Networks, Community Detection, and Robustness: Statistical Inference on Student Enrollment Data

by

Uriah Israel

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Applied Physics)
in The University of Michigan
2020

Doctoral Committee:

Professor Timothy A. McKay, Co-Chair
Professor Scott E. Page, Co-Chair
Assistant Professor Abigail Jacobs
Professor Rada Mihalcea

Uriah Israel

ulisrael@umich.edu

ORCID iD: 0000-0002-3203-3654

DEDICATION

Dedicated to the many people and organizations that supported me throughout my graduate experience.

ACKNOWLEDGEMENTS

I would like to acknowledge and give my most gracious gratitude towards my mentors, Scott E. Page, Timothy McKay, Benjamin Koester, Rada Mihalcea, and Abigail Jacobs. They went out of their way to help me develop as a student and researcher. Scott was my first mentor at the University of Michigan. He is amazing and there is absolutely no way I would be where I am now without his hard work and dedication. Even with his busy travel schedule, he always set aside time to work with me. Tim is responsible for establishing the foundation of this dissertation. He took me on as a graduate student and helped me to focus on progressing through grad school. I was able to publish my first paper as a grad student because of his support. He has been a supportive and patient mentor. Ben has worked with me throughout this entire dissertation project. He was in the trenches with me and helped me master a lot of the data analysis. He familiarized me with the LARC dataset and also helped me gain access to anything I needed in research. Rada has been great in my development as a data scientist. Even though I wasn't a computer science student, she brought me into her lab so that I can learn from her journal club and group meetings. Abbie is new to the University of Michigan, but since her arrival she has been nothing but helpful. She met with me multiple times a week to discuss research ideas and helped me finalize a lot of this thesis.

I am truly grateful for the endless support my parents, Angela and Immanuel, have provided me. Their support and guidance has played a crucial role in my self-development. I would also like to acknowledge my appreciation for all of my friends

who have helped this be an enjoyable experience.

Majority of my career as a graduate student has been within the Applied Physics Program and the Center for the Study of Complex Systems (CSCS). For Applied Physics, I am deeply indebted to Cynthia McNabb, Cagliyan Kurdak, Charles Sutton, and Lauren Segall. As for CSCS, I owe everything to Linda Wood and Mita Gibson. The amount of support I received from these programs exceeded my greatest expectations ten times over!

Finally, I'd like to acknowledge the Rackham Merit Fellowship, Imes-Moore Fellowship, Sloan Scholarship, and the teaching opportunities provided by Tim and Scott, that financially supported me throughout this journey.

Thank you everyone affiliated with the University of Michigan for a truly great graduate school experience!

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

At the heart of higher education is the student experience. The student experience represents the courses students take, the people they interact with, their extracurricular activities provided by the university and much more. Developing methods that are able to measure the student experience is important to many stakeholders. Having better measures of student experience will help university leaders such as presidents, provosts, deans, department chairs, and faculty design better curricula and allocate resources, it will give more context to students about the courses they select, and it help employers better understand the graduates that they will employee.

Previous attempts to quantify the experience are not comprehensive. For example, student surveys and evaluation forms only get a subset of the population. Accreditation committees only get a snapshot impression.

In this study we demonstrate how high resolution student enrollment data can be used to better quantify the student experience. The methods described in this thesis are not unique to the institution studied and are scalable. Thus, they can be applied at other institutions were student enrollment is recorded.

This thesis introduces a new dataset provided by the University of Michigan Information and Technology Services staff. This dataset contains information on student enrollment dating back to 2000. We demonstrate how this data is implicitly networked. The connections between students and courses are explored and analyzed by employing methods common to network science. Student enrollment is represented as a bipartite network that is then flattened into two separate networks, a student network and a course network. Common network measures are made on these in-

dividual networks to gain insights on the structure of the university based on how students enroll in courses. This analysis validates our intuition about the importance large courses such as Introduction to Statistics (STATS-250), Principles of Economics (ECON-101), and Introduction to Psychology (PSYCH-111) play in connecting students from various disciplines across campus. Diving deeper this analysis revealed the importance of some courses with significantly lower enrollments. Aliens (ASTRO-106) has only one fifth the students that STATS-250 does, yet plays almost an equivalent role in connection students from across campus.

Questions related to how to characterize connections lie at the core of social network analysis. How are edges defined? Are they directed? Do they recieve different weight and if so how? In this thesis, we introduce three measures for defining a connection between students. The three types of connections are unique connections, weighted connections, and intensity connections. The first, unique, answers the question: who did you take courses with? The second, weighted, answers: how many courses did you take with an individual? The third measure, intensity, combines the previous question of how many, with the question of: what was the enrollment size of the courses you took with an individual?

We demonstrate how these various definitions can have significant impact on the measures we make on the network. We also show that the relationship between these measures varies depending on the subset of students you're looking at. For example, there is zero correlation between the unique connections and intensity connections a Mechanical Engineering BSE students makes, however, there is relatively high correlation with these two connections for History BA students.

We can employ network analysis to capture the student experience, and we can compare what we learn to more traditional categories and measures. For example, How effective, or informative, are the typical categorizations (or labels) used to describe students? The typical categorizations, which we refer to as legacy labels,

explored in this study are: split students into bachelors of science and bachelors of arts (BS/BA), the next splits students into humanities, social sciences, biological sciences, and natural sciences, and the final categorization splits students by majors. We introduce concepts such as label coherence, strong and weak recoverability, and robustness. We use these new concepts along with classic measures like Rand Index and Normalized Mutual Information to compare the legacy labels to an optimal clustering algorithm and random partitions. Through this analysis we find that BA and BS is not a good representation of courses taken. We also show that majors as a whole performs the best of the legacy labels, however, there is significant difference in performance between the majors.

Finally we explore the link between how connections are defined in a network and the recoverability of a labeling in a network. Here we see little correlation between strong recoverability and unique connections and high correlation between strong recoverability and intensity connections.

CHAPTER I

Introduction

1.1 Social networks in education research

The courses that a student takes during their college studies arise from a combination of institutional requirements and the choices the student makes to fulfill them, along with the student's other personal interests. Each choice creates an opportunity for connection to faculty members, other students, ideas, and experiences. The courses chosen are subsequently recorded on a transcript that describes the activities of this individual: major, courses and grades received, grade point average, and perhaps honors. However, the experience of the student can be better understood when placed in context and understood in relation to other students, which provides insights into interactions with peers and connections to the broader intellectual environment.

The interactions that occur among college students on the same campus have consequences for their experiences as learners in single courses [3] and reflect the eventual academic outcomes at the conclusion of their studies [3, 16]. To varying degrees these connections may be categorized as physical, behavioral, or associative [5]. Physical connections occur when individuals occupy the same space, behavioral connections include the active exchange of information, and associative connections emerge from the shared intellectual experiences and activities of individuals in the same course or major. Collectively these connections can be represened as a social

network [35]. Using network science and analysis to study this social network enhances our understanding of the learning environment of modern higher education.

A network where the nodes represent people or groups of people and the edges represent the interactions or connections among them is known as a social network. Because of the various ways in which people and groups interact or connect, the edges in a social network can take many forms. For example, edges can be formed from communication [18, 26] or relationships [12, 31].

A collaboration between and among groups of scientists is another example of a social network in higher education [10, 25]. In a network of that type the nodes represent individual scientists, and two scientists are connected (i.e., there is an edge between them) if they share co-authorship on a paper. [5] refined the network description to four levels of analysis. The first is the "ego" level, referring to an analysis of one node, the ego, and all of its connections, which are the "alters". The next two levels grow to collections of nodes, either dyads (pairs) or triads (triplets). The fourth level, which is the focus of this manuscript, is the analysis of the complete network, which encompasses sets of actors and ties among them in a bounded sample.

The study we report in this article responds to the call of [3], who argued that social network research in higher education "lacks a rich body of descriptive work portraying the student experience of college from a network perspective." We first provide useful definitions with a literature review. Next we describe the study. Discussion, practical applications, suggestions for future research, and conclusions follow.

1.1.1 Network analysis primer

Within the domain of network science a network is generally understood as a collection of two different types of objects called nodes and edges that depict a system. The nodes in the network represent subcomponents of the system while the edges, which connect the nodes and can also be called connections, represent interactions or

simply relations between and among these subcomponents. The resultant structure of the network can describe the connectedness of the system, the stability of the system, the information flow, and much more. This network structure then allows for the identification of components and interactions of interest. Two examples of large systems that have been represented as networks and that might be familiar to many are power grids [37] and railway networks [29].

There is a set of commonly measured network attributes and statistics that help in understanding network structure and composition. We now define three of these common measurements. Both degree centrality and local clustering coefficient are used in our study.

Degree Centrality. Reports the fraction of all the network nodes to which each node is connected. This measure simply counts the total number of nodes to which each node is connected within the network. This value is normalized by dividing the number of connections (or edges) a particular node has by the total number of possible connections the node could have (i.e., a network with n nodes is normalized by $n-1$). Degree centrality is often considered in relation to information flow through a network because it provides an estimate of the probability that this node will play a role in transmitting ideas (or information) within the network.

Local Clustering Coefficient. The clustering coefficient relies heavily on the idea of network transitivity. Transitivity is a term that reflects relationships: if node u is connected to node v and v is connected to w, then u is also connected to w. This leads us to an idea of partial transitivity, which is a concept in social networks where the fact that u knows v and v knows w does not guarantee that u knows w, but makes it more likely [24]. The clustering coefficient is an attempt to measure the level of partial transitivity of the network, whereas the local clustering coefficient is the clustering coefficient for a specific node. For each node, it probes how often this node provides a unique connection between two other nodes. The value of clustering

coefficient indicates whether a node is embedded in a tightly clustered region of the network. Conversely, it tells us something about how powerful a node is for connecting otherwise unrelated nodes. The lower the value of the clustering coefficient, the more powerful it is at providing unique connections [24].

Eigenvector Centrality This trait refines degree centrality's initial estimate of a node's importance in a network. It does so by assigning relative scores to all nodes, ascribing more importance to nodes which are highly connected to other nodes which are themselves important. Eigenvector centrality gives each node a score proportional to the sum of the scores of its neighbors [24]. This statistic is included here as an example of the increasingly complex refinements that are available to basic statistics like degree centrality, but it is not used in the results.

Community Detection "Loosely speaking, the goal of community detection is to find the natural divisions of a network into groups of nodes such that there are many edges within groups and few edges between them" [24].

1.1.2 Related Research

Due to the benefits of social network analysis, interest in applying it to education research has arisen. There are multiple settings in which networks have been applied in education research, including student relationships and faculty research collaborations within and among departments that are reviewed by [3]. They identified threads of research as "networks as dependent variables," "networks as independent variables," and "descriptive." As dependent variables, common themes have involved the roles of race and ethnicity in the formation of friendships [7, 19] by using surveys and later using Facebook data matched with institutional administrative data [21]. [20] also explored additional demographic traits as predictors of network behavior on Facebook. As independent variables, networks have been notably tested as predictors of GPA [2, 40], student integration and persistence [30], and health outcomes [11].

Social networks as they exist in learning communities have been considered at the classroom level [8, 15] and more extensively in massive open online courses (MOOCs) and E-learning [6, 14, 33]. Additional examples can be found in education research. For instance, Dawson (2008) [8] compared communication logs among students in an online forum to draw connections between them. In that study the student network measures of closeness, degree, and betweenness were used to assess each student's sense of community. Betweeness is defined as the extent to which a node lies on the paths between other nodes [24]. [8] and [4] asked how a student's centrality within a school friendship network affects student performance. [36] looked at friendship network size among faculty members and its relation to perception of work-family culture. [9] used networks to look at faculty hiring patterns at top universities.

The earliest example of a descriptive study of student networks is found in [13], who studied the friendships and communities formed among married veterans in housing at MIT. That work along with later descriptive studies [23, 28] initiated a field that has since been lacking in activity [3].

These studies offer a few examples of ways to characterize and analyze networks in higher education. They highlight how networks play a role in shaping higher education. Much of this research took advantage of only a few network measures to draw powerful conclusions: among them degree, closeness, and betweenness are most common. These studies focused on small networks, that is, no more than a few thousand connections and a small sample of individuals. The limited nature of the available data may therefore reflect a biased subset of the whole.

One of the central challenges of social network studies is how to measure connection (i.e., how edges are defined in social networks). Many forms of connection which we might like to study (e.g., friendships, collaborations, inspirations, conflicts, mutual support, mentoring) are not comprehensively recorded. As a result, social network studies need to rely on methods of estimating these based on context, survey data,

or inference.

With our work we hope to rekindle this field by using large student enrollment datasets, which all colleges have, as well as modern computational techniques that reshape data into a network setting for further evaluation. For this study we took advantage of one area of higher education for which data about substantive connection, that is extensive physical proximity, shared intellectual experience, and a suite of activities, is carefully recorded. Every college and university has maintained careful records of the courses taken by students. These records provide an opportunity to study networks of substantive campus connection in ways which can be replicated across the landscape of higher education.

1.2 Thesis Organization

This thesis is organized as follows: Chapter II introduces the Learning Analytics Data Architecture, or LARC [1], provided by the University of Michigan Information Technology Services. This dataset contains all of the enrollment data from the university dating back the 1990s. Chapter II shows how this data can be represented as a bipartite graph and demonstrates the application of network analysis to student enrollment. In this demonstration we use standard measures made on networks to explore how students and courses are connected across campus.

Chapter III takes a step back to explore the many ways a "connection" in a social network can be defined, specially in the case of enrollment data. It sets out to answer the question: If the number of students enrolled in a course matters, how does this effect measures made in Chapter II? Chapter III demonstrates the subtlety of these definitions in a toy model that is tractable and easy to interpret. It then applies it to a large subset of the LARC data. Chapter III brings to our attention how majors are not affected uniformly by these definitions. This implies that further investigation of majors is necessary.

Chapter IV explores the information content of legacy labels used in a university setting such as a BA/BS, or grouping students by Humanities, Social Sciences, Biological Sciences, and Natural Sciences, or by just using the majors students received as labels. It explores the information content as an abstraction of courses taken at the university. Chapter IV takes a subset of students in the LARC dataset that have certain majors. These majors were selected based on the number of students that were enrolled in them. Chapter IV compares how legacy labels perform to an optimal clustering and a random partitioning. Various comparison criteria are explored. Chapter IV also introduces the concept of recoverability of legacy categories. We define criteria for a legacy category to be strongly or weakly recoverable from data.

CHAPTER II

Student and Course Networks

2.1 Introduction

Residential higher education brings thousands of students together for multiple years and offers them an array of shared intellectual experiences and a network of social interactions. Many of these intellectual and social connections are formed during courses. Students are connected to students through courses they take together, and courses are connected to one another by students who take both. These courses and the students who take them form a bipartite network which encodes information about campus structures and student experiences. Because all institutions of higher education collect and maintain precise records of what courses students take, it is possible to assemble a student-course network that quantitatively describes the interactions among students and courses. We provide an example that demonstrates the identification of courses effective at creating unique connections among students and reveals how students and majors can be strongly connected or dispersed. We show how social network analysis is used to improve our understanding of the learning environment at the University of Michigan.

The purpose of this study was to demonstrate the application of network analysis to a large administrative dataset in order to gain insights into the connections formed among students and courses in higher education.

2.2 LARC Dataset

We used data from a large, selective, public, state university with over 40,000 undergraduate students and several colleges. In an explicit effort to make student information data more easily accessible to researchers, the University of Michigan Information Technology Services staff created the Learning Analytics Data Architecture, or LARC [1]. This dataset, which is curated and distributed by the Office of the Registrar, has enabled a wide range of learning analytics research efforts, including this project. It is available to researchers for a project if their IRB is approved by Michigan, and if there is a signed MOU. It provides an authoritative and complete view of the student data present in the data warehouse; and it is updated every semester, similar to the public data releases of "big science" projects such as the Sloan Digital Sky Survey [38] or the GAIA Satellite [27].

The LARC currently includes four main tables. They contain over 400 columns describing more than 200,000 students who enrolled in roughly 15,000 courses since the year 2000. The content of the tables is reviewed and updated every semester to accommodate researchers' needs and to allow for redefinition of fields as appropriate. All of the characteristics of students and courses used in this study were either drawn directly or derived from the LARC dataset. The data dictionary for LARC is publicly available online. For clarity and ease of understanding, we used more familiar names for data elements in this work rather than the official names used in LARC.

In addition to defining the connective structure of our bipartite network, data from LARC provided insight into labeling both student and course nodes with a variety of metadata. The labeling may include descriptive metadata such as the name and number of the course, offering department(s) and college, credit hours, time and location of meetings, structure (lecture, lab, discussion, seminar), enrollment, history of offerings, prerequisites, and categorization in terms of college requirements. For the metadata about students who take each course, LARC provides insights about students'

backgrounds, including information from campus admission, prior courses taken, and previous academic performance. Demographic information includes records of age, gender, ethnicity, country and state of origin, first or continuing generation status, and intended major at time of enrollment. Also, there is a complete record of subsequent courses taken and of honors and degrees ultimately earned by students.

For this study we used NetworkX [17] to model and analyze the connections among students and courses. NetworkX is a Python package constructed specifically for network analysis. We used it to extract from the networks a set of traits which characterize the role of nodes (students or courses) in the overall networks.

2.3 Methods

In this section, we begin by documenting the construction of our student-student and course-course networks, then describe the extraction of network parameters characterizing the nodes in each of the networks.

Building and partitioning the network: The full student-course network is bipartite, meaning it consists of two types of nodes (i.e., student nodes and course nodes) [24]. A student is connected to another student through courses that they both take, and courses are connected to courses through students who take both. Figure 2.1 is an example of enrollment in one semester and the connections can be interpreted as follows: student 1, $S1$, is connected to student 5 ($S5$) because they were both enrolled in course 2 ($C2$), in that semester. Courses are connected when one student enrolls in both courses during the same semester. Course 1 ($C1$) is connected to course 4 ($C4$) because student 2, $S2$ is enrolled in both courses. In practice, this network is quite complex. Each student may be connected to another student through a variety of courses they take together. The list of shared courses surely colors the nature of that "connection". Likewise, each course may be connected to another by a few or many different students, and the quality of this course-course connection is flavored

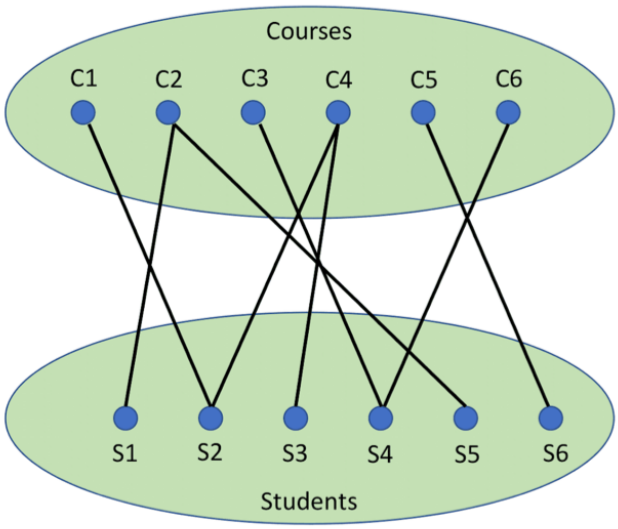


Figure 2.1: Example bipartite (two-mode) network. The top set of nodes are courses labeled $C1, C2, \ldots, C6$ and the bottom set nodes represent the students enrolled in those courses $S1, S2, \ldots, S6$.

by the composition of these students. The network is rich in information and can be used to address many questions about the student experience. The co-enrollment (bipartite) network is then flattened into two separate networks, a student-student network (Figure 3) and a course-course network. In flattening the network into two separate networks there can be some information loss; however, there are still useful insights to gain from the individual networks. Once constructed, we treated these student-student and course-course networks separately for the study

The construction of this network required consideration of boundaries. In any given term, students enrolled in courses may have academic careers which began long before a particular course enrollment and the term being examined. As a student's career progresses, they may be connected to students whose academic careers will continue after they graduate. Likewise, a course offered only once, before or after the term of interest, may be connected to a current course by a student who spends multiple years on campus. In this sense, there were no essential boundaries in time

for our student-student or course-course networks. Every student on campus was connected backward in time through a "friends-of-friends" network which extends back to the founding of the institution and forward in time to students not yet born.

In our construction of the network, we simplified the interpretation and eased the computational burden created by these boundary effects by restricting the study to a well-defined set of students who entered and approximately exited the university contemporaneously. We focused on the ego networks of a cohort of 6,738 students. The ego network refers to the analysis of the individual student and their connections. These students entered University of Michigan as undergraduate students for the first time in the fall of 2011 (including transfer students), and have graduated or were still enrolled by winter 2016. By the end of this five year period, 90.9% of those who entered in fall 2011 had completed a degree. This five-year graduation rate is typical for Michigan undergraduates. Most cohort students (90%) were in their first term of college attendance, and these students typically completed 6-10 terms of coursework during this five year period (the statistical mode is 8). The remainder of the cohort was almost entirely transfer students, who typically took between 2 and 6 terms of courses. The most frequent number of courses completed by the cohort of students was 33, with an average of about 4.1 courses per term.

