attention on a group of students who are studying a very specific subject, such as real estate, cosmetology, or culinary arts, provided that they represent a significant percentage of students in the system. This has a tendency to mix levels of specificity or resolution among programs of study because clustering has some tendency to put like transcripts together whether they are brought together by a single specific CIP code (as is largely the case for these three specific examples), or if they are brought together by a pattern of course-taking, as is the case with the nursing or accounting clusters above, which each consist of both courses in CIP codes specific to these two subjects as well as courses in other CIPs. Generally, those clusters in which a broader range of CIPs are represented logically involve a broader range of studies by the students in them.

Of these four clusterings, we believe that the 20-cluster version makes the best balance between not combining majors that should be distinct and combining clusters that actually represent the same major. For instance, there are two clusters for the accounting

major in the 25-cluster version; these are combined in the 20-cluster version. It also eliminates one of the liberal arts/physical education clusters.[2]

When we move to 15 clusters, cosmetology is combined with culinary arts, and all the computer topics are combined. The computer networking students are combined with the microcomputer applications students. Business administration is combined with business assisting and medical administration. When we go to 10 clusters, the clusters are more aggregated. Note that reducing the number of clusters does not involve, necessarily, combining clusters from a higher level, because the students are assigned differently to clusters in each clustering.

3.2 The Twenty-Cluster Version in More Detail

Now that we have selected the 20-cluster version as the best one for descriptive purposes, let us examine it in more detail. The breakdown of course enrollments by two- (CIP2) and four-digit CIP code (CIP4) by cluster is given in Appendix B.

From the 20-cluster analysis, we can see that the course-taking patterns of students in each cluster are quite distinct.

Consider, for instance, the accounting students, in cluster 5. At the CIP4 level, they took 13% of their courses in accounting (CIP 5203), but they also took 7% in math, 6% in economics, and 5% in English. Looking at the two-digit level, they also took 24% of their courses in business (which includes the 13% in accounting). Clearly, the accounting students took a great number of courses outside of the aforementioned subjects as well, and the cluster brings together students who were diverse in terms of these additional subjects.

At the CIP4 level, the business assisting students (cluster 6) took 29% of their courses in business assisting, 9% in computer classes, 9% in health administrative assisting, and 6% in accounting.

The auto repair students (cluster 7) took a very focused program: 74% of their courses were in auto repair. This was also true of the precision metal working students (cluster 11): 63% of their courses were in this field. A similar pattern held for real estate

[2] There were still a few liberal arts clusters, since the criterion that 75% of one's courses be in liberal arts is a stringent one; physical education is not considered part of liberal arts, so students who took many physical education courses may not have reached that 75% mark.

(cluster 17), parenting education (cluster 15), culinary arts (cluster 20), and cosmetology (cluster 19).

However, the story is quite different for the nursing students (cluster 10). They took 35% of their courses in nursing and 6% in allied health. But looking at the two-digit level, they also took 13% of their courses in biology, as well as significant numbers of courses in other subjects like English and psychology. These nursing students were a mix of students who were training to be nurse's aides and those who were training to be registered nurses, because the four-digit CIP of 51.16 contains both of these.

The business administration students (cluster 14) are similar to the nursing students in that their program was composed of a relatively diverse set of subjects. Looking at the CIP2 level, 44% of their courses were in business. But they also took 11% of their courses in computers and 8% in English.

If one wants to get an even more detailed look at these clusters, one could also look at the most frequent courses taken by the students in each cluster. We do not do that herein because Washington State currently lacks a system-wide course numbering system, and we are looking at system-wide data. But doing so could be a useful tool when looking at the students at a single school or in a system that does have a uniform course numbering system (such as, for instance, the Virginia Community College System).

3.3 Comparing the 20 CTE Clusters with Jenkins and Weiss's Concentrations

Table 2 shows how each of the 20 mainly CTE clusters decompose into Jenkins and Weiss's (2011) attempted concentrations. The table shows the concentrations accounting for most of the clusters, listing at least three in each case. As is shown, there are a number of differences. For instance, in at least six cases—accounting, auto repair, computer networking, early childhood education, parenting education, and real estate—the clustering method found groups that were more specific than the corresponding groups assigned to most of the students by the concentrator method. For the real estate category, students were almost all assigned to business and marketing by the concentrator method, but our description of the cluster found that almost all of their courses were in the more specific field of real estate within business. In two cases, cosmetology and culinary arts, the students found were virtually the same, but the concentrator method had

to be aware of and had to look for these particular concentrations specifically (i.e., it had to include them as a categories of interest), while the clustering method found them automatically. In their study, Jenkins and Weiss used their knowledge of the Washington system to select and detect these concentrations.

Table 2
The 20 CTE Clusters, Decomposed into Jenkins and Weiss's (2011) Concentrations

Cluster	Concentration	N	Share	Cumulative share
Physical education	Arts, humanities, and English	629	44%	44%
Physical education	Mathematics and science (STEM)	134	9%	54%
Physical education	Social and behavioral sciences	119	8%	62%
Physical education	Other career-technical	106	7%	69%
Liberal arts/P.E.	Arts, humanities, and English	363	33%	33%
Liberal arts/P.E.	Not assigned	195	18%	50%
Liberal arts/P.E.	Social and behavioral sciences	149	13%	64%
Liberal arts/P.E.	Allied health	85	8%	71%
Liberal arts/Psychology	Arts, humanities, and English	442	30%	30%
Liberal arts/Psychology	Not assigned	184	13%	43%
Liberal arts/Psychology	Social and behavioral sciences	131	9%	52%
Liberal arts/Psychology	Allied health	119	8%	60%
Liberal arts/Psychology	Mathematics and science (STEM)	109	7%	67%
Liberal arts/Psychology	Protective services	89	6%	73%
Liberal arts/P.E.	Arts, humanities, and English	338	44%	44%
Liberal arts/P.E.	Mathematics and science (STEM)	108	14%	57%
Liberal arts/P.E.	Social and behavioral sciences	81	10%	68%
Accounting	Arts, humanities, and English	462	31%	31%
Accounting	Business and marketing	316	21%	52%
Accounting	Not assigned	183	12%	64%
Business assisting	Secretarial and administrative services	154	29%	29%
Business assisting	Allied health	85	16%	45%
Business assisting	Business and marketing	80	15%	60%
Business assisting	Computer and information sciences	77	14%	74%
Auto repair	Mechanics and repair	534	99%	99%
Auto repair	Not assigned	3	1%	99%
Auto repair	Engineering/science technologies	2	0%	99%
Computers/Design	Computer and information sciences	145	20%	20%
Computers/Design	Engineering/science technologies	98	13%	33%
Computers/Design	Allied health	70	9%	42%
Computers/Design	Business and marketing	64	9%	51%