The complete student-course network of these students included the full complement of their classmates from fall 2011 to fall 2016, a total of 68,946 students who enrolled in a total of 6,152 courses. The student network itself was an aggregation of 6,738 ego networks, one for each individual student within the cohort. This aggregation resulted in a network containing 68,946 nodes: 6,738 egos and their 62,208 alters or classmates. Thus, the thrust of the analysis of the student network concerned this cohort: we observed only some parts of the network for students (alters) who entered before fall 2011 or who graduated after winter 2016. This made information about network structure "fuzzy" at the boundaries. We did not know whether, for exam-

Course	CC	DC	N	Format
STATS-250	0.078	0.523	20231	LEC
ASTRO-106	0.104	0.422	4065	LEC
UC-280	0.099	0.400	12632	REC
ECON-101	0.102	0.411	14065	LEC
PSYCH-111	0.104	0.403	12516	DIS
DANCE-100	0.110	0.392	2761	LAB
ANTHRCUL-101	0.117	0.379	8190	LEC
ENGLISH-223	0.122	0.361	3184	REC
SPANISH-232	0.124	0.342	7206	REC
PSYCH-240	0.134	0.337	6621	LEC

Table 2.1: Top 10 courses by degree centrality

ple, two "older" students might be connected by a course taken prior to fall 2011, or two "younger" students who entered in fall 2016 might later become connected by a course. Only connections observed within the cohort of students who entered in fall 2011 and were still enrolled or graduated by winter 2016 are complete. For this reason, we studied and now report only the networks produced by cohort students in what follows.

2.4 Results

We now describe characteristics of the course network and the student network, exploring along the way how these measurements might be used by various stakeholders in the higher education system.

2.4.1 The Course Network

We begin by examining the courses with highest degree centrality, which is a measure counting the number of connections a node has. Network statistics for the top 10 courses by degree centrality are shown in Table 2.1. For each node we report course name and number, clustering coefficient (CC), degree centrality (DC) total number of students enrolled from fall 2011 to winter 2016 (N), and course format.

Several of these classes (e.g., STATS 250 – Introduction to Statistics and Data Analysis, ECON 101 – Principles of Economics, PSYCH 111 – Introduction to Psychology) have especially large enrollments, which naturally increases their degree centrality and gives them a prominent role in the course network. STATS 250 in particular is the largest course on campus, taken by well over half of all Michigan undergraduates at some point in their careers: 75% of the students in the course have sophomore or junior standing. STATS 250 continues to provide a basic foundation in statistical thinking to students from many disciplines: social scientists in Psychology, Sociology, and Economics; natural scientists in Biology, Chemistry, and Astronomy; and humanities majors in History, English, and Linguistics. The low local clustering coefficient (0.078) reflects the fact that it is especially likely to form the only connection between students whose academic experiences are otherwise remote.

ASTRO 106 – Aliens, provides an interesting and non-intuitive counterexample to the obvious large enrollment trend. While it enrolls only a fifth as many students as STATS 250, it has nearly the same level of degree centrality. ASTRO 106 is a one credit course on extraterrestrial life that is often taken by individuals in the College of Literature, Science, and Arts, either out of interest or as part of fulfilling a quantitative reasoning requirement. As such, it is often taken by students in the latter half of their studies (60% have junior or senior standing) and draws from a very wide variety of majors. ASTRO 106 has a higher local clustering coefficient, indicating that it forms fewer unique connections between students than does STATS 250. However, the quality of those connections is likely different because ASTRO 106 is not taught in a large lecture hall.

Smaller enrollment is also characteristic of DANCE 100 – Introduction to Dance, which is housed in a separate, and smaller college: the School of Music, Theater, and Dance. The enrollments in DANCE 100 are smaller than any of the other courses in Table 2.1. Designed specifically to introduce non-dance majors to the subject, it is

offered in many small "lab" sections, in which students drawn from all over campus are brought into extended, close contact. Such a course plays an outsize social and intellectual role on campus, a role which might be invisible to both students and campus leadership without this network analysis.

UC 280 – Undergraduate Research, provides another dramatically different example. This course, which meets once a week in groups of 40, delivers support for students engaging in undergraduate research in faculty labs during their first and second years on campus. As such, it engages a wide variety of students in the shared experience of working closely with a faculty-led research group. The high degree centrality of UC 280 shows that it connects a large number of students, and the very low clustering coefficient shows that the connections it forms among students are often unique.

The role of these courses as especially powerful connectors should be more widely known, both to students and campus leaders because they are adept at leveraging existing diversity on campus by exposing students to peers with whom they might not otherwise interact. The variety of enrollments and formats also serve as important reminders that the course structure and format of STATS 250 includes several hours a week in large lectures, while ASTRO 106 is offered in smaller, more intimate discussion sections, which classify the connections differently. While they share the same space and engage with the same content, STATS 250 students are likely to engage with only a small fraction of their classmates, while those in ASTRO 106 likely enjoy a setting that may allow for more meaningful interaction.

Other large enrollment introductory courses like ECON 101 – Principles of Economics, PSYCH 111 – Intro to Psychology, and ANTHRCUL 101 – Introduction to Anthropology connect to many other classes; but they provide unique connections among these courses less often than does STATS 250. They are less likely to connect to courses across more substantial disciplinary divides on campus. This distinction

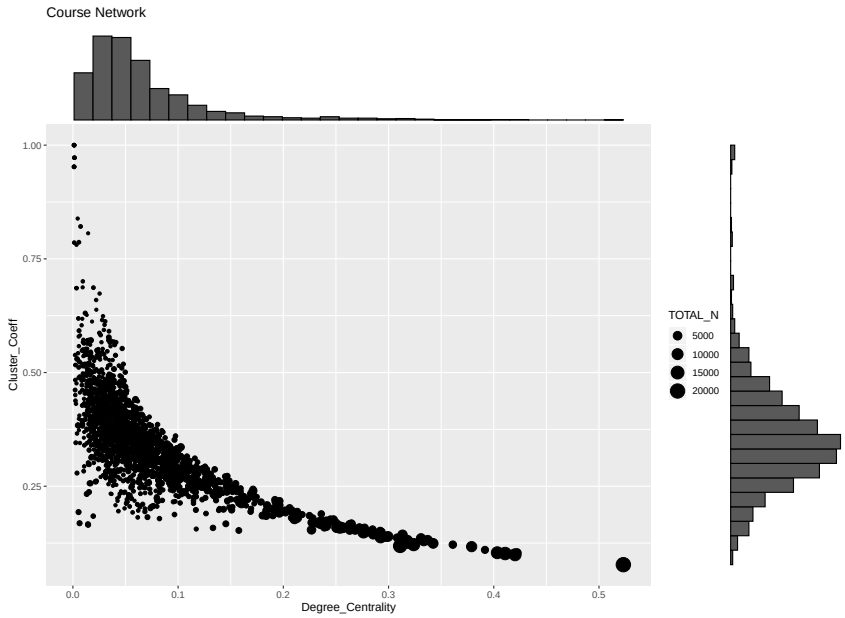


Figure 2.2: The Degree Centrality vs. Clustering Coefficient for 1,878 University of Michigan courses with enrollments N > 100.

becomes stronger for more advanced courses on the list, like PSYCH 240 – Introduction to Cognitive Psychology. This course is taken primarily by students who will major in psychology or one of the several forms of neuroscience, so it is less likely to connect to more distant subject areas. A course does not necessarily facilitate unique connections among other courses by virtue of its size.

Figure 2.2 examines the structure of the course network more generally, we see that the relationships among total enrollment, degree centrality, and the clustering coefficient are strong, with the largest, most highly connected courses also more likely to provide unique connections. However, there is complexity here too. In Figure 2.2 each point represents a course and total course enrollment over the time period considered; the 10 courses in Table 2.1 occupy the high degree centrality low cluster coefficient. Point sizes indicate the relative course size. Marginal histograms show the one-dimensional distributions of degree centrality (top) and clustering coefficient (right).

Some large enrollment courses, for example PHYSICS 240 – General Physics II

remain relatively isolated and do not connect to many other courses; and they rarely connect otherwise unconnected pairs of students. Conversely, some relatively small enrollment courses have strikingly high degree centrality and low clustering coefficients, which shows that they are especially effective at providing connections among otherwise remote courses. A handful of courses live in very tightly clustered environments so that almost every pair of courses they are connected to is also connected to one another. Examples of such courses include studio music courses, along with advanced undergraduate courses in Pharmacy, Classics, and Naval Architecture.

These examples help to illustrate some of the ways in which course network information might inform various audiences on campus. Students could use this information to seek out courses which will connect them to a more diverse array of other students. Faculty members might use these networks of connection to better understand where their students are coming from and where they might go. Administrators might use these networks to better understand the student experience and perhaps to draw together the community of instructional teams working on the most connected courses, supporting them more openly in their efforts to create especially inclusive and equitable experiences for the diverse students whom they serve.

2.4.2 The Student Network

In Figure 2.3 a sample of the student network is shown for a selection of highly populated majors, both egos and alters. Shaded circles represent individual students (nodes), connected by edges whose length is inversely proportional to the strength of the connection between students. Students co-enrolled in many courses (as those in same major often are) cluster tightly within a major, and the majors themselves cluster according to the co-enrollment of students in the two majors. As expected, students cluster by major through the courses in which they co-enroll; and some majors are more tightly clustered in the network than others, reflecting both the rel-

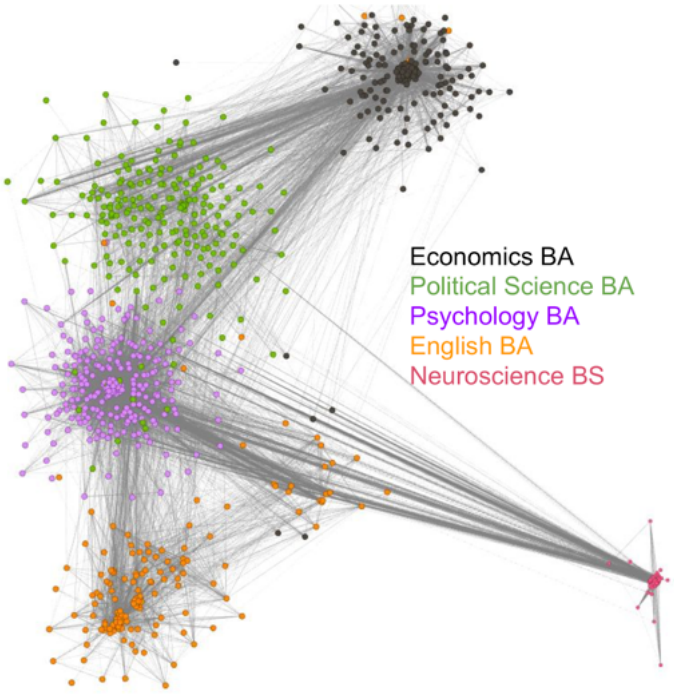


Figure 2.3: Flattened student-student network for large majors

ative flexibility of curricular requirements and the choices made by each individual in course selection. Neuroscience majors, which include many pre-medical students, are tightly clustered, due in part to a large set of prerequisite biology and chemistry courses, while Political Science majors show reduced clustering that reflects the relatively greater freedom they have in fulfilling their degree requirements. The clustering also does not obey hard boundaries. The English majors found among the Economics cluster (and vice versa) are individuals who used freedom in the curriculum to enroll in several courses commonly taken by students in the other major, and perhaps even to double major.

In Figure 2.4 we display the relationship between the full student network degree centrality and clustering coefficient for all of the students in the fall 2011 cohort. Each point represents a student in the cohort. Marginal histograms show the relative distributions of cluster coefficient (right) and degree centrality (top) among students. The correlation between the two is similar to that seen in the course network, though

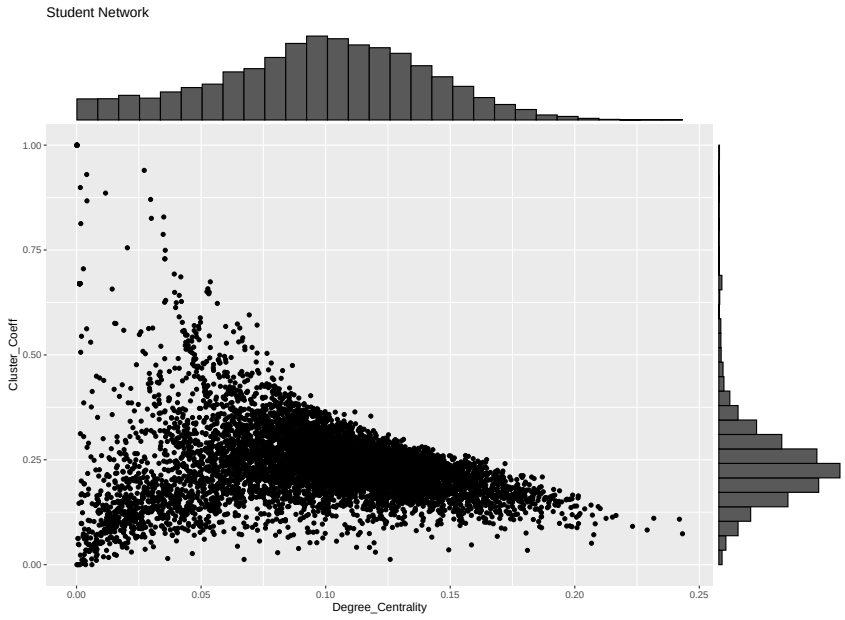


Figure 2.4: Degree Centrality vs. Clustering Coefficient for University of Michigan students

less pronounced, largely because the highest degree centrality seen in the student network is only about half that of the course network. Some basic features of the student network merit further mention.

The students with the highest degree centrality in the fall 2011 cohort are connected through course co-enrollment to almost 25% of all the 6,738 Michigan undergraduates who entered in that term; they each took courses with about 1500 unique individuals. Those with the top five degree centralities all had different majors, though all appeared to be pursuing a track toward medical school. Four had completed their undergraduate degrees: a Biopsychology, Cognitive Science and Neuroscience B.A.; an Asian Studies B.S.; a Biomolecular Science B.S.; and a Biology, Health, and Society B.S. One of the five students had not yet graduated in winter 2016. All of these students were enrolled for the full five years, completing ten terms and taking anywhere from 41-47 courses. Degree centrality can only rise as one adds additional classes, and it does so especially rapidly with large STEM courses, like these which premedical students regularly take.

There are also students with very low degree centrality; they are connected to almost no other students in the fall 2011 cohort. Many are transfer students, who entered in fall 2011 as juniors or seniors; took courses primarily with older, non-cohort students; and graduated without forming extensive connections in this cohort. The most extreme cases, occupying the upper left of Figure 4, are students who returned to school to receive, for example, a second-career B.S. in Nursing. By embedding enrollment in the network setting and without explicitly labeling transfer students, the resultant network statistics easily identify them as outliers. These measures reflect an important reality of their university experience, that is, they interacted with a much smaller and less diverse group of peers.

At each level of degree centrality, we find students whose clustering coefficients cover a broad range. Some are embedded in relatively dense parts of the network, like large majors in the College of Engineering. Such students rarely create unique connections between pairs of other students: they are almost always already connected. Others are vital connectors, regularly providing the only connections between students in deeply connected but otherwise separate neighborhoods. These students are unusual, often majoring in two or more fields, sometimes studying in two separate colleges

2.5 Discussion

2.5.1 Practical Applications of Network Analysis

A complete student-course bipartite network can provide answers to questions raised at many levels within higher education. For example, presidents, provosts, deans, department chairs, and faculty members may seek to better understand how to design their curricula or to allocate resources to courses that might provide opportunities for a greater number of students to experience the benefits of diversity.

Which courses are especially important for creating interdisciplinary, cross-campus connections? Which courses provide especially rich opportunities for connections among students with differing backgrounds, interests, and goals? Where do first year and senior students or traditional and non-traditional, students interact? Campus leaders may also use these same analyses to gain a deeper understanding of the student experience, identifying groups of students who are especially well-connected or especially isolated, exploring the relationships between curricular requirements and student connections, and designing new courses which enhance desirable connections where they are lacking.

There are also practical applications for students, who may now probe the breadth and depth of their academic experience with greater clarity. They will have the opportunity to see what kinds of courses are likely to build their network of connections and to examine and evaluate the extent to which they have contributed to the network of connections across the campus. They can query how similar or diverse they are to other students in the network and then make informed decisions. Students can see how they are connecting otherwise disparate parts of campus (e.g., they could be connecting two departments that do not normally have students interacting). They can also see the reverse and see how isolated they are relative to the possibility of connections on campus. While answers to some of these may seem intuitive, network analysis helps us quantify these ideas.

Once constructed, student-student and course-course networks can be analyzed to identify community structures using any of the variety of community finding algorithms developed by the network science community over the last few decades [24]. Communities finding algorithms partition networks into groups of nodes having high within group connectivity and low between connectivity. Community finding provides insights into the hierarchical structure of networks. If used on communities of students or courses, community finding algorithms will likely identify obvious com-

munities, that is, majors and colleges, while also quantifying similarity and difference in new ways, showing, for example, that the courses taken by economics students more closely resemble that of natural scientists than social scientists. These types of algorithms also probe which elements of the curriculum are important primarily for a discipline and which elements are cross-disciplinary. Community finding will identify the communities that emerge due to the network structure of connections and will allow for a direct comparison to department requirements.

As these results are presented to groups of students, faculty, administrators, and our peers in higher education, future work will be undertaken to understand how they are actually used in practice. Moreover, measures such as these will be of broad interest to members of the research community who may use them to address larger research questions that use network statistics as either independent variables predictive of certain outcomes or in the mode of dependent variables, where the network statistics are the outcomes themselves [3].

2.5.2 Refinements and Extentions

As others have pointed out, networks may not measure precisely what was intended. The process of validation for a measure or instrument is an integral part of social science; for example, [14] and [22] recently treated the matter in detail in the context of social ties formed in MOOC forums. In the framework of [5], co-enrollment in a course certainly creates a physical and associative connection, but what behavioral connections are formed among students is not known. Is it necessary that both students take the course at the same time, or can a meaningful "shared" experience emerge when they both complete the same course in different terms? Should the strength of connection be weighted by credit-hours, time in class, course structure, or measures of difficulty? Is the connection created by a small seminar stronger than that created by a large lecture course? Are two courses equally "connected" when a

student takes them both in the same term or takes one in the first year and one in the fourth year? Sensible answers to these questions depend upon the specific outcomes or phenomena of interest. This work does not directly address all of these questions, but we suggest that it does lay the groundwork for future studies.

Important refinements and extensions of this work remain to be explored. Each choice made in the construction of our bipartite student-course network is open for reconsideration. For example, an immediate task is the exploration of weighting of the connections produced by co-enrollment. This may be especially important if we want our student-student networks to model social connections reasonably well.

Another significant extension will allow for course connections to form when the same students take a course in different terms. This will change the course network in substantial ways, emphasizing course sequences associated with major requirements which are currently absent. In an analogous way we might create a student network meant to reflect only shared intellectual experiences and to connect students who take the same course in different terms. Such a network would describe shared intellectual experiences, social ones less so.

Other opportunities exist to gather richer, more precise measures of campus connections. Perhaps the most important involve the input from the students themselves. A number of learning analytics studies [41] have been done relying on student self-reports of networks of connection, ranging from networks on social media like Facebook to self-reports of who studies with whom.

2.6 Conclusion

In this chapter we have demonstrated the power of the network framework in the context of large institutional datasets. Networks are a direct means to quantify elements of the course-course and student-student interactions for all students and all courses. The features of the network presented thus far only scratch the surface of the

rich description this framework affords. Recent literature [3] has called for reinvigorating social network analysis in higher education as a means of description of higher education. We believe that our analysis responds to this challenge by assembling a bipartite student-course network, which consists of over 68,000 students connected by 6,152 courses. The flattened student and course networks are then turned to describe the relationships of courses to one another through the students that take them and relationships of students to one another through the courses in which they co-enroll. We reached the following conclusions.

- The intuited belief that high enrollment courses uniquely connect students from different academic backgrounds is confirmed for some courses, but is not the rule.
- Specific low enrollment courses can also as serve as equally powerful unique connectors of students.
- Students cluster by major, as expected; and unusual students, such as transfer students, appear with low degree centrality due to the smaller number of courses they have taken on campus.
- Some majors, particularly pre-medical, exhibit higher degree centrality due in part to enrollment in more large courses.
- High degree centrality in students is also not a guarantee that they form unique connections (low local clustering coefficients) among different communities.

These results both confirm that network measures accurately recover intuitive connections and uncover those that are less than apparent. At this point it important to recall that, despite the ability of a course to facilitate unique connections among students, simply occupying a space in a large lecture at the same time is no guarantee of a meaningful exchange. Proper interpretation of these results requires

an acknowledgement that course formats can be more or less conducive to this kind of engagement, and lack of course format should be borne in mind by those using these tools to draw conclusions about particular courses and their students.

Sometimes a lack of connection may be especially troubling, as it might be if students working on algorithmic data science have little chance to encounter ethicists. Other examples of troubling forms of isolation may take place along lines of social class, ethnicity, gender, or nationality. We suggest that institutions could have the opportunity to minimize the systems that perpetuate inequality in higher education by systematically examining networks of campus connections.

While limited in scope, this work provides a first look at the ways in which network measures of student connection through course co-enrollment can provide new insights into how students connect with students and courses connect with courses on the campus of a large public research university. Because the data used to build these networks is available at essentially every university and college, this kind of analysis could be replicated. Doing so would help everyone concerned with higher education better understand how campus connections form and where they do not form.

CHAPTER III

Connections in the Student-Course Network

3.1 Introduction

In this chapter we seek to answer the following question: In social network analysis of student enrollment data, how do you define a connection? We explore three different ways of thinking about student connections. It is important to note that we are only looking at the enrollment data and that we are aware of the numerous other ways students connect on campus from dorms, sporting events, social/academic clubs, and other various extracurricular activities. That being said, restricting our analysis to connections only within the classroom poses some interesting questions about what defines a connection in the class room.

One way that we evaluate a student's connections is count the number of unique students which they take a course with. In our analysis, we will refer to this as the network of *unique connections*. When we think about connection between nodes, we need to think about the edge weights. The unique connections gives every edge a binary weight of zero or one and does not account for how many courses a student takes with another. A reasonable next step would be to say that the edge weight between two nodes (students) is equal to the number of courses they have in common. In our analysis we refer to this as a network of *weighted connections*. In this mode if two students took five courses together their edge weight would equal 5. This method

of defining a connection considers each course to have the same value of weight. However, one can imagine that the connection of two students depends, for instance, on the size of the enrollment in the course they take together. In our final method of defining connections we use this interpretation, with smaller courses getting a higher weight value than larger courses. We call this a network of *intensity connections*. Intensity works similar to weighted connections, but instead of a +1 for every course a pair of students take, a student gets a value that decreases as a function of the course size. All of these methods are described in more detail in section 3.3.

3.2 Data

In this chapter we, use the same LARC dataset introduced in Chapter II. Here we focus on the same cohort of students that enrolled to the University of Michigan in Fall 2011. There are 6738 students in this cohort, they took 17960 unique courses, and had 174 unique majors.

Figure 3.1 shows the distribution of course sizes. This is of particular interest to our calculations because they highlight the potential for unique classmates, frequency of interaction, and size of courses for the three ways in which we define edge weights. A few observation from these distributions: first, just over 75% of the courses taken by this cohort have less than 11 students enrolled. A majority of the courses that have one student enrolled seem to be independent study courses. These independent study courses make up a majority of this distribution. On the opposite end of the spectrum we also see that there are 27 courses with enrollments greater than 500 students. English 125 (College Writing), Math 115 (Calculus 1), and Chem 130 (General Chemistry) all have more than 1000 students enrolled. The remaining courses with greater than 500 students are other large freshmen courses.

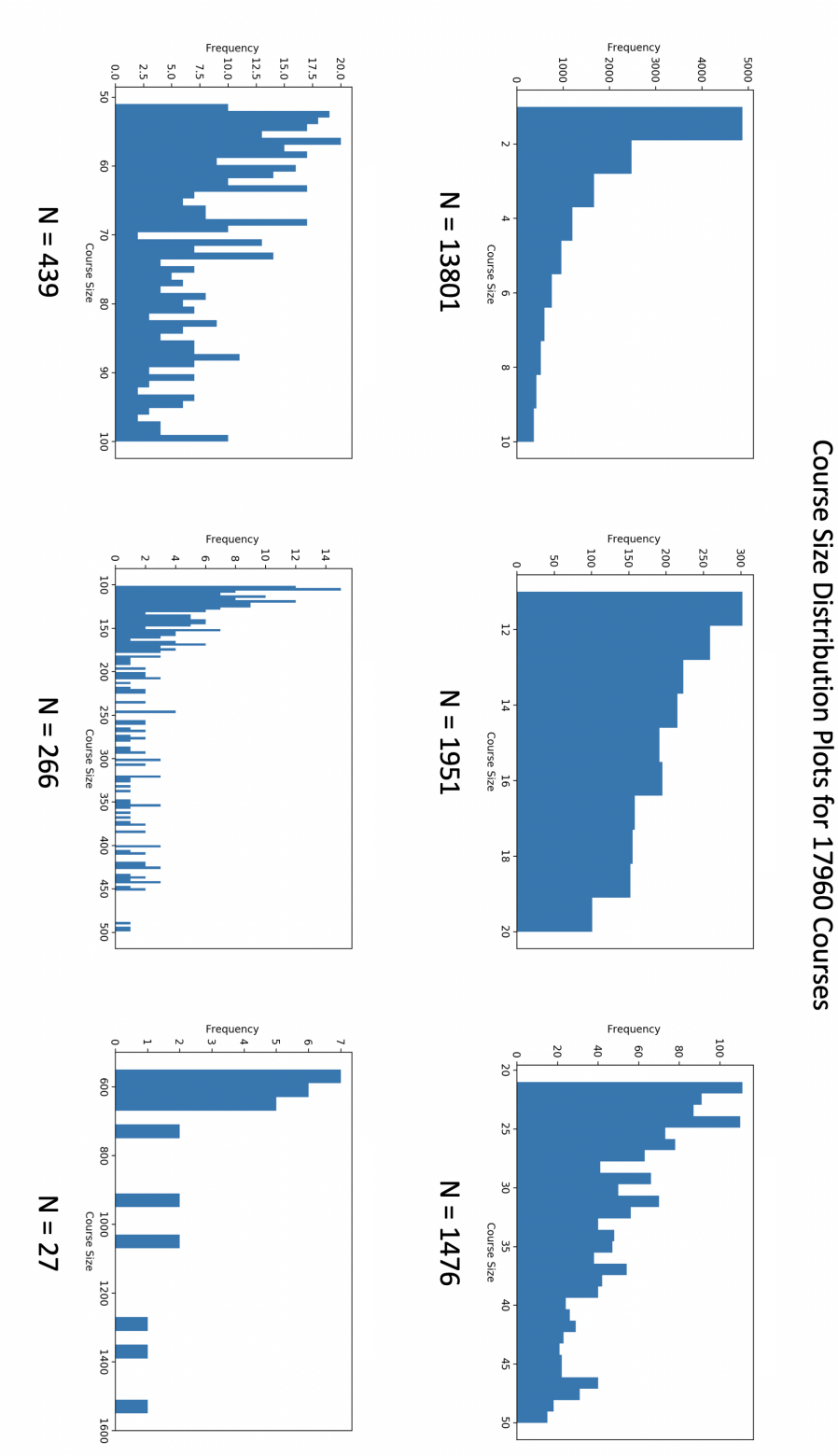


Figure 3.1: Course Enrollment Distributions

3.3 Methods

We represent the enrollment data with a bipartite network. A bipartite network is a special type of network whose nodes divide into two separate courses. In our analysis the two sets of nodes represent the courses and students in the data-set. There exists an edge between a student node and a course node if the student takes that course in a specific semester. An example of a bipartite student network is shown in Figure 3.2.

Figure 3.2 is a toy-model used to help demonstrate the methods and results that are calculated on the real student enrollment data. The toy-model network is composed of 10 students and 10 courses and edges. This representation allows questions about what courses a student took to be answered simply by looking at the edges. For example, in Figure 3.2 student {S1} is enrolled in courses {C1, C6, C9, C10}. Course-centric questions about courses can also be visualized, for instance, {C2} was taken by students {S3, S6}.

In this methods section, we introduce three ways of building the student network. These three methods differ in how they define a connection between two students in the network. Section 3.3.2.1 describes the process of defining unique connections between students. Section 3.3.2.2 describes how weighted connections are defined and finally section 3.3.2.3 describes how intensity between students is defined.

In the Results (Sec. 4.4) we explore how these three different constructions vary in how they are distributed among students and majors. We then explore how these three different ways of drawing connections in a network effect various centrality measures. The centrality measures we consider are degree centrality, eigenvector centrality, and triangle centrality. Degree centrality is a normalized count of the number of edges a node has. Eigenvector centrality gives each node a score proportional to the sum of the scores of its neighbors [24]. Triangle centrality finds the number of triangles that include a node as one vertex.

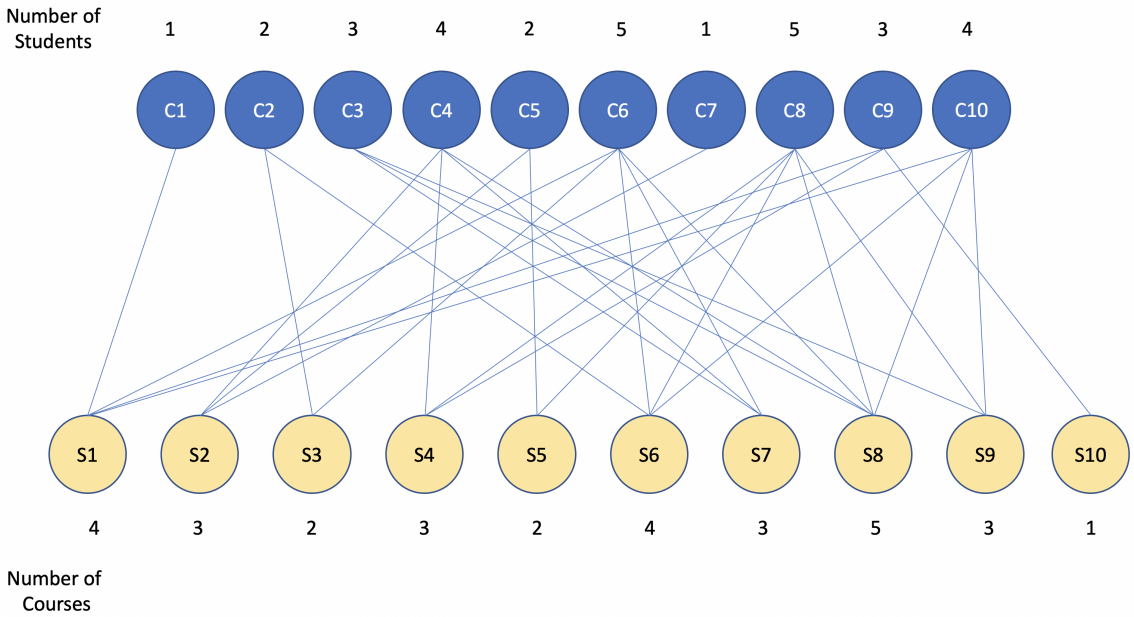


Figure 3.2: A toy model of the bipartite student-course network. Students (upper row) are connected to one another through the courses (lower row) in which they co-enroll. When considering the flattened student-student social network, we imagine scenarios where the edge weights can be binary, up-weighted by the number of courses taken together, or down-weighted by the course enrollment.

3.3.1 Flattening the network

Once the bipartite network is constructed, single mode networks can be extracted. This process is referred to here as "flattening", which in effect produces two single-mode networks. In this case we have a network comprised solely of student nodes which we refer to as the "student network" and a network consisting of only courses which we refer to as the "course network".

To build the student-network from the bipartite enrollment network, edges are formed between students when they are enrolled in the same courses. In Figure 3.2 student S10 only took one course, C9. However, they will be connected to two other students, S1 and S4, who also took course C9. In a similar fashion, to build the course-network, courses will be connected to each other if a student enrolled in both courses. Again we can refer to Figure 3.2 and look at course C2. C2 is connected to

C6 via student S3 and it is also connected to C6, C8, and C10 via student S6.

Our focus is ultimately the flattened student network, where now the influence of particular courses is confused, and so is the nature of the of the connections student form because of the information loss. For instance, student S10 took a single course C9, that has an enrollment of three students. The higher enrollment of C9 (than, say, C2) could mean lower quality connections formed among the students than a lower enrollment course. Large lectures are a limiting case of this scenario.

The connection between students also needs clarification in the one mode projection. Student S1 took courses C1, C6, C9, and C10 and is now connected in some way to peers in all those courses (except C1). The course sizes notwithstanding, S1 took more courses with S9 (two total) than with S8 (one total), so the mere presence or absence of and edge does not suffice to capture a key feature of the two-mode network.

In both cases, weighting schemes that acknowledge these practical realities are worth exploring. Consequently, the student-centric statistics we derive from the one-node network are sensitive to these schemes.

3.3.2 Weight Considerations

Weighting schemes were considered in some detail by [39], and [32] pointed out their effect on network statistics. Weight signals how "strong" the edges between two nodes are. In studying social networks the weight can signify the strength of the connection between two nodes. For example, two students that took multiple courses together should have a higher weight than two students that only took one course together.

Different forms of connection that might be represented in a social network in an educational setting [5]. Some of these may include more intentional and personal interactions while others indicate shared experiences.

In this study we create three different student-networks: unique connections, weighted connections, and intensity connections. Each weighting scheme places different emphasis on the character of connections formed between students.

3.3.2.1 Unique Connection Edge Network

In the unique student-network two students get a weight of 1 if they ever enroll in the same course. This value does not increase with the number of courses they have in common like weighted. It is defined by the following equation

$$u_{i,j} = \begin{cases} 1, & \text{if student } i \text{ and student } j \text{ ver take a course together,} \\ 0, & \text{otherwise} \end{cases}$$

3.3.2.2 Weighted Connection Edge Network

The weighted student-network is similar to the intensity student graph except each course is not weighted by the number of students in each course (i.e., each course has an weighted connection of 1). The weight given between two students in the weighted network is given by

$$w_{i,j} = \sum_{k,k'} \delta_{k,k'}$$

where δ_{ij} is known as the kronecker delta function and is defined as the following

$$\delta_{k,k'} = \begin{cases} 1, & \text{if } k = k', \\ 0, & \text{if } k \neq k'. \end{cases}$$

In the weighted student-network two students get +1 for every course they take together. This situation is useful for when the focus on physical/social interaction is lowered, but still exist. One particular example stands out, when you are trying to quantify similarity of experience between two students. Two students that had 7

courses together (co-enrolled) are more similar than two students who only have 2 co-enrolled courses. Similarity is a relative concept, which why the network setting is useful it allows for computation between nodes and groups of nodes.

3.3.2.3 Intensity Connection Edge Network

The intensity connected edge network is built on the premise that the edge between two students should depend on the size of the course. This is important if we are thinking about students physically/socially interacting. In this case students in smaller courses have a higher chance of interacting and thus have a higher weight than two students in a highly enrolled course. This is demonstrated in in Figure 3.3 below. Figure 3.3 is an example weight function that is only used in the toy model. Using the toy model as an example we know the distribution of course sizes. The largest course size is 5 and the smallest course size is 1. Figure 3.2 shows there are two students students enrolled in course C2. Thus each receives a weight score of 0.75 for that course.

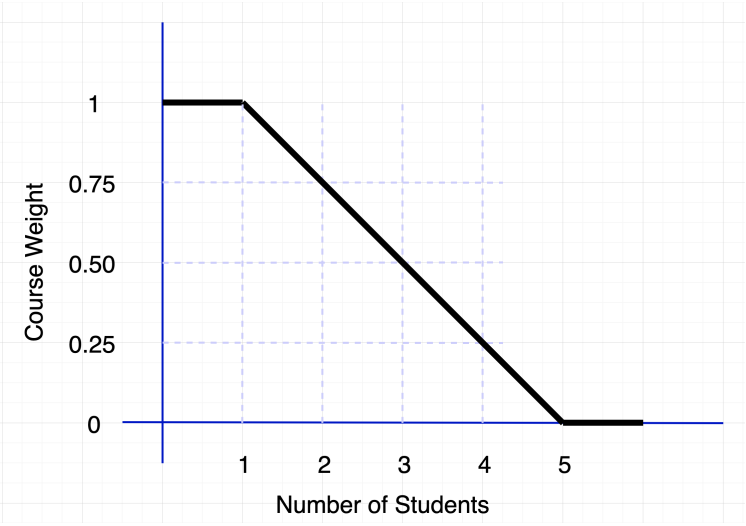


Figure 3.3: Course weight function for the toy model.

The weight of a given course is defined by the function:

$$f(\theta_C) = \begin{cases} 1 & \theta_C \leq 1 \\ \frac{-1}{4}(\theta_C + 5) & 1 < \theta_C < 5 \\ 0 & 5 \leq \theta_C \end{cases}$$

Where θ_C is the number of students enrolled in a course (e.g., using C2 from above we see that $\theta_{C2} = 2$). We can use $f(\theta_C)$ to calculate the weight of a connection between students. The edge weight between two students is the sum of the weight courses that the took together. It is defined by the following equation

$$I_{i,j} = \sum_{Ck \in X} f(\theta_{Ck})$$

where $X = C_i \cap C_j$ and C_i and C_j is defined as the set of courses for student Si and Sj respectively.

3.3.2.4 The interactions of weights

Weights are important when evaluating the connection between two students. Changing the weighting scheme can significantly effect results of network statistics. For example, looking at Table 3.1 we can see how when calculating the eigenvector centrality using the intensity student-network resulted in S6 being ranked 4th whereas using the weighted network S6 is ranked 9th and 8th for the unique network. Using Figure 3.2 we can easily investigate why this is true. S6 has a low intensity score because two of the four courses (C6 and C8) they are enrolled in have 5 students, which in this toy example is a large amount for enrollment. When we apply the intensity function to these two course it returns a value of 0, thus S6 gets no "points" for connections made in these courses.

Intensity EC	Weighted EC	Unique EC
S3 – 0.069	S10 – 0.048	S10 – 0.116
S10 – 0.153	S2 – 0.124	S2 – 0.224
S5 – 0.171	S5 – 0.151	S3 – 0.234
S6 – 0.191	S3 – 0.180	S5 – 0.271
S4 – 0.290	S4 – 0.273	S9 – 0.337
S9 – 0.316	S7 – 0.321	S1 – 0.348
S7 – 0.331	S1 – 0.331	S7 – 0.363
S1 – 0.426	S9 – 0.345	S6 – 0.369
S8 – 0.453	S6 – 0.434	S4 – 0.383
S2 – 0.476	S8 – 0.576	S8 – 0.399

Table 3.1: Eigenvector centrality of students in the toy model (Fig 3.2).

	Weighted	Unique	Intensity
Weight	1	0.804	0.026
Unique	0.804	1	0.146
Intensity	0.026	0.146	1

Table 3.2: Weighted, Unique, and Intensity Correlation Matrix

3.4 Results

We now examine how the three different ways of calculating connections lead to dissimilar inferences from the data. For the real data we use a sigmoid weight function to calculate the intensity score for each course.

$$f(\theta_C) = \frac{1}{1 + \exp\{slope * (coursesize - cutoff)\}}$$

3.4.1 Student Distributions by Network Type

To examine how the students are distributed among our three measures of connection (unique, weighted, and intensity) we sum each row of the student x student matrix. Summing each row gives a total value for a measure for a specific student. We then check to see how these totals are distributed.