Cluster	Concentration	N	Share	Cumulative share
Computers/Design	Construction	63	9%	59%
Computers/Design	Not assigned	59	8%	67%
Computers/Design	Agriculture and natural resources	53	7%	74%
Computer networking	Computer and information sciences	293	57%	57%
Computer networking	Not assigned	65	13%	69%
Computer networking	Engineering/science technologies	42	8%	77%
Nursing	Nursing	226	37%	37%
Nursing	Mathematics and science (STEM)	138	23%	60%
Nursing	Not assigned	77	13%	72%
Precision metal working	Manufacturing	370	88%	88%
Precision metal working	Mechanics and repair	19	5%	92%
Precision metal working	Engineering/science technologies	8	2%	94%
Industrial production	Engineering/Science technologies	176	38%	38%
Industrial production	Construction	77	17%	55%
Industrial production	Agriculture and natural resources	31	7%	61%
Industrial production	Allied health	27	6%	67%
Industrial production	Mechanics and repair	18	4%	71%
Industrial production	Other career-technical	17	4%	75%
Early childhood education	Education and child care	248	76%	76%
Early childhood education	Arts, humanities, and English	39	12%	88%
Early childhood education	Not assigned	16	5%	92%
Business administration	Business and marketing	332	59%	59%
Business administration	Arts, humanities, and English	42	7%	67%
Business administration	Allied health	35	6%	73%
Parenting education	Education and child care	197	88%	88%
Parenting education	Arts, humanities, and English	10	4%	92%
Parenting education	Allied health	3	1%	94%
Physical education	Arts, humanities, and English	452	44%	44%
Physical education	Social and behavioral sciences	185	18%	62%
Physical education	Mathematics and science (STEM)	131	13%	75%
Real estate	Business and marketing	212	98%	98%
Real estate	Not assigned	1	0%	99%
Real estate	Allied health	1	0%	99%
Allied health/Medical assisting	Allied health	398	91%	91%
Allied health/Medical assisting	Not assigned	14	3%	94%
Allied health/Medical assisting	Mathematics and science (STEM)	7	2%	96%
Cosmetology	Cosmetology	227	98%	98%
Cosmetology	Arts, humanities, and English	1	0%	99%

Cluster	Concentration	N	Share	Cumulative share
Cosmetology	Communications and design	1	0%	99%
Culinary arts	Culinary services	210	97%	97%
Culinary arts	Arts, humanities, and English	2	1%	98%
Culinary arts	Construction	1	0%	98%

3.4 Demographics of the Twenty CTE Clusters

Table 3 below shows the demographics of the 20 CTE clusters. As we can see, there is substantial variation across the clusters on the four demographics shown, which are percent falling in the two lowest SES quintiles, percent female, percent White, and mean age. Gender is the variable that shows the most variation.

Table 3
Demographics of the 20 CTE Clusters

Cluster	% Low SES	% Female	% White	Mean age
Physical education (1)	43%	47%	72%	20
Liberal arts/P.E. (1)	41%	57%	71%	21
Liberal arts/Psychology	35%	53%	71%	22
Liberal arts/P.E. (2)	28%	57%	71%	20
Accounting	34%	47%	65%	22
Business assisting	48%	75%	66%	31
Auto repair	46%	5%	74%	22
Computers/Design	44%	42%	75%	27
Computer networking	32%	23%	78%	26
Nursing	44%	72%	67%	25
Precision metal working	46%	5%	80%	26
Industrial production	44%	28%	75%	27
Early childhood education	46%	94%	66%	25
Business administration	36%	65%	66%	28
Parenting education	25%	89%	71%	34
Physical education (2)	35%	40%	73%	19
Real estate	34%	62%	92%	48
Allied health/Medical assisting	39%	87%	60%	27
Cosmetology	40%	96%	67%	22
Culinary arts	32%	48%	69%	24

Looking first at the variation in SES, we can see that business assisting, auto repair, precision metal working, and early childhood education had the highest percentage of low-SES students. Industrial production is not far behind. However, two fields where one might not expect a high proportion of low-SES, nursing and

computers/design, did have a relatively high fraction of low-SES students. However, as we have noted, some of these nursing students were training to be nurse's aides, while some were training to be associate or bachelor's level registered nurses.

At the other end of the spectrum, we find that computer networking, culinary arts, liberal arts/psychology, liberal arts/P.E., and parenting education all had relatively low numbers of low-SES students. None of these are surprising; liberal arts is known to attract higher-SES students, and the parenting education courses were co-op courses that allow parents (largely mothers, which comprised 89% of this cluster) to get credit for helping out at their child's day care center.[3]

Turning to gender, we see that there is still substantial divergence between what male and female students studied, at least in this sample. Auto repair and precision metal working were 95% male. At the other extreme (disregarding the highly female parenting students), we have the cosmetology students, who were 96% female, followed by the early childhood education students, who were 94% female. It appears that there was an interaction between a program of study being relatively low SES and it being highly gendered, since of these four areas, only one was not low-SES (cosmetology, which had exactly as many lower-SES students, 40%, in its population as existed in the population as a whole). Other fields that were notably female were nursing, business assisting, and allied health/medical assisting. Other fields were more balanced. Generally, the liberal arts fields were more balanced, including the physical education fields, as was accounting.

The liberal arts fields and the physical education courses were dominated by younger students; the other CTE fields were comprised of somewhat older students, generally, although this composition varied somewhat; cosmetology and auto repair tended to have young students as well. Real estate had relatively older students (age 48 on average) and they tended to be White women. On race, other than real estate, the most notable outlier was precision metal working, which was 80% White and almost all men, as noted above. Despite years of effort to get women and minorities into fields like this one, we do not see much evidence of change, at least in this data.

[3] See, for instance, the webpage for the Parent Education program at Edmonds Community College, described at http://www.edcc.edu/pared/Parent%20Cooperative%20Preschool/

3.5 Credential Attainment and Transfer by Cluster

Table 4 shows credential attainment by cluster for the 20 CTE clusters. One should note, first off, that this was a population of students who completed at least four college-level classes, so the overall credential attainment rates were higher than they would have been if all students who enrolled in the system had been included (many of these students took very few classes, so it was difficult to classify them into a major).