Figure 3.4 shows the results for all three. In Figure 3.4 the unique connection distribution shows that majority of the students have more than 2000 unique connec-

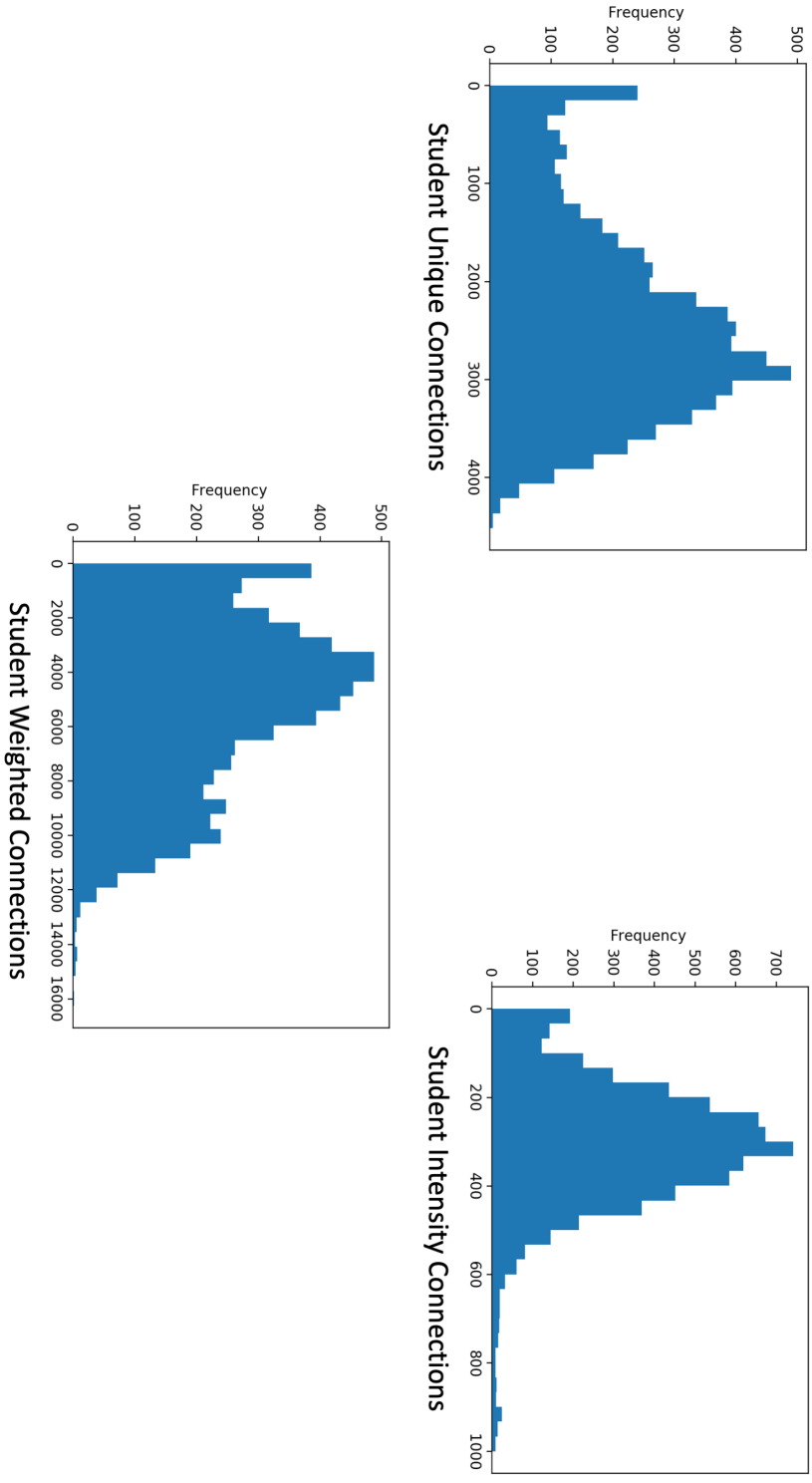


Figure 3.4: Distribution of student unique, weighted, and intensity values

	Weighted	Unique	Intensity
Weight	1	0.915	0.673
Unique	0.915	1	0.585
Intensity	0.673	0.585	1

Table 3.3: History - Weighted, Unique, and Intensity Correlation Matrix

	Weighted	Unique	Intensity
Weight	1	0.790	-0.202
Unique	0.790	1	0.069
Intensity	-0.202	0.069	1

Table 3.4: Mechanical Engineering - Weighted, Unique, and Intensity Correlation Matrix

tions.

Table 3.2 shows the correlation between the measures, we see that unique and weighted have a correlation of 0.80. Intensity has near zero correlation with weighted and a low correlation of .15 with unique.

To investigate this phenomenon further we explore how these different definitions of connection alter the distribution for specific majors.

Figure 3.5 and Figure 3.6 show how majors vary significantly in the three results. The differences in the three distributions tell different narratives about how students take courses within each of the majors. In Figure 3.5 we see that the distribution of unique connections for History students is spread between 50 and 80. The distribution of weighted connections, this value is between 60 and 160. This says that History students tend to take one to two courses with the students they meet. The distribution for intensity connections tells us that one of those two courses is a large lecture because the intensity score is close the unique score.

Figure 3.6 tells a different story. Almost all Mechanical Engineering students make 225 unique connections. The distribution of weighted connections says that they take 10 courses with the unique connections. Form the intensity connections we see that a majority of these 10 courses are large courses.

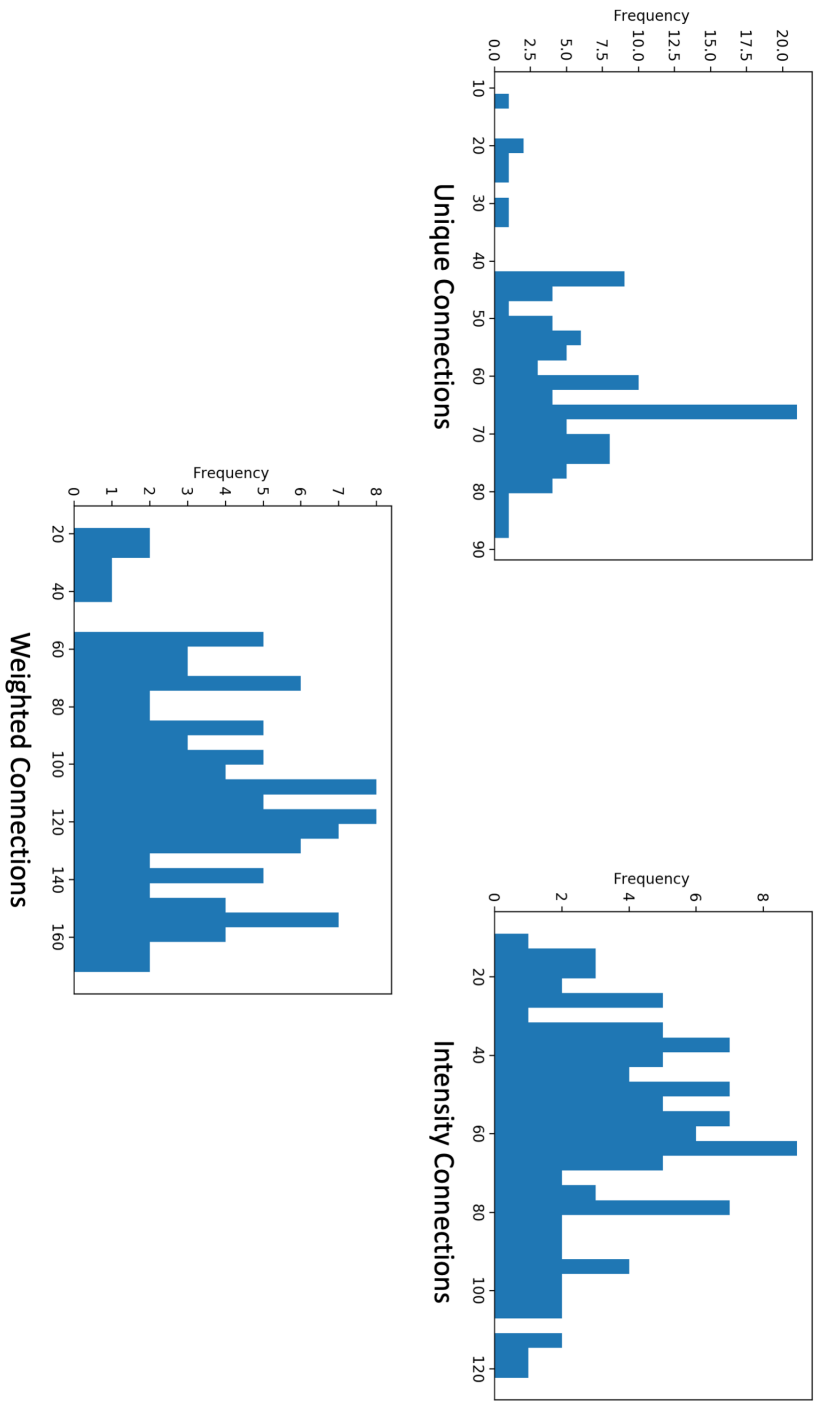


Figure 3.5: History Major Distribution - Unique, Weighted, Intensity

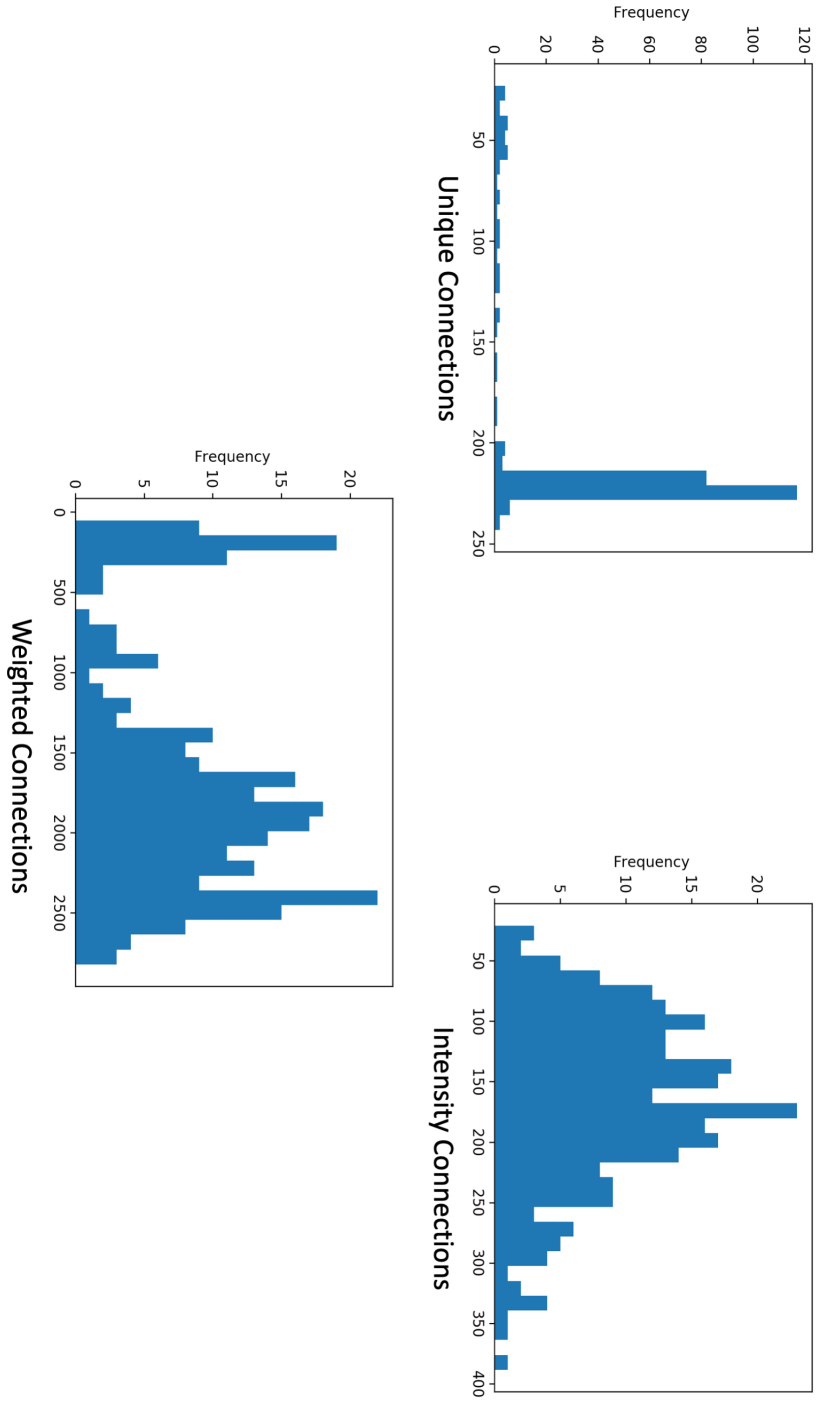


Figure 3.6: Mechanical Engineering Major Distribution - Unique, Weighted, Intensity

Recall how unique connections and intensity connections have almost zero correlation overall. This is true in aggregate but not at the level of individual majors. Table 3.3 shows History BA major has a positive correlation of 0.67 between the unique and intensity and Table 3.4 shows that the Mechanical Engineering BSE has a negative correlation of -0.20. We also notice a max intensity in History of 120 while Mechanical Engineering has a max intensity of 375.

Another noticeable difference in two figures is in the number of unique connections the student makes which is shown in the top right of both figures. Mechanical Engineering students have a very tight distribution around 225 while the History students are not only lower in magnitude, but also display higher variance. The lower left corner of the figure displays the student weighted distributions. Majority of the Mechanical Engineering students have greater than 1000 weighted connection values.

This indicates that not only do the these students make a lot of unique connections, they also take multiple courses with these students. We notice about an order of magnitude difference in the unique and weighted graphs.

This is not the story with History majors. A majority of the history majors have between 60 and 160 for their weighted connection values. When we compare this to their number of unique connections, we notice they are about the same order of magnitude. This implies that not only do History majors have a low number of unique connections, they only encounter most individuals in at most 2 courses on average.

The distributions for all of the other majors can be found in [A.1]. The mean of unique, weighted, and intensity formulations for each major are displayed in figures 3.7, 3.8, and 3.9. The actual values are in Table 3.5. This summary highlights a the following: First, looking at the range of values. For unique connections, the lowest value of 60.58 for History. While Psychology has about 3.5 times larger value of 222.62. Yet, they have almost equal Intensity scores.

Major	Unique Mean	Unique Var	Weighted Mean	Weighted Var	Intensity Mean	Intensity Var
Mechanical Engineering	197.83	3080.42	1638.28	604309	165.57	4985.92
Economics	174.93	1702.34	504.41	43951.55	76.01	693.82
Political Science	147.65	858.53	274.59	8415.97	78.39	735.58
Computer Science	183.77	1530.41	1089.04	180865	94.15	1138.03
Biopsych, Cognit & Neurosci	192.13	354.58	729.99	65973.17	46.65	346.62
Neuroscience	196.84	111.45	928.94	51264.92	54.88	658.91
Industrial & Oper Eng	152.44	1395.79	1247.89	253527.94	231.39	6794.50
Communication	151.61	280.86	402.46	9964.09	106.50	1082.72
International Studies	125.52	394.79	253.66	5081.48	52.35	420.22
Mathematics	81.69	373.13	180.07	4572.71	59.32	468.31
English	92.77	368.71	178.14	3124.33	80.98	681.07
Chemical Engineering	111.46	280.97	1114.98	149271.02	97.15	3015.39
Movement Science	109.00	46.30	632.96	27546.31	184.94	2023.49
Biomolecular Science	93.39	70.61	382.74	10040.62	54.53	997.96
History	60.58	213.70	107.25	1359.50	58.50	657.34
Psychology	222.62	2558.24	524.21	47472.47	63.05	663.21

Table 3.5: Major unique, weighted, and intensity means and variances

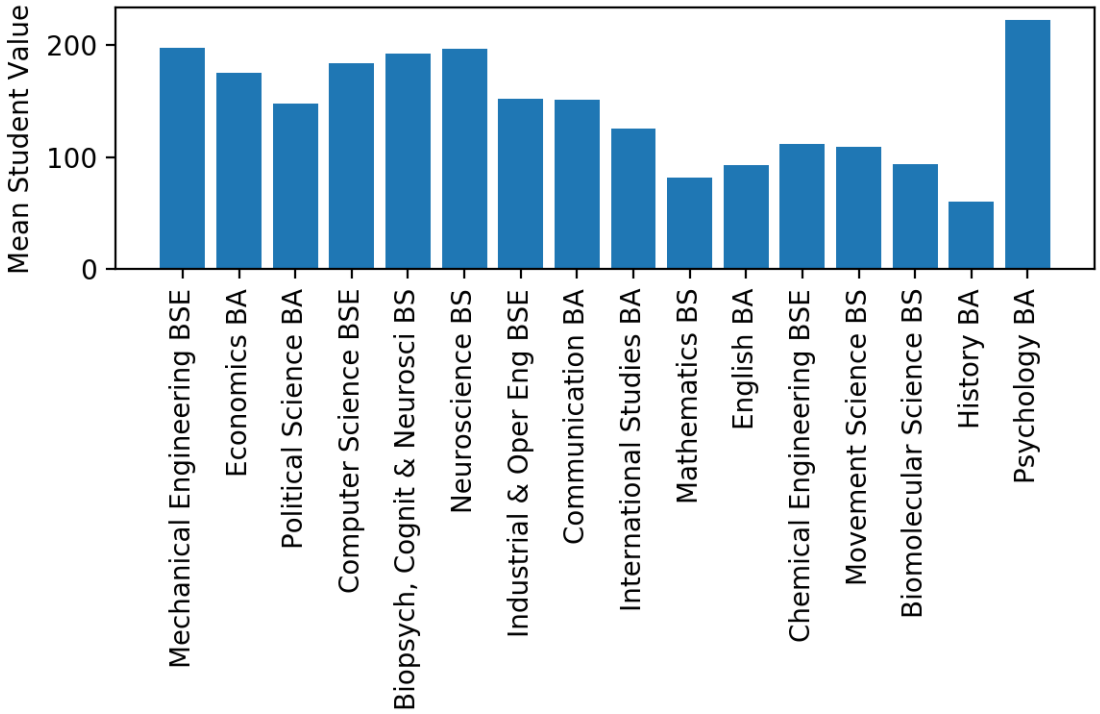


Figure 3.7: Average mean of student unique connections for each major

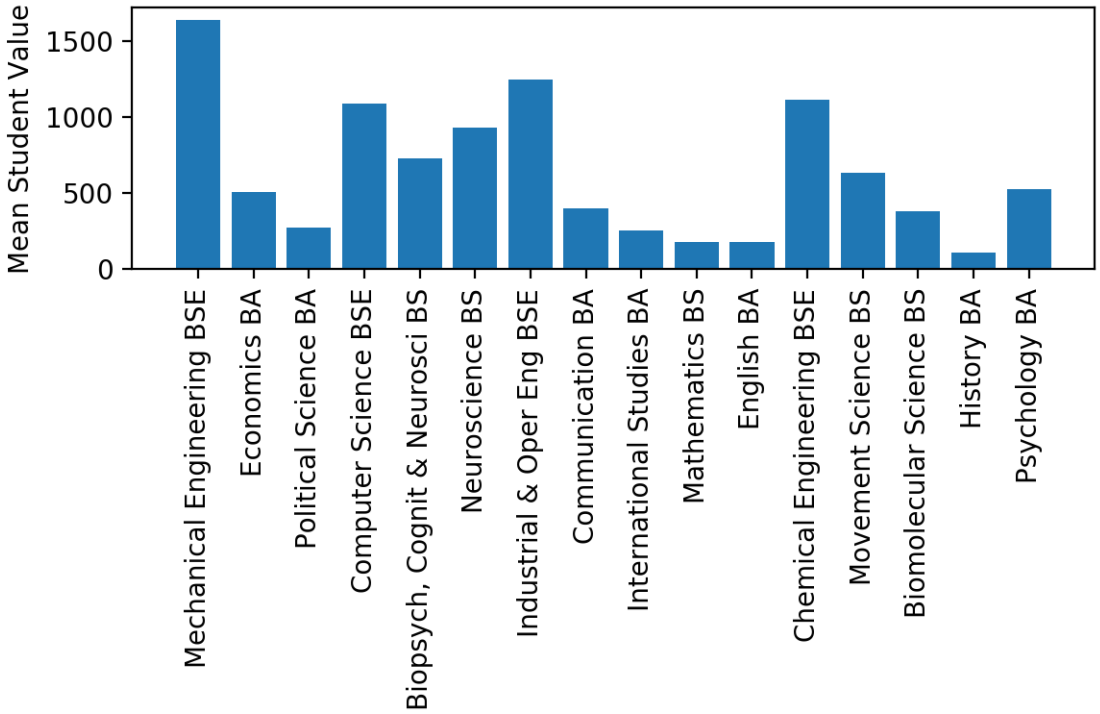


Figure 3.8: Average mean of student weighted connections for each major

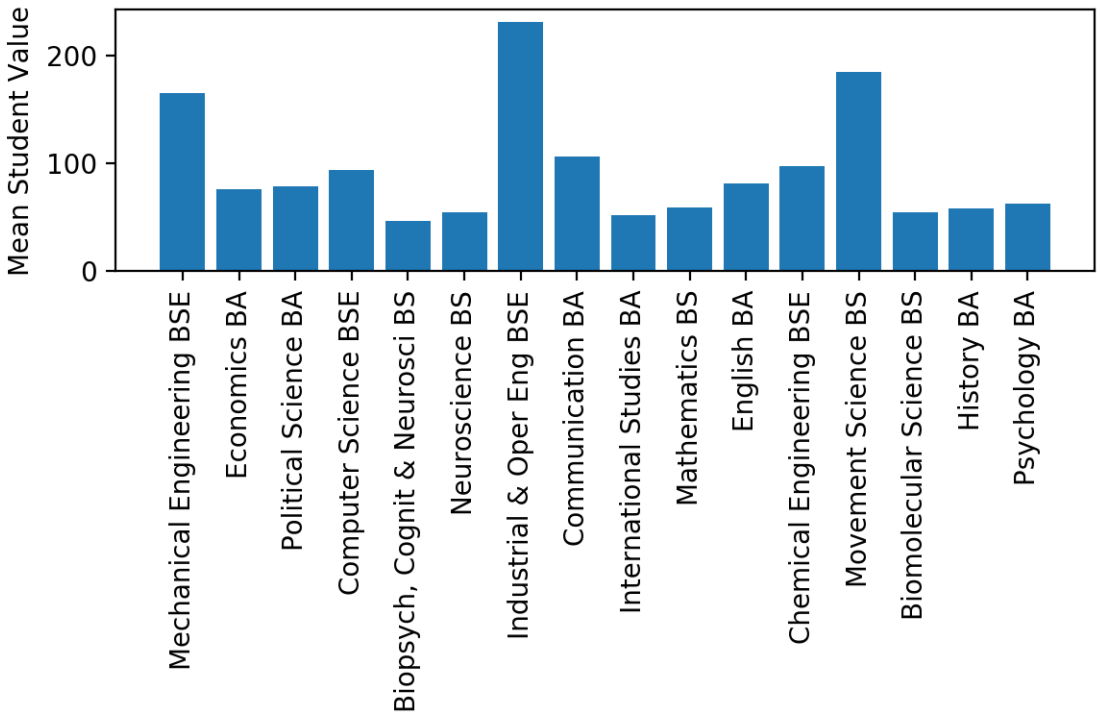


Figure 3.9: Average mean of student intensity connections for each major

3.4.2 Network Measures

We now examine how these three definitions effect common network statistics.

In this context, degree centrality is a measure of the connectedness of a student to other students. In practice, the connectedness can indicate the integrated influence of peers, or the converse. Strictly speaking, degree centrality is a measure of the number of connections, regardless of their weights. Figure 3.10 shows degree centrality measured by the unique connection student graph on the x-axis and the intensity student graph on the y-axis. Each point represents a student with degree centrality measured as represented by either graph. A feature of this plot is the rightward scattering from the line of equality. The weighting scheme, removes the effect of large courses, so that degree centralities in Ww can only be equal or less.

This line forms because of how degree centrality is calculated in NetworkX. NetworkX does not consider edge weight when calculating degree centrality, this means the number of times a pair of students take the same course doesn't matter (i.e., every edge is weighs the same). Since every edge weighs the same, the only attribute that will change the structure of the graph is the existence of an edge. Because the unique connection graph has an edge between two students independent of course size, it's an upper bound on the degree centrality measure. The intensity graph has a chance to have less edges because there is a chance two students are co-enrolled into a large course setting a weight value close to or equal to zero thus resulting in no edge connecting the two students.

Figure 3.10 is broken into four quadrants. Students whose unique and intensity degree centrality measures are closely related to one another are located near the line of equality. They took primarily courses with smaller enrollments, or at least small enough for an edge to be added. The farther a student scatters to the right, the greater number of large enrolled courses contributed to their unique degree centrality. Generally, those found further to the left took courses that had relatively lower

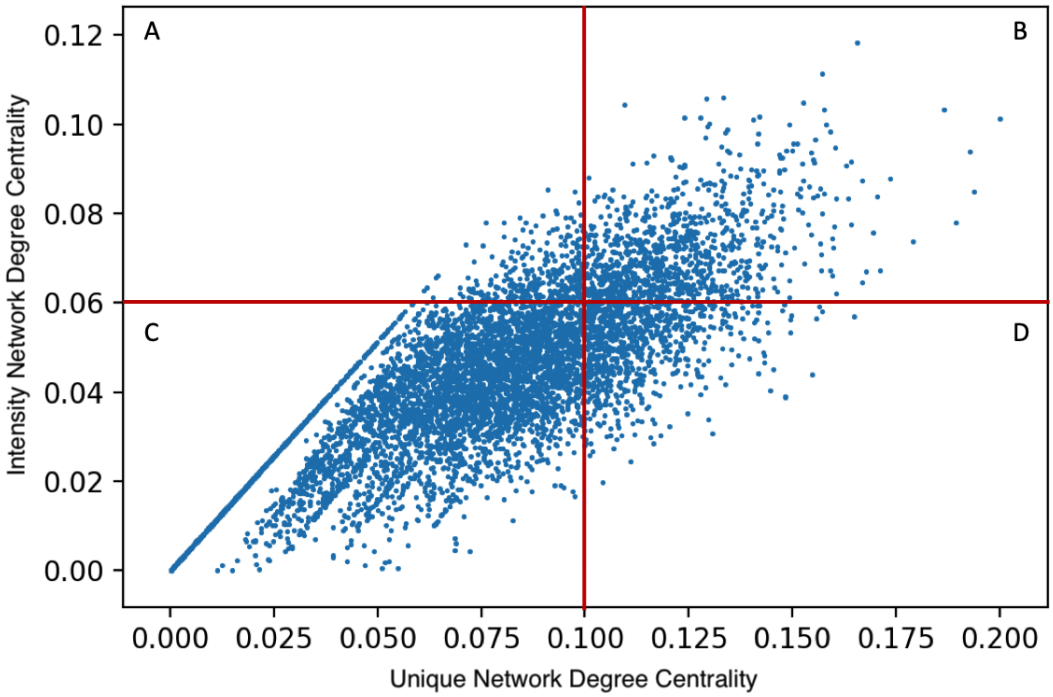


Figure 3.10: Student intensity degree centrality vs Student unique degree centrality

enrollments than students to the right.

Most notably, the upper right quadrant contains majors with high centrality for either measure. Pre-medical majors (Biopsych, Cognit & Neuroscience BS degrees and Neuroscience BS degrees) comprise more than a quarter of these. Their curricula include many prerequisite chemistry, biology, physics, and psychology courses that have large enrollement, however, not large enough where they lose their weight.

Students existing in the lower right corners of quadrants B, C, and D all are examples of students who had larger enrolled courses contribute to their degree centrality.

Examining A of Figure 3.10 (Intensity DC $>$ .06 and unique DC $<$ 0.1): Because of where the line that bounds degree centrality is located, the students that exist in A don't have that much mobility due to weighting. As with all the quadrants students in the bottom right exhibit the most mobility, but relative to the other quadrants this quantity is small.

In this section there are 328 students. The following table shows how they are

Major	Number of Students
Psychology	38
Computer Science	22
Business Administration	21
Biopsych, Cognit & Neurosci	17
Economics	13
Movement Science	12
Mathematics	11
Cellular & Molec Biology	10
International Studies	9

Table 3.6: Caption

distributed among the top 10 majors in this sections:

Part B (the top right) (Intensity DC $>$ .06 and unique DC $>$ 0.1): This quadrant contains the most mobility and spread. There are 973 students in this section. Majority of the students (4631) are located in part C of Figure 3.10. Bottom Right (Intensity DC $<$ .06 and unique DC $>$ 0.1)

Now we want to look at students whose intensity degree centrality almost equals the unique degree centrality. This is done by looking at students with weighted and unique values are within 1% of each other. This is displayed in the Figure 3.11.

3.4.3 Eigenvector Centrality

Like degree centrality, eigenvector centrality attempts to measure node importance. Below Figure 3.12 is a plot of intensity versus unique eigenvector centrality scores for students. Looking at the figure we see that the majority of the students are concentrated near zero, thus, for the first part of the analysis we looked at the outliers (students with intensity scores above 0.05), we then incrementally looked at slices approaching zero where majority of the students are. As displayed in the data below, moving through slices shows that students majoring in Dance, Performing arts, Musical Theatre, etc. get a significantly higher intensity eigenvector centrality score than other majors.

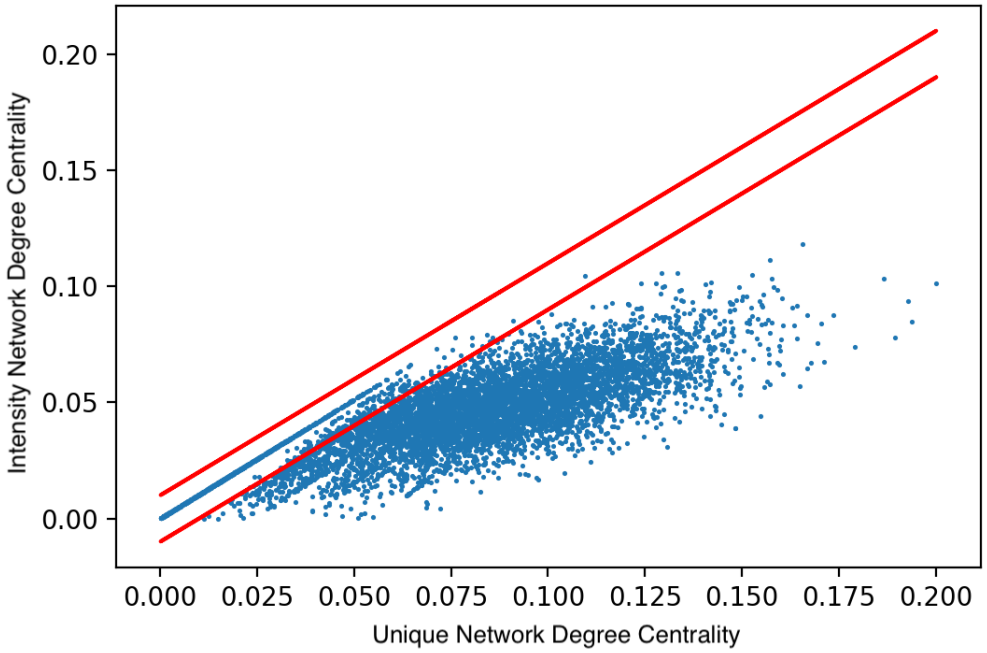


Figure 3.11: Highlight of intensity = unique portion of degree centrality graph

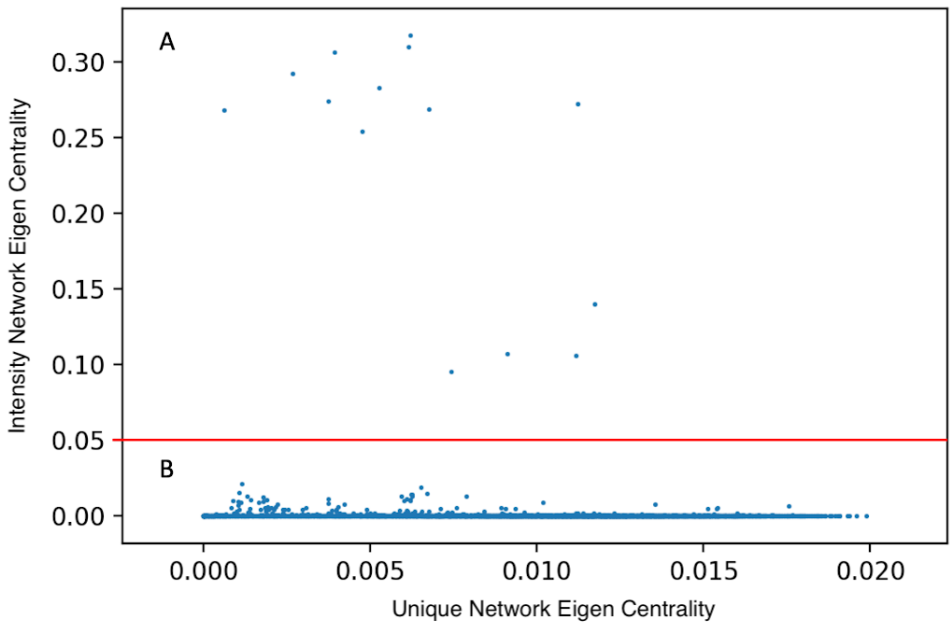


Figure 3.12: Intensity Network Eigenvector Centrality vs Unique Network Eigenvecotr Centrality

Intensity Eigen Range	Number of Students	Avg # Courses
$0.5 < EC$	14	59.64
$0.005 < EC < .05$	39	51.97
$0.001 < EC < .005$	66	43.94
$0.0001 < EC < .001$	264	41.01
$EC < .0001$	2365	32.36

Table 3.7: Slices of intensity network eigenvector centrality

This plot is broken up into two areas. Part A and Part B. Part A consists of all students with a intensity eigenvector centrality above 0.05. There are 14 students belonging to the following majors: Dance BFA (8), Psychology BA (2), no major (1), Business Administration BBA (1), Cellular & Molec Biology BS (1), Nursing BS Soph Transfer (1). The average number of courses taken by these students was 59.64. When we dive into these students a few things stand out. For the Dance BFA students majority of the Dance courses offered are less than 3 credits, so these students take more courses per semester. The Cellular & Molec Biology BS student double majored and also received a Dance BFA. The Nursing BS Soph Transfer student transferred from Dance. Though the remaining students didn't double major it would appear that they all started off as Dance BFA students because there first two semester was majority Dance courses. Changing their major probably didn't require transferring like the Nursing student had to

To analyze part B we'll zoom in the plot. First looks is .005 < intensity Eigen Score < .05. There are 39 students in this section. Continuing to zoom in .001 < intensity Eigen Score < .005 we add another 66 students. When we look at what these students study we notice that they are almost all from the Arts and Humanities. These students also have a higher number of courses than the remaining. Table 3.7 shows that students with the highest eigenvector centrality take 1.8x more courses on average than the students with near zero intensity eigen centrality.

3.4.4 Triangle Centrality

In NetworkX, Triangles is the number of triangles that include a particular node as a vertex. When we look at intensity Triangle vs Unique Triangles, it seems to follow the same pattern as degree centrality which makes sense. They both do not consider weight. As explained above, the only thing affect the sigmoid function can have in this result is the removal of edges in the adjacency matrix. Since the number of triangles that include a node is also proportional to the degree of the node, it makes sense that these two measures have similar patterns. However, it must be noted, that we see more spread in the triangle calculation, this is due to the large magnitude of the number, which allows for more variation. The removal of an edge in the normalized calculation has less an effect than it does on the triangle calculation.

3.5 Conclusion

This chapter explored how to define a connection in student enrollment data. Not only does this construction have a differential effect on majors, it significantly impacts the outcomes of various node measures on the network. First, when looking at the two example majors in Figure 3.5 and 3.6 is the extreme difference in correlation between the unique connection measure and the intensity connection measure. Table 3.8 shows the correlation for 16 majors. These majors were selected because they are the focus of Chapter IV for reasons explained there. In this table we see that the correlation varies by a large amount, from a positive 0.67 for English to a -0.66 for Chemical Engineering, to almost no correlation (0.02) for Neuroscience. Of these majors have a similar number of students enrolled in them. This highlights how majors with similar size enrollments can still exhibit differences in the courses students take in the number and intensity.

Figure 3.5 and 3.6 also highlight the drastic differences in student distributions

Major	Unique-Intensity Correlation
Chemical Engineering	-0.663
Computer Science	-0.14
Neuroscience	0.019
Biopsych, Cognit & Neurosci	0.040
Mechanical Engineering	0.069
International Studies	0.071
Biomelecular Science	0.211
Communication	0.260
Industrial & Oper Eng	0.285
Economics	0.315
Psychology	0.364
Political Science	0.461
Mathematics	0.525
Movement Science	0.598
English	0.668
History	0.673

Table 3.8: Correlation of majors unique and intensity connections

along the three modes of connection. Some majors have well-defined number of unique connections (e.g., Mechanical Engineer - Figure 3.3) while others (e.g., History -Figure 3.5) have a wider distribution. These figures also highlight the differences in typical course sizes and average number of courses you take with classmates.

Finally in this chapter we explored how the using the unique, weighted, and intensity graphs affect three different centrality measures. The measures examined are the degree centrality, eigenvector centrality, and triangle centrality. The important take away from this section was examining where students, and by proxy - majors, exist on the Intensity vs Unique network measure plot. For degree centrality (Figure 3.10) it was useful to divide the plot into quadrants. Doing this allowed for the easy identification of types of students, for example, pre-medical students are in the upper right corner of Figure 3.10. This also allowed us to define high-level characteristics of each quadrant. An example of this is that students located in section A of Figure 3.10 take courses with smaller enrollments. Exploring eigenvector centrality in Figure 3.12 highlights even more information about the students. Notice that students the average

number of courses is correlated with intensity measured eigenvector centrality. The students with the highest intensity measured eigenvector centrality have almost twice as many courses on average than those with the lowest. Upon further investigation we notice that students in the Arts tend to have a higher intesity measured eigenvector centrality.

CHAPTER IV

Clusters and Label Robustness

4.1 Introduction

Categorizing, lumping like with like, enables individuals to make sense of the world, organizations and teams to coordinate actions and scientists to derive causal and correlative relationship. Any given collection can often be categorized in a variety of ways. We can categorize people by age, gender, race, ethnicity, income or by any combination of those attributes. We can categorize cars by manufacturer or by body type, and student by major or grade point average.

The usefulness, or information content, of a categorization hinges on the relevant similarity of the objects within each category, the relevant dissimilarity of the objects in different categories, and whether variation in variables of interest exhibit a correlative or causal relationship with the categories constructed. For example, if we are trying to explain voter turnout, categorizing by age (in decades) reveals that older people are more likely to vote. If, instead, we categorized people by the color of their cars, we might find no differences in turnout across categories.

As mentioned above, categorizations also facilitate communication. A stock portfolio can be described by its allocation across growth and value stocks. The student enrollment data analyzed in this chapter includes majors as a category. Setting aside the possibility of double majors for the moment, we can interpret student majors as

a set of categories. When a college graduate describes themself as a physics major, they provide information about the types of courses they likely took and suggest that they have analytic skills.

The communicative function of categorizations creates stickiness. We may rely on a less than optimal legacy categorization because there exists a shared understanding of what the categories mean.

In this chapter, we evaluate the information content of college majors as a categorization. We evaluate whether they have become less informative over time by comparing them to categorizations derived using community detection (clustering) algorithms. We measure the informativeness of a categorization by calculating student similarity within categories and between categories.

Community detection algorithms decompose networks into communities such that within each community the network shows a high edge density but across communities there exists a low edge density.

There exist several types of algorithms to detect communities including cut methods, which divide the network into groups of specific sizes that minimize the number of connections between groups and modularity algorithms, which maximize a modularity function. Here, we use a spectral clustering algorithm and measure the distance between students by the cosine similarity between them. Spectral clustering and cosine similarity are described in the methods section.

4.2 Data

In this study, we use the same LARC dataset described in Chapter 2. Again we look at one cohort of students that enrolled in the University of Michigan in the fall semester of 2011. This cohort contains 6,738 students. Because we are interested in legacy labeling of students we examine the distribution of majors for this cohort. Figure 4.1 shows the top 40 majors.

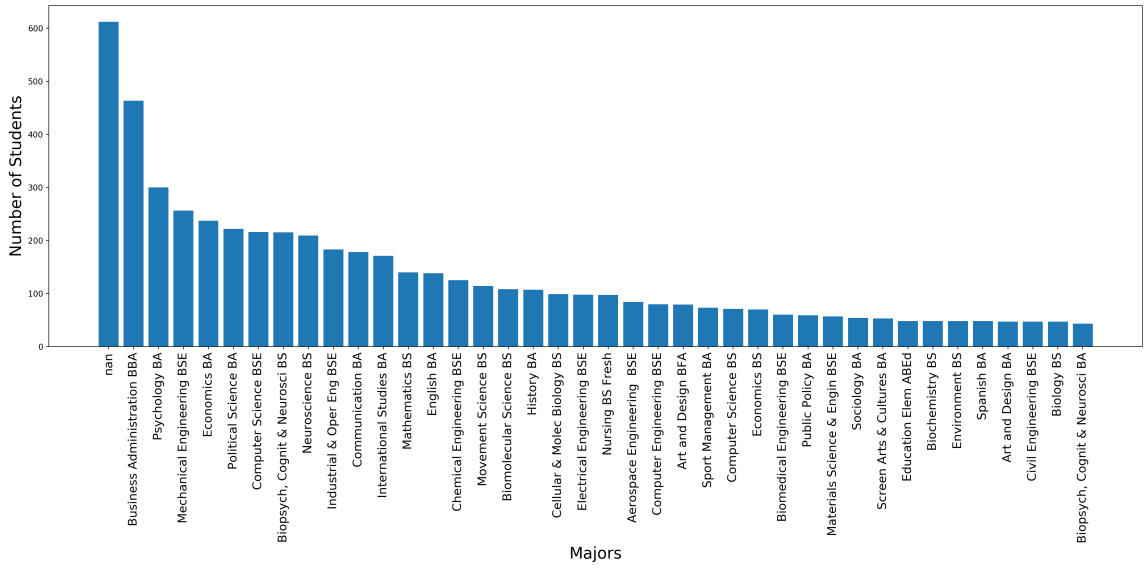


Figure 4.1: Top 40 Majors for the Fall 2011 cohort of students

For our analysis we consider only those majors that have a substantial and similar number of students. All the majors selected have at least 100 students and no more than 300. This range of enrollment was selected because we need group sizes to be comparable. The reason for this restriction will be come clear in Section 4.4.3, when we explore how students in certain majors are distributed among clusters. In brief, a major with a small enrollment could never be a majority of a cluster of significant size.