It is clear from Table 4 that there is a large difference between fields in the type of credentials earned. The first four fields are not really CTE, and are oriented to associate degrees and transfer. Accounting students also mainly earned associate degrees and also were oriented toward transfer. Business assisting students had a high rate of credential attainment, and they mainly earned short certificates. Auto repair students earned associate degrees as well, but these were mainly terminal degrees (in that they are not designed to transfer to a baccalaureate program). Computer networking students earned credentials at all three levels. Nursing students mainly earned short credentials, indicating that they were mainly becoming nurse's aides. Precision metal working students mainly earned short credentials, as did industrial production students and early childhood education students. Business administration students earned credentials at each level and appeared to be somewhat oriented toward transfer. Real estate students earned almost no credentials. Allied health and medical assisting students mainly earned long certificates, although they earned significant numbers of short certificates and associate degrees as well. This was also true of the cosmetology students, although their associate degrees typically do not transfer. The culinary students also often earned long certificates and associate degrees, along with some short certificates; again, these do not typically transfer.

Table 4
Credential Attainment by Cluster for the 20 CTE Clusters, Within Three Years of Student Entry

Cluster	Short certificate	Long certificate	Associate degree	Any subbaccalaureate credential
Physical education	2%	2%	16%	19%
Liberal arts/P.E.	3%	3%	17%	20%
Liberal arts/Psychology	4%	4%	17%	22%
Liberal arts/P.E.	3%	0%	15%	17%
Accounting	3%	1%	24%	26%
Business assisting	24%	10%	11%	38%
Auto repair	9%	5%	27%	36%
Computers/Design	6%	10%	20%	32%
Computer networking	9%	6%	10%	20%
Nursing	24%	10%	6%	36%
Precision metal working	19%	9%	9%	33%
Industrial production	16%	2%	3%	20%
Early childhood education	13%	2%	9%	22%
Business administration	10%	11%	16%	32%
Parenting education	4%	0%	2%	4%
Physical education	1%	1%	21%	22%
Real estate	1%	0%	0%	1%
Allied health/Medical assisting	16%	28%	13%	53%
Cosmetology	11%	20%	8%	39%
Culinary arts	7%	10%	12%	25%

Table 5 shows the transfer rate for each of the clusters. The first four clusters had a relatively higher transfer rate, because they were more liberal-arts oriented, as did the physical education cluster further down in the table. The accounting cluster actually had the highest transfer rate of all. It is notable that the business administration students did not have a very high rate, despite the fact that one might expect that such courses would lead to a transfer pathway.

Table 5
Transfer Rate for the 20 CTE Clusters

Cluster	Transfer rate
Physical education	23%
Liberal arts/P.E.	15%
Liberal arts/Psychology	12%
Liberal arts/P.E.	21%
Accounting	28%
Business assisting	3%
Auto repair	4%
Computers/Design	3%
Computer networking	9%
Nursing	7%
Precision metal working	1%
Industrial production	2%
Early childhood education	6%
Business administration	5%
Parenting education	4%
Physical education	27%
Real estate	0%
Allied health/Medical assisting	3%
Cosmetology	2%
Culinary arts	3%

3.6 Results of Clustering the Liberal Arts Students

The results of clustering the liberal arts students is shown in Table 6. As the table shows, the 5,610 liberal arts students (with at least 75% liberal arts content in their transcripts) were clustered into 5, 10, and 15 clusters.

Here, we see a much less clean separation than we did in the case of the CTE clusters. This could be driven by the fact that liberal arts students tend to take similar classes: math, English, history, psychology, sociology, and so on. The most distinctive clusters within these sets are the music clusters and the science clusters. There is one music cluster at the 5-cluster and 10-cluster level; this breaks into two at the 15-cluster level. There are also three science clusters at the 15-cluster level, but only one at the 5-cluster level. At the 5-cluster level, there are two clusters that are very similar, containing English, math, and psychology coursework (noted by *2x* in the table.) In our judgment, the five clusters, since they contain one math and one science cluster as well as three other liberal arts clusters, represent the diversity of liberal arts coursework by this cohort of students reasonably well.

Table 6
Liberal Arts Cluster Contents and Cluster Presence by Number of Clusters

Description of cluster contents	5	10	15
English/History/Psychology		x	x
English/Philosophy/History	x		x
English/Speech/Sociology/Psychology		x	x
Biology/Chemistry/English			x
Math/Chemistry/Biology	x		x
Math/English/History		x	x
Math/Chemistry/Physics		x	x
Art/Romance languages			x
Art/English/Math/History		x	
History/English/Political science		x	x
English/Math/Sociology		x	x
English/Math/Psychology	2x	x	
English/Biology/Sociology		x	x
English/Physical education/Philosophy		x	x
Music	x	x	x
Music			x

The details of the content of these five clusters is given in Appendix C. Looking at these details, we can see that the music students took over half of their courses in music, but that they were not a large group: just 5%. The science students were a larger group, and they took 11% of their courses in chemistry, 11% in biology, and 9% in math. They do not appear to have taken much physics. The remaining three clusters are somewhat generic; two are characterized by English, math, psychology, and history, with one containing some physical education, and the other some chemistry. There is also one that that is characterized by history, English, philosophy, speech, and physical education.

3.7 Demographics of the Five Liberal Arts Clusters

Table 7 shows the demographics of students in the five liberal arts clusters. There is not as much variation here as we saw among the CTE clusters. All of them contained young students, as is characteristic of liberal arts students at community colleges generally. All of them were higher SES than average, and all were predominantly White, although the chemistry/biology cluster was slightly less so. The music cluster was primarily male, and the chemistry/biology cluster was 60% women; some of the women were likely pursuing health careers. The remaining three clusters were roughly evenly split between men and women. The second and fourth clusters were particularly close and could probably be combined.

Table 7
Demographics of the Five Liberal Arts Clusters

Cluster	% Low SES	% Female	% White	Mean age
History/Philosophy/Speech	27%	51%	76%	20
English/Math/Psychology (1)	23%	51%	72%	19
Chemistry/Biology	28%	60%	63%	21
English/Math/Psychology (2)	35%	54%	72%	20
Music	24%	27%	74%	20

3.8 Credential Attainment and Transfer for the Five Liberal Arts Clusters

Table 8 shows credential attainment for the students in the five liberal arts clusters. As would be expected, this group of transfer-oriented liberal arts students earned hardly any certificates. The science and music students appear to have earned associate degrees at the lowest rates, followed by the history/philosophy/speech cluster students, who earned at an intermediate rate, and the two English/math/psychology clusters, which earned at the highest rate, but which is still relatively low, in the 30–40% range. However, it is interesting to see that even within groups of liberal arts students there is variation in completion rates based on what programs of study the students were undertaking, which is to be expected.

Table 8
Credential Attainment by Cluster for the Five Liberal Arts Clusters, Within Three Years of Student Entry

Cluster	Short certificate	Long certificate	Associate degree
History/Philosophy/Speech	0.4%	0.1%	21.4%
English/Math/Psychology (1)	0.2%	0.1%	31.4%
Chemistry/Biology	0.7%	0.0%	14.3%
English/Math/Psychology (2)	0.2%	0.0%	36.9%
Music	0.3%	0.0%	12.6%

Table 9 shows the transfer rate for the five clusters. The music and science clusters had the lowest rates, while the other three had higher rates. Thus, the data again indicate that students' fates were related to their programs.

Table 9
Transfer Rate by Cluster for the Five Liberal Arts Clusters

Cluster	Transfer rate
History/Philosophy/Speech	22%
English/Math/Psychology (1)	31%
Chemistry/Biology	14%
English/Math/Psychology (2)	37%
Music	13%

3.9 Comparing the Five Liberal Arts Clusters with Jenkins and Weiss's Concentrations

Table 10 shows how each of the five liberal arts clusters decompose into Jenkins and Weiss's (2011) attempted concentrations. The table shows the concentrations accounting for most of the clusters, listing at least three in each case. Note that there is overlap in classification, but it is far from perfect. For instance, only about half of the students in the science cluster were classified as science students by Jenkins and Weiss. On the other hand, virtually all of the music students were classified as arts, humanities, and English students by Jenkins and Weiss; the clustering method has detected their activity more specifically. The other three clusters were primarily classified as arts, humanities, and English students as well, although not exclusively.

Table 10
The Five Liberal Arts Clusters,
Decomposed into Jenkins and Weiss's (2011) Concentrations

Cluster	Concentration	N	Share	Cumulative Share
History/Philosophy/Speech	Arts, humanities, and English	757	63%	63%
History/Philosophy/Speech	Social and behavioral sciences	284	24%	87%
History/Philosophy/Speech	Not assigned	71	6%	93%
English/Math/Psychology	Arts, humanities, and English	1024	56%	56%
English/Math/Psychology	Mathematics and science (STEM)	376	20%	76%
English/Math/Psychology	Social and behavioral sciences	248	13%	90%
Chemistry/Biology	Mathematics and science (STEM)	611	51%	51%
Chemistry/Biology	Arts, humanities, and English	375	31%	82%
Chemistry/Biology	Social and behavioral sciences	104	9%	90%
English/Math/Psychology	Arts, humanities, and English	643	60%	60%
English/Math/Psychology	Mathematics and science (STEM)	216	20%	81%
English/Math/Psychology	Social and behavioral sciences	107	10%	91%
Music	Arts, humanities, and English	277	92%	92%
Music	Mathematics and science (STEM)	12	4%	96%
Music	Social and behavioral sciences	9	3%	99%

4. Conclusion

We can see from examining this clustering in some detail is that this is a powerful way to describe the course-taking patterns of students in a community college or a college system. It has the advantage of imposing very little in the way of prior assumptions about what is in the data; rather, it lets the data tell what is going on. It also relieves the analyst of the work of determining what the program concentration categories should be. Furthermore, clustering has the advantage of being able to detect *patterns* in student transcripts across different programs that go beyond single subjects. For instance, a nursing student may take biology, psychology, and computer courses as well as nursing courses. By detecting these patterns, it can do a better job of putting students together who have similar patterns of course activity. Note that while we have applied this method to community colleges, it is applicable to schools at all levels. Also, if the method is

applied to a single school or a system with uniform courses, the clustering can be done at the course level rather than at the CIP level.

The main disadvantage of this method is that clustering, particularly of data of this type, is, as we have said, as much of an art as a science; we put substantial effort both into filtering students to feed into the clustering algorithm and into examining several distinct clusterings to find one that seemed to fit the pattern of activity well. Yet, we have seen from the analysis here that there were often substantial differences in the student demographics and credentials awarded among different program areas. In the absence of extensive information about program requirements or better information on student intent or declared major, it would be useful to have a tool that can provide a more in-depth understanding of students' programmatic pathways, so that we could see not only what coursework students are undertaking but also what types of students are undertaking the coursework. We believe that the results are quite fruitful and could be applied in many college settings by institutional researchers, administrators, and other interested faculty and staff.

In future work, we plan to look in more detail not only at the aggregate course-taking activity of students irrespective of time but also at the sequencing of this course-taking activity. Examining course sequences in aggregate will allow us to uncover typical pathways that students follow. We also plan to look more at the course-taking patterns of students who actually completed programs. By comparing the transcripts of completers (who are a minority in the typical community college) with non-completers, we may be better able to identify the programs that the non-completers had attempted to undertake. We may also be able to identify individual courses that pose obstacles to completion in particular programs, so that colleges can take steps to reduce these obstacles.

References

Bailey, T., Jenkins, D., & Leinbach, D. T. (2006). *Is student success labeled institutional failure? Student goals and graduation rates in the accountability debate at community colleges* (CCRC Working Paper No. 1). New York, NY: Columbia University, Teachers College, Community College Research Center.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3), 803–821.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4-5), 993–1022.

Hartigan, J. A. (1978). Asymptotic distributions for clustering criteria. *Annals of Statistics*, *6*(1), 117–131.

Jacobson, L., & Mokher, C. (2010, May). *The effect of career and technical education on employment and earnings: Evidence from Florida.* Paper Presented at the Annual Conference of the American Educational Research Association, Denver, CO.

Jenkins, D., & Weiss, M. J. (2011). *Charting pathways to completion for low-income community college students* (CCRC Working Paper No. 34). New York, NY: Columbia University, Teachers College, Community College Research Center.

Kaufman, L., & Rousseeuw, P. J. (2008). Partitioning around medoids (Program PAM). In *Finding groups in data: An introduction to cluster analysis* (pp. 68–125). Hoboken, NJ: John Wiley & Sons.

Mitchell, T. (1997). *Machine learning*. New York, NY: McGraw-Hill.