Our restriction results in a total of 16 majors and they are displayed in Table 4.1 along with their abbreviations and the number of students within each major. Restricting our focus to only students with these majors gives a total 2919 students in this study.

Now that the set of students has been established, we ask the question of "how do we go about categorizing them?" Here we compare three types of categorization: legacy labels, algorithmic (clusters identified by community detection algorithms), and random. Legacy labels refers to the traditional ways we group students in the

Major	Abbreviation	Number of Students
Psychology BA	PSY	300
Mechanical Engineering BSE	ME	256
Economics BA	ECN	237
Political Science BA	PS	222
Computer Science BSE	CS	216
Biopsych, Cognit & Neurosci BS	BCN	215
Neuroscience BS	NEU	209
Industrial & Oper Eng BSE	IOE	183
Communication BA	COM	178
International Studies BA	IS	171
Mathematics BS	MTH	140
English BA	ENG	138
Chemical Engineering BSE	CE	125
Movement Science BS	MS	114
Biomolecular Science BS	BMS	108
History BA	HIS	107

Table 4.1: 16 majors with their abbreviation and enrolment numbers

university setting. Algorithmic categorizing refers to partitioning the students based on a clustering algorithm that does not take into consideration legacy labeling. Random categorization is a method of putting the students into groups at random with the constraint of having the distribution match the legacy labeling.

The legacy categories can be captured at three levels. At the first level, which is the most abstract, the legacy category of BA or BS splits the student population into two groups. At second level, students are clustered into four groups based on whether a student focused in the humanities, social sciences, biological sciences, or natural sciences (H,S,B,N). This is done by assigning each major to the four categories as follows:

- Humanities (H): History BA, English BA
- Social Sciences (S): Economics BA, Political Science BA, Communication BA, International Studies BA, Psychology BA
- Biological Sciences (B): Biopsych Cognit & Neurosci BS, Neuroscience BS,

Movement Sciences BS

- Natural Sciences (N): Mechanical Engineering BSE, Computer Science BSE, Industrial & Oper Eng BSE, Mathematics BS, Chemical Engineering BSE

At the third, and finest categorization, students are identified by the sixteen majors.

4.3 Clustering Algorithms

In this section, we describe how we generated random and algorithmic categories of different granularity. We introduce the following notation *Algorithm(K)* refers to the categorization/clustering created by an algorithm constrained to create K clusters, and *Random(K)* refers to a categorization broken into K clusters with the same distribution as the historical categorization (e.g., Notice that in Table 4.2, Random (2) has the same distribution as BA-BS). Thus, at this level where we are splitting the students into two groups we use Algorithm(2) and Random(2).

Random(k) is implemented by taking the assignment of the legacy labels for student and randomly permuting the sequence.[1]

Before discussing Algorithm(k) we have to introduce the concept of student similarity. Our dataset contains a list of every course taken by the 2919 students. There exist 3620 possible courses, so each undergraduate can be represented as a binary string of length 3620, where each entry in the string corresponds to a course. If a student enrolled in a course, they are assigned a one for the corresponding entry in the string. Otherwise they are assigned a zero. Students take between 10 and 61 courses, so each student is represented a binary string of length 3620 with between 10 and 61 ones. Let A and B be two binary strings representing the courses taken by two

[1]We create random permutations using using the numpy library in python. The function numpy.random.permutation(sequence) returns a permuted version of the sequence. This method satisfies our requirement that Random(k) shuffles the assignments while keeping the same distribution.

students. One common way to measure similarity between vectors is cosine similarity. The cosine similarity between A and B is given by the following expression:

$$\text{CosSim}(A, B) = \frac{A \cdot B}{\mid A \mid \cdot \mid B \mid}$$

If students A and B took none of the same courses CosSim(A,B) = 0, and if they took identical courses then CosSim(A,B) = 1. We use this to measure the similarity between all of the students. The similarities are represented in a $N \times N$ affinity matrix, where $N = 2919$. Each element in this matrix represents the similarity between those two students.

Algorithm(k) uses this affinity matrix to split the students into k clusters based on their similarities. In this project we use spectral cluster for this task. Spectral clustering works by performing eigen-decomposition on the affinity matrix and using the eigenvectors with high eigenvalues to re-represent the affinity matrix. This new representation is in a lower dimensional space, where the dimensionality is equal to the number of large eigenvalues. The process of mapping the data to a lower dimensional space is known as embedding. The embedded data is easily clustered using a standard method, in our case k-means. The reason we use spectral cluster instead of k-means in the beginning, is because spectral cluster does not make any assumptions about the shape of the clusters and because the embedding process is not sensitive to initial conditions like most iterative methods are.

4.3.1 Clustering Distributions

We first present benchmark results on the distributions of students across the clusters. Table 4.2 shows that the algorithmic clustering creates clusters of much less equal sizes that the legacy labeling of BA-BS. This finding suggests that the BA-BS distinction may not be very informative, or at least not informative for some students.

Table 4.3 shows the categories when the students are broken into four groups. Here

BA/BS	Number of students	Algorithm(2)	Number of students	Random(2)	Number of students
BA	1353	Cluster 1	2122	Random 1	1353
BS	1566	Cluster 2	797	Random 2	1566

Table 4.2: BA/BS, Algorithm(2), and Random(2) student counts

again, the distribution across the four categories in the algorithmic categorization differ markedly from those in the legacy categories.

HSBN	Number of students	Algorithm(4)	Number of students	Random(4)	Number of students
H	245	Cluster 1	1314	Random 1	245
S	1108	Cluster 2	450	Random 2	1108
B	646	Cluster 3	897	Random 3	646
N	920	Cluster 4	258	Random 3	920

Table 4.3: HSBN, Algorithm(4), and Random(4) student counts

The last level splits the students into 16 categories. The legacy labeling at this level is the students major. The majors and the number of students in each major is displayed in Table 4.1. This level uses Algorithm(16) and Random(16). The Random(16) has the same student distribution as the majors. Algorithm(16) enrollment numbers are in Table 4.4/ This distribution of students also differ from the legacy category distribution. Notable, three categories have 30 or fewer students.

Algorithm(16)	Number of students	Algorithm(16)	Number of students
Cluster 1	106	Cluster 9	258
Cluster 2	28	Cluster 10	137
Cluster 3	223	Cluster 11	114
Cluster 4	154	Cluster 12	164
Cluster 5	632	Cluster 13	30
Cluster 6	321	Cluster 14	126
Cluster 7	156	Cluster 15	102
Cluster 8	17	Cluster 16	351

Table 4.4: Number of students in each of the 16 clusters created by Algorithm(16)

4.4 Results

In this section, we compare the legacy labels to Algorithm(k) which is "best" under a specific optimization function and to Random(k) which establishes a baseline.

4.4.1 How well does a labeling predict course agreement?

One aspect of labels that are often applied to students is that they represent the courses a student took throughout there time at the university. Using the example from earlier, when a student mentions they majored in Physics, you can imagine the courses they took while in school (e.g., Electricity and Magnetism, Quantum Mechanics, Thermodynamics, etc...).

Taking this into consideration, we first explore the following questions. What is the number of courses in common for two students belonging to the same label? What is the number of courses in common between two students from different labels? We will call the answer to the first question *Agreement* and the answer to the second question *Agreement Across*.

Agreement for a labeling can be found by first separating the students into their respective labels. Then for every pair of students belonging to a specific label we sum the number of courses in common. Agreement equals the weighted average of the number of courses in common for each label and is equal to

$$\text{Agreement}_{\text{label}} = \sum_{i=1}^{k} \frac{n_{label_i}}{N} \sum_{j,z} \frac{1}{n_{label_i}} nc_{j,z}$$

where $nc_{j,z}$ is the number of courses student j and student z had in common. n_{label_i} is the number of students that are in label i and N is the total number of students. The Agreement Across for a labeling is similar except students j and z belong to different labels.

The following three tables show the Agreement and Agreement Across for the three

levels of categories. Higher Agreement and lower Agreement Across imply a stronger clustering. For a random clustering the two measures should be approximately equal, which is true in each table.

Classification	Agreement	Agreement Across
BA-BS	4.395	1.800
Algorithm (2)	6.388	1.602
Random (2)	3.216	3.218

Table 4.5: Agreement and Agreement Across for BA-BS Classification

Classification	Agreement	Agreement Across
H,S,B,N	6.459	2.003
Algorithm (4)	10.430	2.548
Random (4)	3.182	3.195

Table 4.6: Agreement and Agreement Across for H, S, B, and N Classification

Classification	Agreement	Agreement Across
16 Majors	11.395	2.706
Algorithm (16)	10.411	1.706
Random (16)	3.202	3.223

Table 4.7: Agreement and Agreement Across for Majors as a Classification

The Tables 4.5, 4.6, and 4.7 shows how much agreement a labeling method has. Agreement is our first order measure for how good a labeling method is compared to another. This is because we expect students in the same group to have a higher number of courses in common. Table 4.5 shows the results for the legacy labels of BA and BS. If we were to pick two students from the same set, either BA or BS, they would only have slightly more agreement than if two students were picked at random. This indicates that using the legacy labels of BA and BS might not be appropriate in the context having an approximation to what courses they have taken at the university. However, we do see a jump in Agreement when we look at Algorithm(2). If two students are picked from one of the two clusters created by Algorithm(2) they will have on average 6.3 courses in common.

An alternative way to interpret these results imagines the random clustering as a worst case sorting and the algorithmic clustering as a best case. The Agreement of any clustering should lie in the interval $[3.21, 6.38]$. The BA-BS legacy labels have an Agreement of 4.395 are closer to random than optimal. This implies that the legacy labels are not particularly informative.

Table 4.6 has the results for the H, S, B, N agreement. Here we see a jump in agreement in the legacy labeling, going from BA/BS to H, S, B, N indicating there is a little more signal in this labeling comparatively. It is interesting to see that using Algorithm(2) has about the same Agreement is legacy labeling H, S, B, N. Even though H, S, B, N is increasing the resolution (going from 2 groups to 4), we do not see an improvement on agreement with Algorithm(2). However, H, S, B, N is about twice better than Random(4). Algorithm(4) has the best Agreement among the labelings at this level, it also improves on Algorithm(2).

Table 4.7 shows how well majors do at predicting commonality in courses. Here we see that major labeling has the highest Agreement, however, it is just marginally better than Algorithm(16). That being said, we do not see an improvement from Algorithm(4) to Algorithm(16). Agreement Across gives us some more information though. As mentioned earlier, a labeling should have high Agreement and low Agreement Across. While the major labeling shows an improvement in Agreement from the previous legacy labelings, we see an increase in Agreement Across. Keeping that in mind notice that Algorithm(16) has the same Agreement as Algorithm(4), but has a decrease in Agreement Across. This phenomenon is due to the averaging that is used to calculate Agreement and Agreement Across. Some labels within a classification scheme may be better than others (e.g., some majors may have be good predictors of Agreement while others poor). Later in section 4.4.3 we will explore this in more detail.

4.4.2 How do legacy labels compare to algorithmic labels?

The previous section showed how well a label performs at predicting the number of courses in common students will have. This section will explore how similar/different are the legacy labels to Algorithm(k). To measure the differences, we use two standard measures, *Rand Index* (RI) and *Normalized Mutual Information* (NMI) along with their adjusted values.

We compare the historical classifications and random classifications to those found by the spectral clustering algorithm. To define RI and NMI we use the following construction. Given a set of N objects $S = \{o_1, o_2, \ldots, o_N\}$ and a clustering of S into R non-overlapping subsets $U = \{U_1, U_2, \ldots, U_R\}$. Given another clustering of S into C non-overlapping subsets $V = \{V_1, V_2, \ldots, V_C\}$ we compare the two partitions U and V. Both U and V have the following properties. $\cup_{i=1}^{R} U_i = S = \cup_{j=1}^{C} V_j$ and $U_i \cap U_{i'} = \emptyset = V_j \cap V_{j'}$ for $1 \leq i \neq i' \leq R$ and $1 \leq j \neq j' \leq C$. This results in the following $R \times C$ contingency table:

$U \setminus V$	V_1	V_2	$\ldots$	V_C	sums
U_1	n_{11}	n_{12}	$\ldots$	n_{1C}	a_1
U_2	n_{21}	n_{22}	$\ldots$	n_{2C}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
U_R	n_{R1}	n_{R2}	$\ldots$	n_{RC}	a_R
sums	b_1	b_2	$\ldots$	b_C	$\sum_{ij} n_{ij} = N$

Table 4.8: Contingency Table, $n_{ij} = |U_i \cap V_j|$

When we compare the $\binom{N}{2}$ pairs of objects (students) in the S there are four different results that can occur. These four different results are equivalent to the four categories of a confusion matrix:

- type(i) or True Positive (TP): the pair are in the same subset of U and in the same subset of V
- type(ii) or True Negative (TN): the pair are in different subsets of U and in

different subsets of V

- type(iii) or False Negative (FN): the pair are in different subsets of U and the same subset of V
- type (iv) or False Positive (FP): the pair is in the same subset of U and in different subsets of V

Using the above set we are now ready to define RI and NMI. RI is defined as the number of TP and TN divided by the total number of pairs. $\text{RI} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{TP+TN}{\binom{N}{2}}$ using the elements of the Table 4.8 we can define RI as:

$$\text{RI} = \left(\binom{N}{2} + 2\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \left[\sum_{i=1}^{R}\binom{a_i}{2} + \sum_{j=1}^{C}\binom{b_j}{2}\right]\right) \Big/ \binom{N}{2}$$

RI can be thought of in probabilistic terms, $(TP + TN)/\binom{N}{2}$ is the probability of the two partitions U and V agree on the classification of a pair of elements from S. The Adjusted Rand Index (ARI) is an correction to the RI by taken into account agreement that occurred by chance. This correction is done by using the expected value of the RI. $\text{ARI} = (RI - \mathbb{E}(RI))/(\max(RI) - \mathbb{E}(RI))$. Using terms from Table 4.8 we get the following equation:

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_{i=1}^{R}\binom{a_i}{2}\sum_{j=1}^{C}\binom{b_j}{2}\right] \Big/ \binom{N}{2}}{\frac{1}{2}\left[\sum_{i=1}^{C}\binom{a_i}{2} + \sum_{j=1}^{C}\binom{b_j}{2}\right] - \left[\sum_{i=1}^{R}\binom{a_i}{2}\sum_{j=1}^{C}\binom{b_j}{2}\right] \Big/ \binom{N}{2}}$$

This construction of expected value is under the null hypothesis that the contingency table is randomly generated from a permutation model of clustering. Specifically, it is constructed from the generalized hyper geometric distribution which assumes partitions of U and V are picked at random but have the original number of classes and objects in each. RI has values with a range of $[0, 1]$, but [34] point out that empirically RI has a high baseline and values typically fall under the narrow

range of $[0.5, 1]$. This is exactly what we experience in our studies. ARI however can go negative if the expected value is greater than the given value.

Before characterizing NMI, we must introduce two additional concepts. The first is entropy, H. Entropy is a measure of uncertainty in a random variable or the amount of information required on average to describe the random variable. Using the partition U, from above, entropy is defined as $H(U) = -\sum_{i=1}^{R} \frac{a_i}{N} \log \frac{a_i}{N}$. The second concept needed is Mutual Information, I, between two variables describes the amount of information that one random variable contains about another random variable. Mutual information is symmetric $I(X, Y) = I(Y, X)$. The mutual information between the two partitions U and V is given by $I(U, V) = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{n_{ij}}{N} \log \frac{n_{ij}/N}{a_i b_j / N^2}$. As the name suggest, NMI is the normalized version of I. The normalization sets the range of values to lie within a fixed range $[0, 1]$. NMI can then be defined as:

$$\text{NMI} = \frac{2 * I(U, V)}{H(U) + H(V)}$$

Similar to the RI, NMI can be corrected for chance by incorporating the expected value, this is called the Adjusted-for-Chance Mutual Information (AMI)

$$\text{AMI} = \frac{I(U, V) - \mathbb{E}(I(U, V))}{\frac{1}{2}[H(U) + H(V)] - \mathbb{E}(I(U, V))}$$

The expected mutual information $\mathbb{E}(I(U, V))$ between two clusters U and V is given by

$$\mathbb{E}(I(U, V)) = \sum_{i=1}^{R} \sum_{j=1}^{C} \sum_{n_{ij}=max(a_i+b_j-N,)}^{min(a_i,b_j)} \frac{n_{ij}}{N} \log \frac{N n_{ij}}{a_i b_j} \frac{a_i! b_j! (N-a_i)! (N-b_j)!}{N! n_{ij}! (a_i - n_{ij})! (b_j - n_{ij})! (N - a_i - b_j + n_{ij})!}$$

Table 4.9 shows the values of these three measures when we partition the students into two groups. We see low values for the ARI, NMI, and AMI indicating that using

the legacy labels of BA and BS differs from using a clustering algorithm to split the students into two groups based on the courses they've taken.

This provides even stronger evidence that using BA and BS may not be the best way to represent students if it is intended to be an abstraction of the courses taken. This finding is in agreement with the data from Table 4.5 which shows relatively low agreement among the students within BA/BS clusters especially when compared to the Algorithm(2) and Random(2). BA/BS agreement is closer to the Random(2) partitioning. When we look at the second row of Table 4.9 comparing the Random(2) to the Algorithm(2) we see very little agreement, close to 0 and even negative for ARI, as expected.

Classification	RI	ARI	NMI	AMI
BA-BS to Algorithm (2)	0.609	0.216	0.318	0.294
Random (2) to Algorithm (2)	0.500	-0.001	0.001	0.001

Table 4.9: Distance to Algorithmic Clustering BA-BS Classification

Another way to compare BA/BS to Algorithm(2) is to look at the intersection of BA students with cluster 1 and cluster 2 and also BS students with cluster 1 and cluster 2. This can be displayed in the following contingency table:

	BA	BS	sum
cluster 1	1348	774	2122
cluster 2	5	792	797
sum	1353	1566	

Looking at the contingency table we see that the two clusters do a great job at splitting BAs. Majority of the BA students were assigned to cluster 1. However, 36% of the cluster 1 are BSs. We also see the the two clusters split BS evenly. It's important to note that this does not declare right or wrong, but lack of agreement between the two.

Table 4.10 examines the H,S,B,N labels we typically attribute to students and compares it the Algorithm(4). This comparison shows an increase in the agreement. The contingency table showing the intersection of the sets is in Table B.1 in the Appendix.

Classification	Rand Index	Rand Adjusted	NMI	AMI
H,S,B,N to Algorithm (4)	0.789	0.511	0.542	0.531
Random (4) to Algorithm (4)	0.568	-0.002	0.0003	-0.001

Table 4.10: Distance to Algorithmic Clustering H, S, B, N Classification

Table 4.11 shows an significant increase in agreement. Using the more granular major to label groups of students aligns closely with the clustering algorithm. All of these results align with our intuition, however when we examine these results more closely we notice that not every major makes for a good label. This is demonstrated in Table B.2 in the Appendix

Classification	Rand Index	Rand Adjusted	NMI	AMI
16 Majors to Algorithm (16)	0.935	0.588	0.777	0.737
Random (16) to Algorithm (16)	0.842	0.001	0.014	-0.001

Table 4.11: Distance to Algorithmic Clustering Majors as a Classification

4.4.3 How robust are legacy labels to the number of clusters?

For the remainder of the study we will restrict our analysis of legacy labels to majors only. This is due to the relative performance of the other two legacy labels. BA/BS performed poorly (near random) in predicting agreement and have very low values of ARI and AMI when comparing it's partitions to optimal clustering partitions. While H,S,B,N performed better as a legacy label than BA/BS, on closer examination they do not hold up. Looking back at Table 4.6, we notice an increase in predicting agreement, however, this measure is still closer to random than the algorithm. With this result and the fact that majors performed well on all of these task they had the best chance of performing well with the more strict criteria.

In this section we explore the "robustness" of a legacy label with respect to a clustering algorithm. We restrict our analysis to majors since they performed the best in the analyses from the earlier sections.