Stuart, G. R. (2009). *A benefit/cost analysis of three student enrollment behaviors at a community college: Dropout, transfer and completion of an associate's degree/certificate* (Doctoral dissertation). Retrieved from http://etd.ohiolink.edu

Appendix A: Details of the Cluster Analysis

A fairly comprehensive overview of the clustering methods we used is given in the main text. The purpose of this appendix is to provide additional details regarding the decision process that led to the choices we made.

A.1 Choice of Clustering Method

Clustering is a statistical method for organizing a collection of objects under study (in our case, students) into meaningful groups. The assignment to groups is based on some notion of similarity: two students in the same group should be more alike than two students assigned to different groups. Clustering procedures are designed to maximize the separation between groups and minimize the separation between subjects within groups.

Roughly speaking, there are three types of clustering methods: agglomerative, partitioning, and model-based. Agglomerative clustering builds the assignment from the ground up: students are initially placed in their own unique cluster consisting just of themselves. Two of these clusters are joined if they are nearest each other as compared to all other pairs of existing clusters. This process is repeated until the current set of clusters meets some objective criteria, such as maximizing the ratio of between- versus within-sum-of-squared distances between subjects. There are many variants of agglomerative clustering, mostly involving the choice of which two clusters to merge at any given step. The second approach, partitioning, begins with some (often random) assignment of subjects to clusters and then tries to improve the partition by moving subjects in and out of their current cluster assignment. If moving a subject to a different cluster improves a global measure of fit more than any other move, then that move is chosen. The most commonly used partitioning algorithm is called *k*-means (Hartigan, 1978), and it is favored because it is statistically consistent.[4] Model-based clustering (MBC) (Banfield & Raftery, 1993) begins by assuming that the data are generated from a pre-specified distribution, typically a mixture of multivariate normal distributions. Essentially, the assignment to a cluster is a latent variable, or in this case, a latent membership in a specific class. MBC first identifies the parameters governing the process, such as the

[4] In this context, this means that if the data are derived from *k groups* with certain properties, the *k-means* algorithm will correctly identify the means of each of those k groups, given enough observations.

means and covariances of the multivariate normal densities. These parameters are typically chosen using the method of maximum likelihood: the parameters are the ones most likely to have given rise to the observations obtained. Subjects are then assigned to the group to which they are most likely to belong, given the parameters governing the groups.

For our application, we used a partitioning algorithm known as partitioning around medoids (PAM) (Kaufman & Rousseeuw, 2008), discussed below. We considered using an agglomerative clustering method, but decided that the results would be too dependent on the many choices that are required by this approach. Moreover, those choices are often based on intuition about how well separated the subjects are from one another. Since the transcript, or set of courses, is the feature set being compared, we lack a visualization technique to gauge the extent to which students are more or less similar to one another. There simply is no natural metric for comparing collections of course choices (the courses are nominal data). We ruled out model-based clustering because there are very few models for collections of non-numeric choices. A notable exception involves the classification of documents using variants of the *bag of words* model (see Blei, Ng, & Jordan, 2003). To cross-check our findings, we compared them to results obtained from a customized, model-based approach akin to the classification model mentioned above, and the resulting partitions were similar.[5]

PAM is known as a *k-medoids* method, and is quite similar to the *k*-means algorithm. The idea behind PAM is that clusters may be represented by a single, central, representative element. The medoid is the element in the cluster that is central in the sense that the average distance from it to every other element in the group is smaller than what would be obtained using a different element. PAM, like *k*-means, searches through possible partitions until an overall objective criterion is minimized. This criterion and thus the method rely heavily on the choice of similarity measure, to which we now turn.

A.2 Choosing the Similarity Measure

Agglomerative and partitioning methods both rely on being able to measure the similarity between any two subjects. If the characteristics of these subjects are numeric,

[5] Borrowing the conditional independence assumptions used in naïve-Bayes classifiers, we built an analogous model-based clustering model based on a mixture model (details omitted).

such as age and height, then one can measure the distance between pairs of characteristics (Euclidean distance, e.g.), and similarity is just the complement of distance. If distance is between 0 and 1, then similarity could be defined as *1 minus distance*.

As discussed in the main text, collections of courses contained in transcripts are harder to compare. In this analysis, we constructed a similarity measure in several stages. First, in a pairwise manner, we counted the number of courses that are found on both transcripts. Call this the agreement count between students 1 and 2, A_{12}. For the first student in the pair, there may be courses that cannot be found in the second student's transcript. Call this the disagreement count for student 1, B_{12}. The corresponding measurement for student 2 we will call B_{21}. We constructed a similarity measurement from these three numbers in several different ways. First we took the average fraction of agreement from two perspectives (student 1 or student 2):

$$SimAvg_{12} = \text{avg}\left\{\frac{A_{12}}{A_{12}+B_{12}}, \frac{A_{12}}{A_{12}+B_{21}}\right\}.$$

Next, we considered maximum and minimum agreement:

$$SimMax_{12} = \max\left\{\frac{A_{12}}{A_{12}+B_{12}}, \frac{A_{12}}{A_{12}+B_{21}}\right\}$$

$$SimMin_{12} = \min\left\{\frac{A_{12}}{A_{12}+B_{12}}, \frac{A_{12}}{A_{12}+B_{21}}\right\}.$$

These three measures range from a conservative measure using the minimum to a more liberal measure using the maximum. The average proportion agreement lands somewhere in the middle of these two extremes, and as such was our preferred measure. We ran our clustering algorithm using all three similarity measures and found them to yield comparable clusterings, with the average similarity measure performing somewhat better with respect to the criterion we wished to maximize. As a reminder, distance, or dissimilarity, is taken as 1 minus similarity.

A.3 Details of the PAM Algorithm

While *k*-means clustering minimizes the sum of the squared Euclidean distances between observations and their cluster mean, PAM minimizes the sum of *dissimilarities*

of the observations with all other observations in their assigned cluster. Like *k*-means, PAM requires the researcher to choose the number of clusters, *k*. We discuss our procedure for making that choice below. We first discuss how the PAM algorithm assigns medoids and subjects to clusters.

There is an initial build phase in which *k* "centers" or medoids are assigned. As mentioned above, the medoid is a specific observation that is most similar to a subset of the other observations. The initial set of medoids is not quite randomly chosen, but some random choices are made in their determination, and thus the algorithm may converge to slightly different solutions depending on the initial values of a random number generator. For this reason, we ran our clustering algorithm several times using a different initial random number seed and compared the findings. Next, there is a swap phase, in which alternative choices for medoids are examined. The algorithm will swap a medoid choice if that swap improves the total similarity of all subjects to their assigned cluster. Note that changing a medoid changes the subject cluster assignment, which we now describe. With a set of *k* assigned medoids, each observation is assigned to a group based on the nearest medoid. So an observation may be closest to group 1, but its average dissimilarity with observations in group 1 is actually larger than its average dissimilarity with observations in group 2. In this case, it simply is not close enough to the medoid of group 2 to be placed in that cluster. This process of reassigning medoids and then reassigning cluster membership continues until there is no change in the objective function. Upon completion, each subject is assigned uniquely to one of *k* groups.

A.4 Choosing the Number of Clusters

A useful measure of fit for a clustering solution is the average *silhouette width*, the latter being a measure of fit for a single observation that we now describe. For each observation, we can evaluate its average dissimilarity, now labeled *d* in our formulas, from every other point in the cluster. We will write $d(i,l)$ to denote the dissimilarity between observations *i* and *l*, both in the m^{th} cluster, named C_m. Then let:

$$a(m,i) = \frac{1}{n_m - 1} \sum_{l \in C_m \setminus i} d(i,l).$$

In words, $a(m,i)$ is the average dissimilarity of point i in the m^{th} cluster from all remaining points in that cluster. For any other cluster, C, we can take that same point i in the m^{th} cluster and evaluate its average dissimilarity from all the points in cluster C. We call this measurement D:

$$D(C,m,i) = \frac{1}{n_C}\sum_{j \in C} d(i,j).$$

In words, $D(C,m,i)$ is the average distance of point i in m^{th} cluster C_m from every point j in cluster C. Next, we find the cluster C (excluding C_m) that is "closest" to point i in cluster m, and call the associated average dissimilarity b:

$$b(m,i) = \min_{c \neq m} D(c,m,i).$$

Here, $b(m,i)$ is the average distance of point i in the m^{th} cluster from every point in the *nearest* cluster to i other than the cluster it is in. The idea behind the silhouette width measurement is to compare $a(m,i)$ and $b(m,i)$. Homogenous clusters have large $a(m,i)$ relative to $b(m,i)$, because the nearest cluster is far away. The *silhouette width* at point i in cluster C_m is defined as:

$$SW_i = (b(m,i) - a(m,i)) / \max(a(m,i), b(m,i)).$$

This rescales the measure to be between -1 and 1. Based on the prior discussion, we know that placement of a subject in a cluster depends on the nearest medoid, not the silhouette width. The average dissimilarity of a subject with the *second nearest* medoid may actually be smaller than that with its nearest medoid. Thus, silhouette width can be negative, indicating poor fit of that observation. Average silhouette width averages these single goodness of fit measures across the entire clustering. It may seem counterintuitive, but negative silhouette width for an observation does not imply that we should move that single, poorly fitting observation, as this might worsen the average fit in the new cluster and thus the overall measure.

While silhouette width can be used to assess fit for both individuals, clusters, and the whole ensemble of clusters, it can also be used to select the number of clusters. We

assess average silhouette width for a series of different clustering solutions, such as k = 5, 10, 15 , 20, 25, 30, and we increase the increment between choices once we reach 100. In many clustering problems, the silhouette width increases steadily as k increases, but then it levels off and then declines. The intuition behind this pattern is that we are discovering structure in the data up to a point, after which we are actually splitting up groups that ought to be together. For example, clustering might reveal a biological sciences group for smaller k, but as k increases, a nursing cluster splits off. However, for very large k, the nursing cluster may be further divided into those who pursued the degree part time (taking fewer courses) and those who did not. Unfortunately, the average silhouette criterion applied to Washington state transcript data using the average agreement measure did not peak for well over k = 100 clusters. We took this as a signal that transcript data—at least at community colleges—is very heterogeneous and thus hard to separate into distinct groups (we might find transcripts at four-year competitive colleges fairly easy to cluster, e.g.). Nevertheless, one can form any number of clusters, and these will represent our best attempt at forming that number of distinct groups. Thus, in this study we examined a smaller number of clusters than one might choose if one were only interested in maximally separating the groups. Put another way, we were not interested in discovering subtle differences between students pursuing degrees in nursing part time versus full time.

Given some of the technical challenges associated with clustering transcript data, we cross-checked our cluster solutions with those obtained via an implementation of model-based clustering using a naïve Bayes, conditional independence simplifying assumption (Mitchell, 1997). We used these results to inform the PAM analysis. In particular, we assessed the number of clusters using this alternative approach. The model-based approach uses a large number of parameters to represent different clusters, so there was a tradeoff between fit and parsimony. The model-based approach favored on the order of 30 clusters, and thus our own choices that were in that range seem appropriate.

Appendix B: Twenty Clusters of CTE Student Transcripts

Cluster 1. Physical education: 1425 students (11%); mean transcript length: 16.
Top 5 CIP4s: (3105, Health and physical education/Fitness: 20%); (2301, English language and literature, General: 7%); (2310, Speech and rhetorical studies: 3%); (2701, Mathematics: 3%); (4201, Psychology, general: 3%)
Top 5 CIP2s: (31, Parks, recreation, leisure, and Fitness studies: 20%); (23, English language and literature/Letters: 13%); (50, Visual and performing arts: 7%); (45, Social sciences: 7%); (51, Health professions and related clinical Sciences: 5%).

Cluster 2. Liberal arts, Physical education: 1109 students (8%); mean transcript length: 15.
Top 5 CIP4s: (2301, English language and literature, general: 8%); (3105, Health and physical education/Fitness: 7%); (4201, Psychology, general: 5%); (3701, Personal awareness and self-improvement: 5%); (4511, Sociology: 5%)
Top 5 CIP2s: (23, English language and literature/Letters: 13%); (51, Health professions and related clinical sciences: 11%); (45, Social sciences: 8%); (31, Parks, recreation, leisure, and fitness studies: 7%); (50, Visual and performing arts: 7%).

Cluster 3. Liberal Arts, Psychology 1467 students (11%); mean transcript length: 14.
Top 5 CIP4s: (2304, English composition: 7%); (3701, Personal awareness and self-improvement: 5%); (2310, Speech and rhetorical studies: 5%); (4201, Psychology, general: 5%); (5106, Dental support services and applied professions: 3%).
Top 5 CIP2s: (23, English language and literature/Letters: 15%); (51, Health professions and related clinical sciences: 11%); (11, Computer and information sciences and support services: 8%); (42, Psychology: 6%); (45, Social sciences: 5%).