In this section we answer the question, what would a "good" cluster look like? To do this we introduce some additional concepts from the ones mentioned above to help us measure the relationship between a legacy label and clusters defined by an algorithm. The first two concepts are conditional probabilities defined by a major (e.g., M_{Major}) and a cluster (e.g., C^k_{Major}).

The following example explains how to read these terms, $C^{k=4}_{\text{PSY}}$ states that of the 4 clusters generated by the Algorithm(4), C_{Major} contains more PSY majors than any other cluster. Thus, the two probabilities are the following $P(M_{\text{Major}}|C^k_{\text{Major}})$ and $P(C^k_{\text{Major}}|M_{\text{Major}})$. In words $P(M_{\text{Major}}|C^k_{\text{Major}})$ ask given the cluster that contains the largest fraction of a specific major, what is the probability that major is selected from the elements of that cluster? While $P(C^k_{\text{Major}}|M_{\text{Major}})$ is asking given that an element belongs to a specific major, what is the probability it is in the cluster that contains the largest number of those majors.

Another measure of interest is the *coherence* of a major M_{Major} which is given by the following equation:

$$\text{Coherence}_{Maj} = \sum_{i=1}^{k} \frac{N_{Maj,C_i}}{N_{Maj}} \frac{N_{Maj,C_i}}{N_{C_i}}$$

Where N_{Maj,C_i} is the number of students with major Maj that are in cluster i. N_{Maj} is the number of students in with major Maj and N_{C_i} is the number of students in cluster i.

The first part of this analysis examines how Coherence_{Maj} and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ changes changes with k (number of clusters). This is shown in Figure 4.2. Note that the % in largest (the green line) is referring to $P(M_{\text{Maj}}|C^k_{\text{Maj}})$. Figure 4.2 only shows the result of four majors the others can be found in Appendix B.2. After some

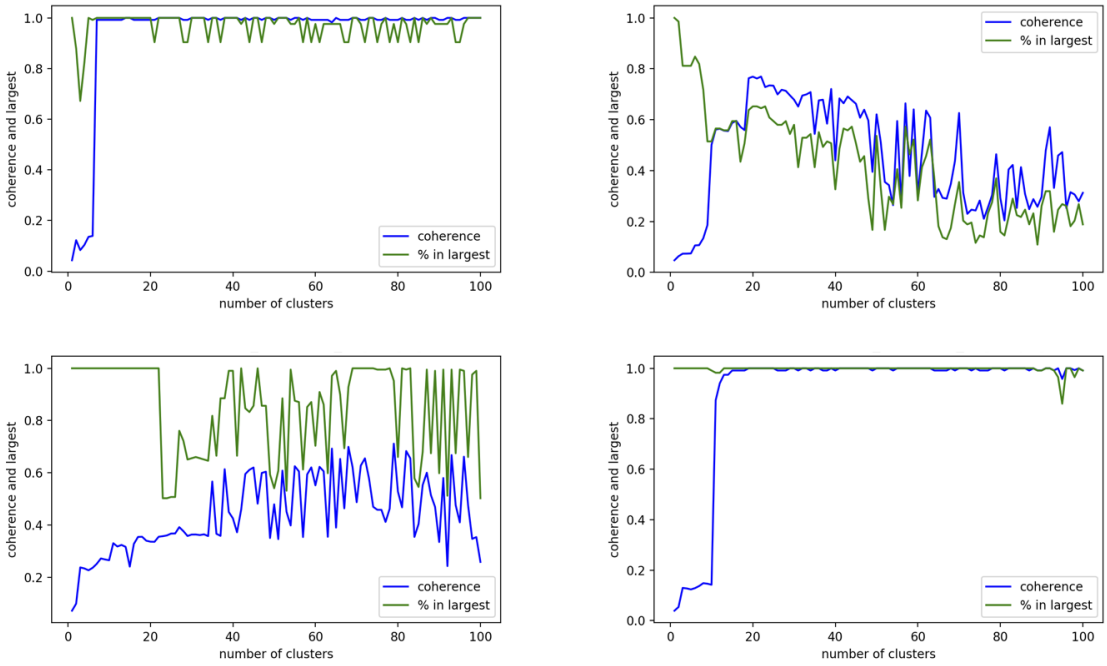


Figure 4.2: Coherence$_{Maj}$ and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ change with k

minimum number of clusters both Chemical Engineering and Movement Science have high values with little fluctuations.

When we look at English and Neuroscience we see very different behavior, not only are the values lower, they vary significantly for each value of k. First, looking only at $P(M_{\text{Maj}}|C^k_{\text{Maj}})$, we see that not only are majority of the Movement Science students in the same cluster, they constitute a majority of that cluster. We also know that this is not affected by k. This is a different story when it comes to English and Neuroscience. For English we notice a downward trend as k is increased. This highlights a trend we have seen throughout this thesis, majors show a differential repsonse to certain measures and forumulations. All of these come together in the measures we introduce in the remainder of this thesis. For higher resolution images of these plots and to see the Coherence$_{Maj}$ and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ for other majors please see Appendix B.2 where all 16 majors indivual plots are located.

Another approach to understanding the robustness of a label when compared to

a clustering algorithm is to ask how recoverable is a major. We call this measure *Recoverability R*. We define two types of recoverability, weak recoverability WR and strong recoverability SR. R is defined by the following equation:

$$R_k(X,Y) = \begin{cases} 1, & \text{if } P(C^k_{\text{Major}}|M_{\text{Major}}) > X \text{ and } P(M_{\text{Major}}|C^k_{\text{Major}}) > Y, \\ 0, & \text{otherwise.} \end{cases}$$

In the analysis below we use the following definitions $WR_k = R_k(0.6, 0.6)$ and $SR_k = R_k(0.8, 0.8)$. Figure 4.3 shows how Economics two conditional probabilities varies with k. The two red lines at probability 0.6 and probability 0.8 represent where the cut offs for weak recoverability and strong recoverability are respectively. To be weakly recoverable both the lines need to be above the dashed red line at 0.6. The major is only strongly recoverable for values of k where both lines are above the line at 0.8. Recoverability looks at the symmetry between $P(C^k_{\text{Major}}|M_{\text{Major}})$ and $P(M_{\text{Major}}|C^k_{\text{Major}})$. For a major to be considered recoverable more than 60% (80%) of the students with that major need to be in same cluster after the clustering algorithm is applied to the enrollment data.In addition these students also need to be at least 60% (80%) of the constituents assigned to that cluster.

Figure 4.4 and Figure 4.5 displays the strong and weak recoverabilty respectively. This analysis goes through each major and ask if it is weakly recoverable, strongly recoverable, both, or neither, for $k = 1$ to $k = 50$ clusters. Figure 4.4 shows that all of the engineering majors are strongly recoverable majority of the time, with the highest given to IOE and ME which both are strongly recoverable 46 out of the 50 times. Figure 4.4 also shows that seven majors never meet the criteria of strong recoverability.

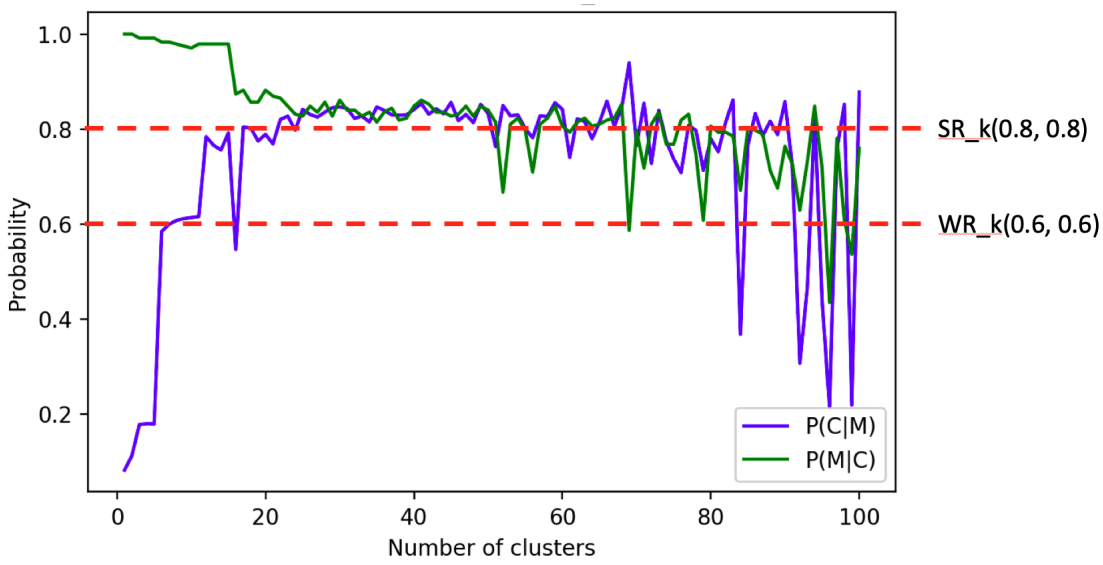


Figure 4.3: $P(M_{\text{ECN}}|C^k_{\text{ECN}})$ and $P(C^k_{\text{ECN}}|M_{\text{ECN}})$ with respect to k

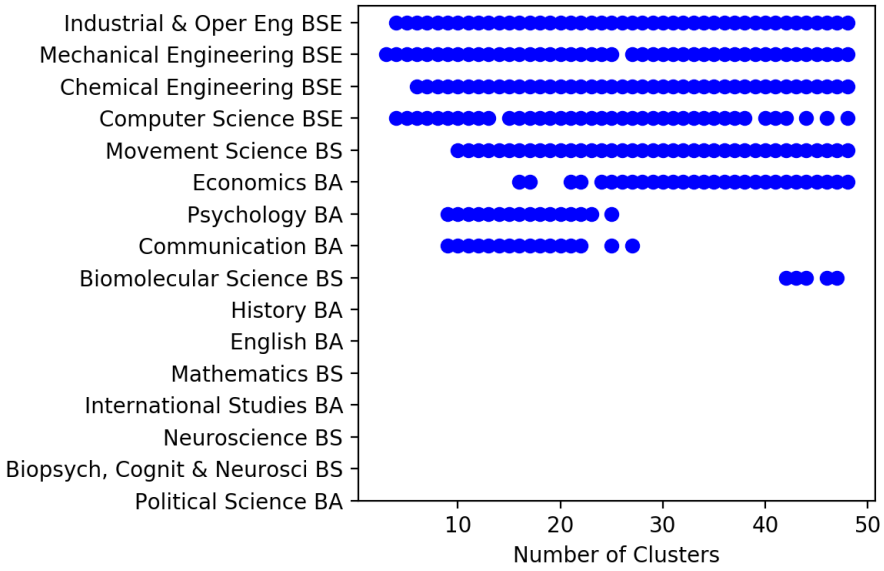


Figure 4.4: Strong recoverability (SR) for $k = 1 : 50$ clusters

Given that weak recoverability is more easily satisfied than strong recoverability, there are more occurrences of recoverability than strong recoverability. As shown in Figure 4.5 only two majors, HIS and BCN are not weakly recoverable for at least some values of k

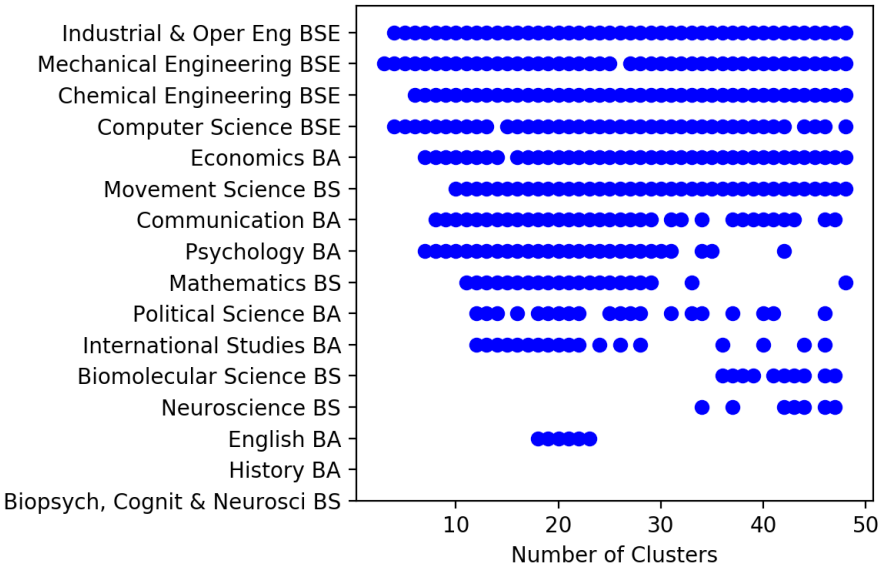


Figure 4.5: Weak recoverability (WR) for $k = 1 : 50$ clusters

Figure 4.6 shows the number of majors that are strongly recoverable and weakly recoverable as a function of the number of clusters k. If you take Figures 4.4 and 4.5 and for each k placed a vertical line, this Figure 4.6 displays the number of points the vertical line will intersect. This plot is useful because it highlights a range for number of clusters where majors are recoverable. Keeping in mind that there are 16 majors, we see that we need at least 10 clusters for more than 50% of the majors to be weakly recoverable. This makes sense because below a certain number of clusters it is impossible for a major to be greater than 60% of the clusters members. This is due to the fact that with a low number of clusters you are forcing students into a lower dimensional space.

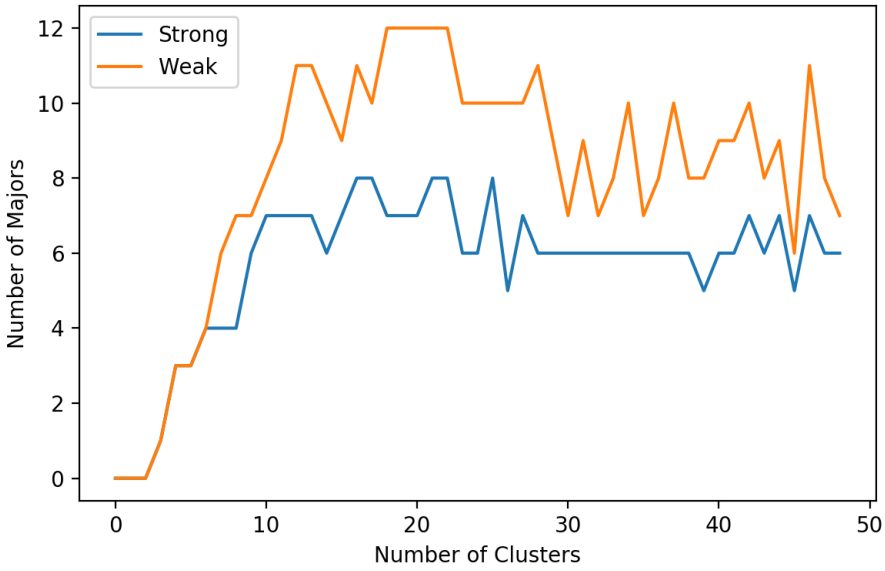


Figure 4.6: Major counts for SR and WR for a given k

4.4.4 Recoverablity of Algorithm(16) Categories

In this section, we compare the recoverability of the clusters produced by Algorithm(16) to the legacy labels. This process works by running spectral clustering on the original enrollment data to generate 16 clusters. We use these cluster assignments to replace the major as the "true" label. We then run this relabeled data through the recoverability process for $k = 1 : 50$. Both Figures 4.7 and 4.8 show a significant increase in recoverability.

As shown in Figure 4.7 every label is strongly recoverable at least once! Figure 4.8 shows that in the relabeled paradigm, not only are all the majors weakly recoverable, but they are weakly recoverable for a wide range of k.

The significant increase in both strong and weak recoverability due to the relabeling suggest that reexamining legacy labels is an important task. It is important to note that there exist a mapping between our legacy labels and our new Algorithm(16) labels. Taking into consideration how the legacy labels are distributed among the clusters and the recoverability of the new labels helps in identifying a balance between

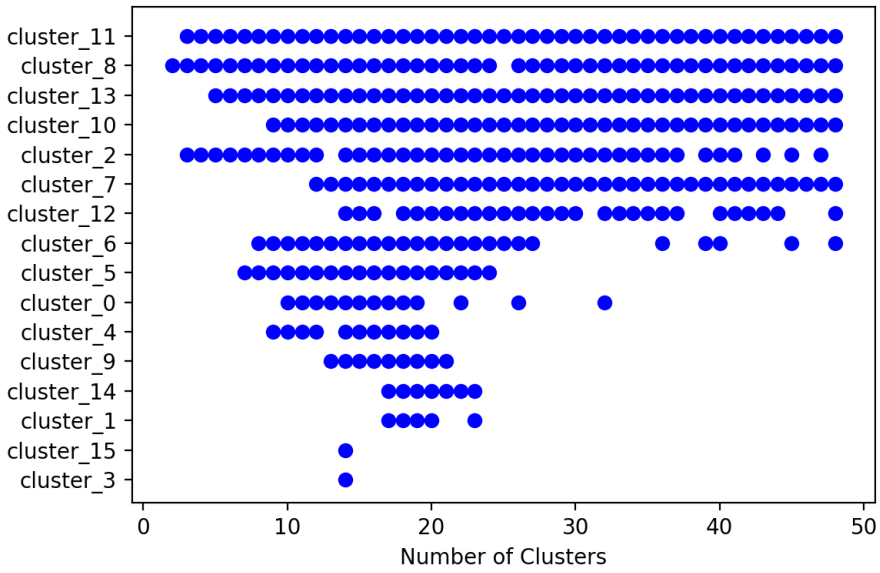


Figure 4.7: Algorithm(16) SR_k

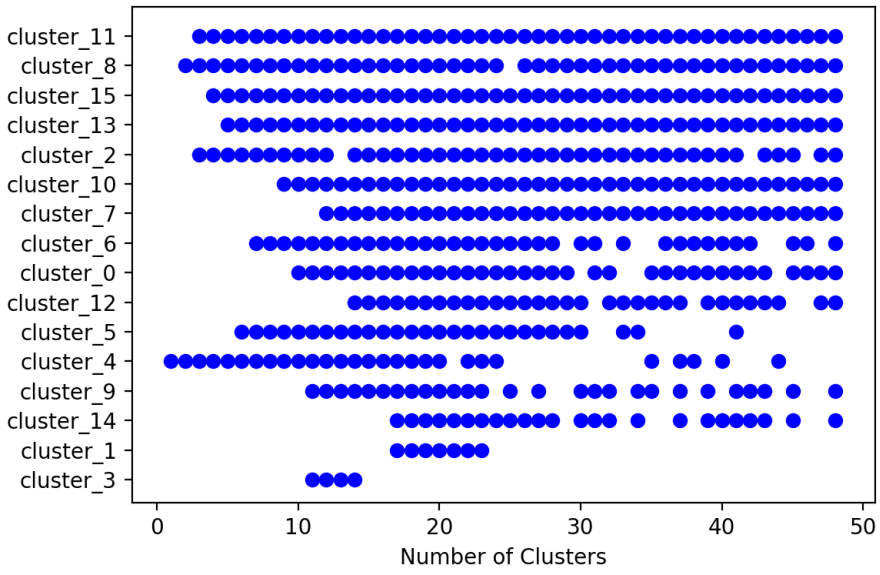


Figure 4.8: Algorithm(16) WR_k

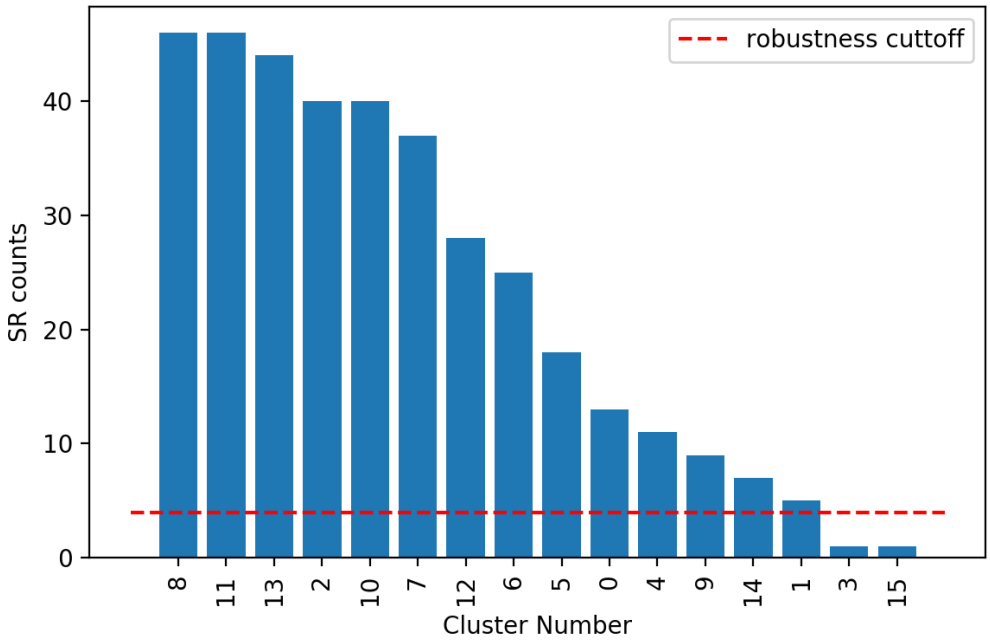


Figure 4.9: Number of times a cluster is strongly recoverable. Robustness cutoff to drop last two

complete algorithmic approach and sticking with what you know.