Cluster 4. Liberal arts, physical education: 777 students (6%); mean transcript length: 15.
Top 5 CIP4s: (2304, English composition: 12%); (3105, Health and physical education/Fitness: 9%); (4201, Psychology, general: 5%); (2701, Mathematics: 4%); (5116, Nursing: 3%).
Top 5 CIP2s: (23, English language and literature/Letters: 15%); (31, Parks, recreation, leisure, and fitness studies: 9%); (51, Health professions and related clinical sciences: 9%); (50, Visual and performing arts: 8%); (45, Social sciences: 7%).

Cluster 5. Accounting: 1492 students (11%); mean transcript length: 14.
Top 5 CIP4s: (5203, Accounting and related services: 13%); (2701, Mathematics: 7%); (4506, Economics: 6%); (2301, English language and literature, general: 5%); (3105, Health and physical education/Fitness: 4%).
Top 5 CIP2s: (52, Business, management, marketing, and related support services: 24%); (23, English language and literature/Letters: 14%); (45, Social sciences: 11%); (27, Mathematics and statistics: 10%); (50, Visual and performing arts: 7%).

Cluster 6. Business Assisting: 536 students (4%); mean transcript length: 14.
Top 5 CIP4s: (5204, Business operations support and assistant services: 29%); (1106, Data entry/Microcomputer applications: 9%); (5107, Health and medical administrative services: 9%); (5203, Accounting and related services: 6%); (5202, Business administration, management and operations: 5%).
Top 5 CIP2s: (52, Business, management, marketing, and related support services: 43%); (11, Computer and information sciences and support services: 16%); (51, Health professions and related clinical sciences: 13%); (23, English language and literature/Letters: 7%); (37, Personal awareness and self-improvement: 3%).

Cluster 7. Auto repair: 542 students (4%); mean transcript length: 18.
Top 5 CIP4s: (4706, Vehicle maintenance and repair technologies: 74%); (4805, Precision metal working: 4%); (2399, English language and literature/Letters, other: 3%); (2799, Mathematics and statistics, other: 2%); (5109, Allied health diagnostic, intervention, and treatment professions: 2%).
Top 5 CIP2s: (47, Mechanic and repair technologies/Technicians: 74%); (23, English language and literature/Letters: 6%); (48, Precision production: 4%); (27, Mathematics and statistics: 3%); (11, Computer and information sciences and support services: 2%).

Cluster 8. Computers/Design 741 students (6%); mean transcript length: 15.
Top 5 CIP4s: (1106, Data entry/Microcomputer applications: 11%); (2399, English language and literature/Letters, Other: 6%); (5004, Design and applied arts: 6%); (2799, Mathematics and statistics, other: 5%); (4603, Electrical and power transmission installers: 3%).
Top 5 CIP2s: (11, Computer and information sciences and support services: 17%); (15, Engineering technologies/Technicians: 13%); (51, Health professions and related clinical sciences: 9%); (23, English language and literature/Letters: 8%); (50, Visual and performing arts: 7%).

Cluster 9. Computer networking: 517 students (4%); mean transcript length: 12.
Top 5 CIP4s: (1109, Computer systems networking and telecommunications: 19%); (1103, Data processing: 13%); (1102, Computer programming: 8%); (1503, Electrical engineering technologies/Technicians: 7%); (1110, Computer/Information technology administration and management: 7%).
Top 5 CIP2s: (11, Computer and information sciences and support services: 56%); (15, Engineering technologies/Technicians: 10%); (23, English language and literature/Letters: 8%); (27, Mathematics and statistics: 5%); (52, Business, management, marketing, and related support services: 5%).

Cluster 10. Nursing: 610 students (5%); mean transcript length: 15.
Top 5 CIP4s: (5116, Nursing: 35%); (5109, Allied health diagnostic, intervention, and treatment professions: 6%); (2604, Cell/Cellular biology and anatomical sciences: 6%); (2601, Biology, general: 3%); (2301, English language and literature, general: 3%).
Top 5 CIP2s: (51, Health professions and related clinical sciences: 47%); (26, Biological and biomedical sciences: 13%); (23, English language and literature/Letters: 8%); (42, Psychology: 6%); (40, Physical sciences: 4%).

Cluster 11. Precision metal working: 421 students (3%); mean transcript length: 12.
Top 5 CIP4s: (4805, Precision metal working: 63%); (1513, Drafting/Design engineering technologies/Technicians: 4%); (2799, Mathematics and statistics, other: 3%); (2399, English language and literature/Letters, other: 2%); (3701, Personal awareness and self-improvement: 2%).
Top 5 CIP2s: (48, Precision production: 65%); (15, Engineering technologies/Technicians: 7%); (23, English language and literature/Letters: 5%); (47, Mechanic and repair technologies/Technicians: 4%); (27, Mathematics and statistics: 4%).

Cluster 12. Industrial production: 462 students (3%); mean transcript length: 11.
Top 5 CIP4s: Industrial production: (1506, Industrial production technologies/Technicians: 36%); (0106, Applied horticulture and horticultural business services: 5%); (1513, Drafting/Design engineering technologies/Technicians: 4%); (4699, Construction trades, Other: 3%); (5214, Marketing: 3%).
Top 5 CIP2s: (15, Engineering technologies/Technicians: 42%); (46, Construction trades: 11%); (52, Business, management, marketing, and related support services: 6%); (01, Agriculture, agriculture operations, and related sciences: 5%); (51, Health professions and related clinical sciences: 5%).

Cluster 13. Early childhood education: 328 students (2%); mean transcript length: 13.
Top 5 CIP4s: (1312, Teacher education and professional development, specific levels and methods: 49%); (2002, Child care and guidance workers and managers: 5%); (1907, Human development, family studies, and related services: 4%); (2304, English composition: 3%); (4201, Psychology, general: 2%).
Top 5 CIP2s: (13, Education: 52%); (23, English language and literature/Letters: 8%); (20, Vocational home economics: 5%); (19, Family and consumer sciences/Human sciences: 5%); (45, Social sciences: 4%).

Cluster 14. Business administration: 562 students (4%); mean transcript length: 15.
Top 5 CIP4s: (5202, Business administration, management and operations: 30%); (5203, Accounting and related services: 6%); (1106, Data entry/Microcomputer applications: 6%); (5107, Health and medical administrative services: 3%); (1103, Data processing: 3%).
Top 5 CIP2s: (52, Business, management, marketing, and related support Services: 44%); (11, Computer and information sciences and support services: 11%); (23, English language and literature/Letters: 8%); (51, Health professions and related clinical sciences: 7%); (15, Engineering technologies/Technicians: 4%).

Cluster 15. Parenting education: 224 students (2%); mean transcript length: 8.
Top 5 CIP4s: (2001, Consumer and homemaking education: 81%); (5116, Nursing: 2%); (3701, Personal awareness and self-improvement: 1%); (2301, English language and literature, general: 1%); (5107, Health and medical administrative services: 1%).

Top 5 CIP2s: (20, Vocational home economics: 81%); (51, Health professions and related clinical sciences: 4%); (23, English language and literature/Letters: 2%); (52, Business, management, marketing, and related support services: 2%); (13, Education: 1%).

Cluster 16. Physical education: 1024 students (8%); mean transcript length: 16.
Top 5 CIP4s: (3105, Health and physical education/Fitness: 13%); (5401, History: 6%); (2304, English composition: 6%); (2310, Speech and rhetorical studies: 5%); (2701, Mathematics: 5%).
Top 5 CIP2s: (23, English language and literature/Letters: 15%); (31, Parks, recreation, leisure, and fitness studies: 13%); (45, Social sciences: 8%); (50, Visual and performing arts: 7%); (54, History: 6%).

Cluster 17. Real estate: 216 students (2%); mean transcript length: 5.
Top 5 CIP4s: (5215, Real estate: 94%); (5207, Entrepreneurial and small business operations: 1%); (5202, Business administration, management and operations: 1%); (5203, Accounting and related services: 1%); (1106, Data entry/Microcomputer applications: 0%).
Top 5 CIP2s: (52, Business, management, marketing, and related support services: 97%); (23, English language and literature/Letters: 1%); (26, Biological and biomedical sciences: 0%); (45, Social sciences: 0%); (11, Computer and information sciences and support services: 0%).

Cluster 18. Allied health/Medical assisting: 436 students (3%); mean transcript length: 18.
Top 5 CIP4s: (5108, Allied health and medical assisting services: 53%); (5107, Health and medical administrative services: 13%); (5204, Business operations support and assistant services: 2%); (5106, Dental support services and applied professions: 2%); (2304, English composition: 2%).
Top 5 CIP2s: (51, Health professions and related clinical sciences: 72%); (23, English language and literature/Letters: 5%); (52, Business, management, marketing, and related support services: 4%); (26, Biological and biomedical sciences: 3%); (11, Computer and information sciences and support services: 2%).

Cluster 19. Cosmetology: 231 students (2%); mean transcript length: 15.
Top 5 CIP4s: (1204, Cosmetology and related personal grooming services: 81%); (2799, Mathematics and statistics, Other: 1%); (2399, English language and literature/Letters, Other: 1%); (3701, Personal awareness and self-improvement: 1%); (5214, Marketing: 1%).
Top 5 CIP2s: (12, Personal and culinary services: 81%); (23, English language and literature/Letters: 4%); (51, Health professions and related clinical sciences: 2%); (27, Mathematics and statistics: 2%); (50, Visual and performing arts: 2%).

Cluster 20. Culinary arts: 217 students (2%); mean transcript length: 21.
Top 5 CIP4s: (1205, Culinary arts and related services: 79%); (2399, English language and literature/Letters, Other: 2%); (2799, Mathematics and statistics, other: 2%); (5209, Hospitality administration/management: 1%); (4599, Social sciences, other: 1%).
Top 5 CIP2s: (12, Personal and culinary services: 79%); (23, English language and literature/Letters: 4%); (52, Business, management, marketing, and related support services: 3%); (27, Mathematics and statistics: 2%); (45, Social sciences: 2%).

Appendix C: Five Clusters of Liberal Arts Student Transcripts

Cluster 1. History/Philosophy/Speech: 1200 students (21%); mean transcript length: 13.
Top 5 CIP4s: (5401, History: 8%); (2301, English language and literature, general: 8%); (3801, Philosophy: 5%); (2310, Speech and rhetorical studies: 5%); (3105, Health and physical education/Fitness: 5%).
Top 5 CIP2s: (23, English language and literature/Letters: 20%); (45, Social sciences: 15%); (50, Visual and performing arts: 10%); (54, History: 8%); (40, Physical sciences: 7%).

Cluster 2. English/Math/Psychology: 1838 students (33%); mean transcript length: 16.
Top 5 CIP4s: (2304, English composition: 12%); (2701, Mathematics: 8%); (4201, Psychology, general: 5%); (5401, History: 5%); (4005, Chemistry: 4%).
Top 5 CIP2s: (23, English language and literature/Letters: 19%); (45, Social sciences: 11%); (40, Physical sciences: 10%); (50, Visual and performing arts: 10%); (27, Mathematics and statistics: 9%).

Cluster 3. Chemistry/Biology: 1206 students (21%); mean transcript length: 12.
Top 5 CIP4s: (4005, Chemistry: 11%); (2701, Mathematics: 9%); (2601, Biology, general: 6%); (2604, Cell/Cellular biology and anatomical sciences: 5%); (2304, English composition: 5%).
Top 5 CIP2s: (23, English language and literature/Letters: 15%); (40, Physical sciences: 14%); (26, Biological and biomedical sciences: 14%); (27, Mathematics and statistics: 12%); (45, Social sciences: 9%).

Cluster 4. English/Math/Psychology: 1065 students (19%); mean transcript length: 16.
Top 5 CIP4s: (2301, English language and literature, general: 12%); (2701, Mathematics: 9%); (4201, Psychology, general: 6%); (3105, Health and physical education/Fitness: 5%); (5401, History: 4%).
Top 5 CIP2s: (23, English language and literature/Letters: 18%); (45, Social sciences: 11%); (27, Mathematics and statistics: 10%); (50, Visual and performing arts: 10%); (40, Physical sciences: 9%).

Cluster 5. Music: 301 students (5%); mean transcript length: 17.
Top 5 CIP4s: (5009, Music: 53%); (2304, English composition: 4%); (3105, Health and physical education/Fitness: 3%); (2701, Mathematics: 3%); (2301, English language and literature, general: 2%).
Top 5 CIP2s: (50, Visual and performing arts: 57%); (23, English language and literature/Letters: 9%); (45, Social sciences: 5%); (40, Physical sciences: 4%); (31, Parks, recreation, leisure, and fitness studies: 3%).