Due to Algorithm(16) great performance on recoverability, we can use it to find a cutoff to define robustness. We Figure 4.9 and 4.10 to define a cutoff for recoverability. We define the cutoff at the top 85% (i.e., top 14 of the 16). Thus for Figure 4.9 we use cluster 1 as our cutoff. Cluster 1 is strongly recoverable 5 times. This will set our robustness on the strong recoverability graph at 4. We can now use this value to see what majors meet this criteria. This same process can be applied to the weakly recoverable threshold. We see from Figure 4.10 that the cutoff is at cluster 14. Cluster 14 is weakly recoverable 23 times. We can use this our robustness cutoff to see what majors meet this critera. The robustness of the majors are show in Figures 4.11 and 4.12. This results in eight strongly and weakly recoverable majors

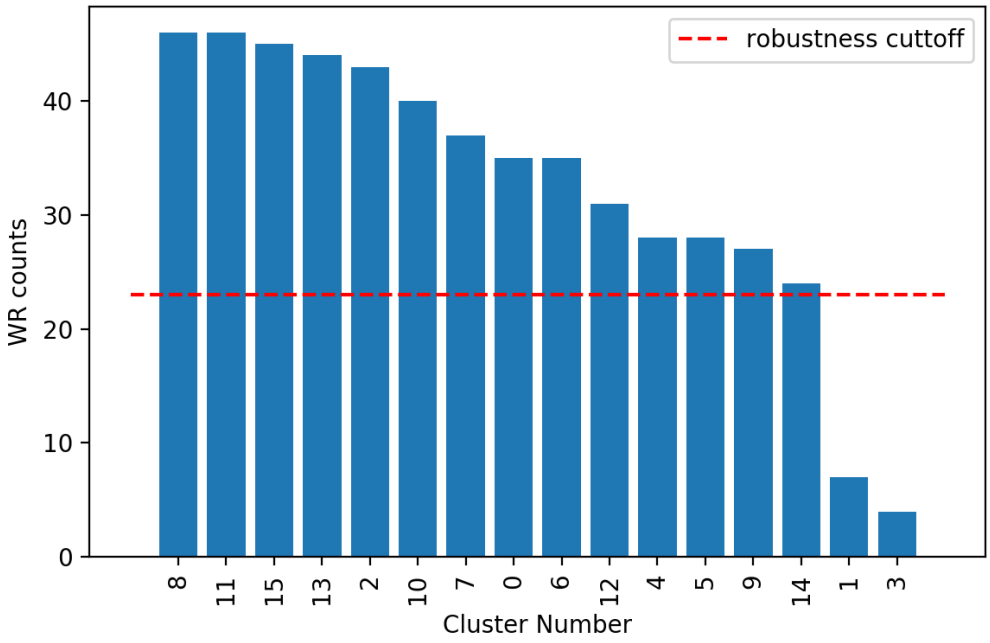


Figure 4.10: Number of times a cluster is weakly recoverable. Robustness cutoff to drop last two

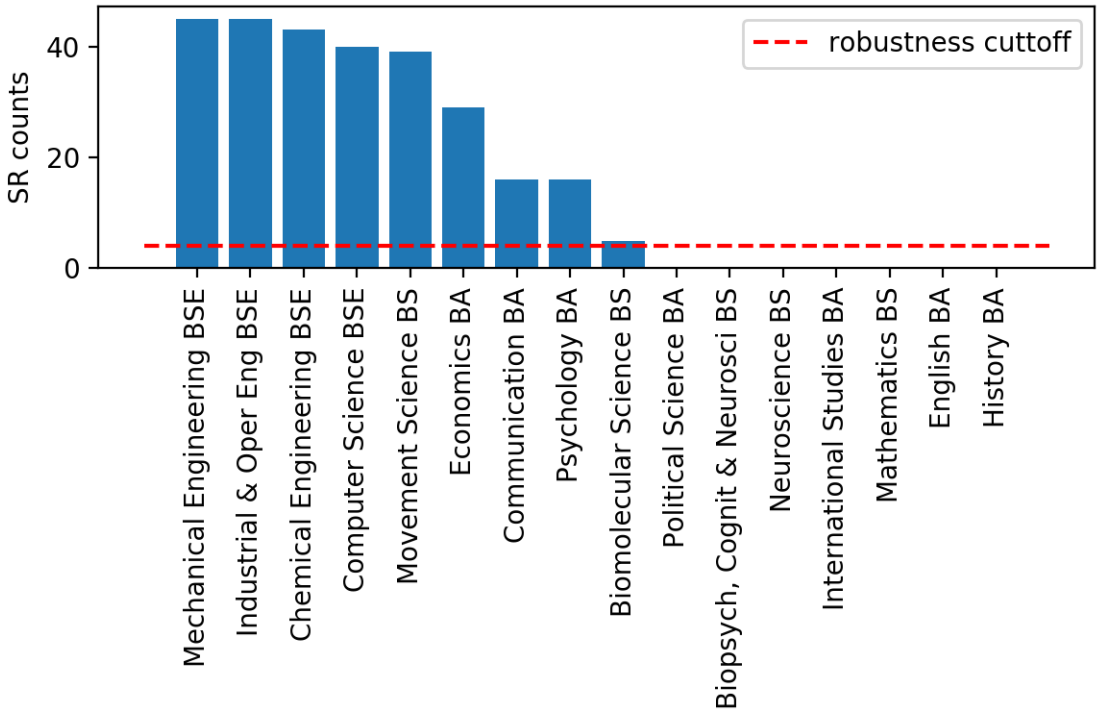


Figure 4.11: Major strong recoverability with robustness cutoff

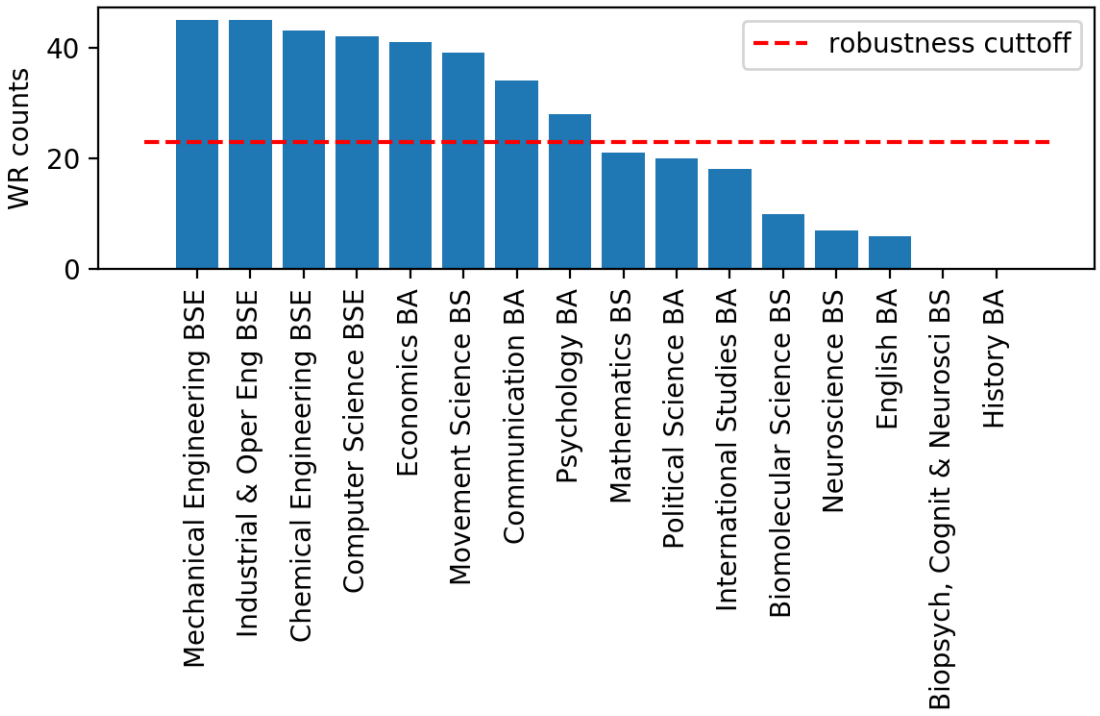


Figure 4.12: Major weak recoverability with robustness cutoff

4.5 Conclusion

In this chapter we set out to answer the following question: Do legacy categories make sense? We analyzed three levels of legacy categories. The first level was looking at if a student received a BA or BS. The next level was seeing if the student studied in the Humanities, Social Sciences, Biological Sciences, or Natural Sciences. The last level was to use a students major as a label. We compared how these labels performed to optimal algorithmic cluster and random. The first criteria for evaluation was to compare the distribution of students in the legacy label and compare it to the algorithm. We didn't compare the distribution of students to random because, Random(k) was forced to have the same distribution. At this stage we noticed that all three levels of legacy labels had a different distribution of students compared to the algorithm. However, this assessment isn't too informative in answering the question of do they make sense, this was more of a baseline comparison. The first approach to try and answer the question of do legacy labels makes sense was to see how well a categorization predicts course agreement. This criteria measured the number of courses in common between students of the same category. It then measured the number of courses in common between students from different categories. It was in this section that we got the first clue that using the legacy label of BA/BS might not be a good idea. Looking at student agreement, BA/BS was much closer to the performance of the Random(k) than to the Algorithm(k). It was also in this section that we get the idea that major labels may be a legacy categorization system that makes sense. In the next step we compare how students are assigned labels in the legacy system to the Algorithm(k) using Rand Index and Normalized Mutual Information. We again see BA/BS and HSBN perform less than stellar when compared to Algorithm(k). It is at this point a reader may draw the conclusion that BA/BS and HSBN are not good labels, but majors are. However, it is important to note that these are aggregate measures. Due to this we use measures of recoverability and robustness to evaluate

each individual major. This analysis was not applied to BA/BS and HSBN because of the poor performance in the earlier sections. It is in this section we notice that some labels do make sense. The engineering majors, movement science, and economics are all recoverable and robust by the strongest criteria. However, two majors,BCN and history,never make the robustness or recoverability criteria.

CHAPTER V

Conclusion

In this thesis we described the LARC dataset provided by the university and do some exploratory data analysis. We then investigate how to define connections within this dataset. We also study how students are grouped together, both by legacy labels and by clustering algorithms. Below all of this summarized in more detail by chapter. There is also a brief discussion on future direction.

5.1 Summary

Chapter II introduced the LARC dataset which was an essential component to this thesis. This dataset contains information on more than 200,000 students who enrolled in approximately 15,000 courses. We focused on a subset of this data, specifically, a cohort of 6,738 students who enrolled in the university in the fall of 2011 and graduated by winter 2016. These students enrolled in a total 6,152 courses throughout this period. The main purpose of this chapter was to demonstrate how this enrollment data can be represented by a bipartite network and to show how this representation makes it easy to draw various conclusions about students and courses at the university. To reiterate some points from the chapters conclusion, we were able to show and quantify that the intuited belief that high enrollment courses uniquely connect students from different academic backgrounds. However, we also discovered that this

is not a rule. We saw that some low enrollment courses also served as unique connectors of students, with similar quantitative values as the larger courses. Looking at the local clustering coefficient we found how certain students and courses were extremely effective bridges for typically disconnected parts of campus. This chapter mainly focused on how individual courses and students are embedded in the network. However, the nature of this embeddedness depends significantly on how an edge is defined in the bipartite network.

Chapter III looks at various definitions for calculating the connection between students (the existence and weight of an edge between two nodes). We setup three ways of calculating a students connectedness. The first was *unique connections*, in this formulation students got a value of 1 for every student they took a course with. The next was *weighted connections* takes into consideration the number of courses you take with students. The third measure was *intensity connections*. This is similar to weighted, but instead of attributing the same value to every course, the courses value is a function of the size with smaller courses getting a larger weight than smaller ones. A toy model was introduced to help demonstrate the construction of the various connection definitions.

The analysis in this chapter is in two parts. The first examines how the various definitions (unique, weighted, and intensity) are distributed among the students. Looking at the entire dataset we see that unique connections and weighted connections had a correlation of 0.80 among the students. We also see that intensity connections had a near zero correlation of 0.026 with unique connections and a correlation of .15 with weighted connections. However, upon further investigation we uncovered that these distributions and correlations varies significantly among the majors. Looking specifically at the correlation between unique and intensity connections we notice an opposite measure between Chemical Engineering (-.663) and History (.673).

The next part of the analysis focused on how unique, weighted, and intensity

connections affect network centrality measures. The centrality measures explored were degree centrality, eigenvector centrality, and triangle centrality. As expected from the previous work, this analysis showed that a students change in these measures were dependent on their majors. For example, we noticed that students that had a higher average number of courses taken throughout the period of investigation received higher intensity connected eigenvector centrality scores. Examining these students showed that they all belonged to the arts (e.g., Dance BFA).

Chapter IV explores the clusters that exist in the affinity matrix provided by the student network and compares them to how students are classically grouped together. We refer to the way students are classically grouped as legacy labels. The legacy labels that we investigate are Bachelors of Science or Bachelors of Arts (BA/BS), the grouping of humanities, social science, biological science, or natural science (H,S,B,N), and finally majors. The study of majors was restricted to only 16 with similar number of student enrollment. The 16 majors are listed in Table 4.1.

The first analysis completed looked at how students were distributed among the legacy labels and compared them to how they are distributed to clusters with the same number of possibilities (e.g., BA/BS was compared to partitioning the students into two groups using a clustering algorithm).

In the results section we set out to answer the following questions. The first question answered was: How well does a legacy labeling predict course agreement? Course agreement is the number of courses in common between two students. The finding here was that the clustering algorithm and majors proved the best predictors. Next we looked at: How do legacy labels compare to algorithmic label using the Rand Index and Normalized Mutual Information? Here we see that agreement was only found between the majors and algorithms. For the last two analysis ran in this chapter we focused on majors, because it was the only legacy label that was performing well. The next question we answer is: How robust are majors to the number of clusters?

We first introduce the idea of major coherence, which measures how well do students belonging to certain majors stay together when clustered by an algorithm. We also introduced two conditional probabilities, $P(M_{\text{Major}}|C^k_{\text{Major}})$ and $P(C^k_{\text{Major}}|M_{\text{Major}})$. In words $P(M_{\text{Major}}|C^k_{\text{Major}})$ ask given the cluster that contains the largest fraction of a specific major, what is the probability that major is selected from the elements of that cluster? While $P(C^k_{\text{Major}}|M_{\text{Major}})$ is asking given that an element belongs to a specific major, what is the probability it is in the cluster that contains the largest number of those majors.

Until now, we only focused on majors as a collective. It was at this point in the analysis where we look at individual majors. Answering this question of robustness we see the performance varies wildly. Some majors, for example, Chemical Engineering and Movement Science show high coherence and $P(M_{\text{Major}}|C^k_{\text{Major}})$ with k having little effect on the value of these measures. Other majors, for example, English and Neuroscience, values vary wildly based on k.

In the next part of analysis we introduce the idea of recoverability with a weak and strong constraint. We define weak recoverability to be a situation where both $P(M_{\text{Major}}|C^k_{\text{Major}})$ and $P(C^k_{\text{Major}}|M_{\text{Major}})$ have values greater than 0.6 for a specific value of k. We define strong recoverability to be where the same probabilities have values greater than 0.8 for values of k. The goal of weak and strong recoverability is to explore the following idea: in a hypothetical world where students were clustered by an algorithm, what is the chance that I would end up defining a cluster with the same name given by the legacy labeling? We find that the engineering majors exhibit strong recoverability for almost any value of k, while some majors like History, are never recoverable (weak or strong).

Next we come up with a definition of robustness that is a function of the recoverability of a cluster algorithm. Here we see that two majors, Biopsych, Cognit & Neurosci BS and History BA are not robust. This means that students in this ma-

jor cannot be seprated from students in others algorithmically based on similarity in course co-enrollment.

5.2 Tying Category Recoverability to Intensity

In Chapter III a differential response to how connections were defined in the network was observed among the majors. In Chapter IV some majors showed high coherence, strong recoverability, and robustness, while some showed near zero for these measures. In this section of the thesis we'll explore some connections between the majors behavior in these two chapters.

First we examine the how the mean value of unique connections for students within a major and see if that major is strongly recoverable. This is shown in Figure 5.1. Here and in Figures 5.2 and 5.3 a bar is colored green if the corresponding major is strongly recoverable. So looking at how unique connections are distributed about the majors, we don't any pattern emerge.

Some majors with high values for unique connections are strongly recoverable and some are not. The opposite is true too, some majors with low mean value for unique connections are strongly recoverable while others are not. Moving to Figure 5.2 we start to see the emergence of a pattern. That is, mostly majors with high values of weighted connections are strongly recoverable. The strongest relationship emergence when we look at Figure 3.3.2.3. There seems to be a strong relationship between intensity connections and strong recoverability.

Indeed, this is what we find when we look at the correlation between the three formulations of connections and strong recoverability. The results are shown in Table 5.1. Strong recoverability and intensity have a correlation of 0.772. This is really quite amazing. Intensity, a way of defining connections between students, is highly correlated with the robustness of a major. These two separate studies that look at the relationship between students in two different lenses give to similar conclusions

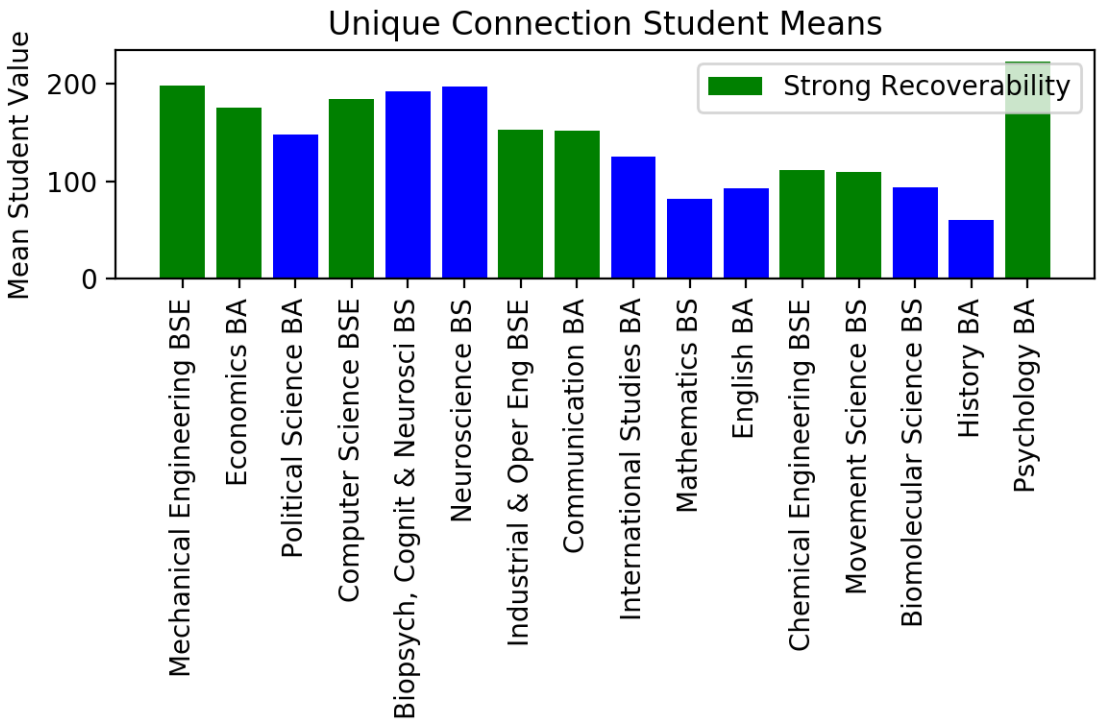


Figure 5.1: Average number of unique connections by major with the strongly recoverable majors highlighted

	Correlation
SR-Unique	0.249
SR-Weighted	0.691
SR-Intensity	0.772

Table 5.1: Strong Recoverability - Unique/Weighted/Intensity Correlations

about certain majors. Using both frameworks, we can re-examine how we think about majors at the university.

5.3 Future work

One area of particular interest is to extend this dataset to include location data based on students connecting to the internet throughout campus. This is a completely different way to define student connections that will give more information on the

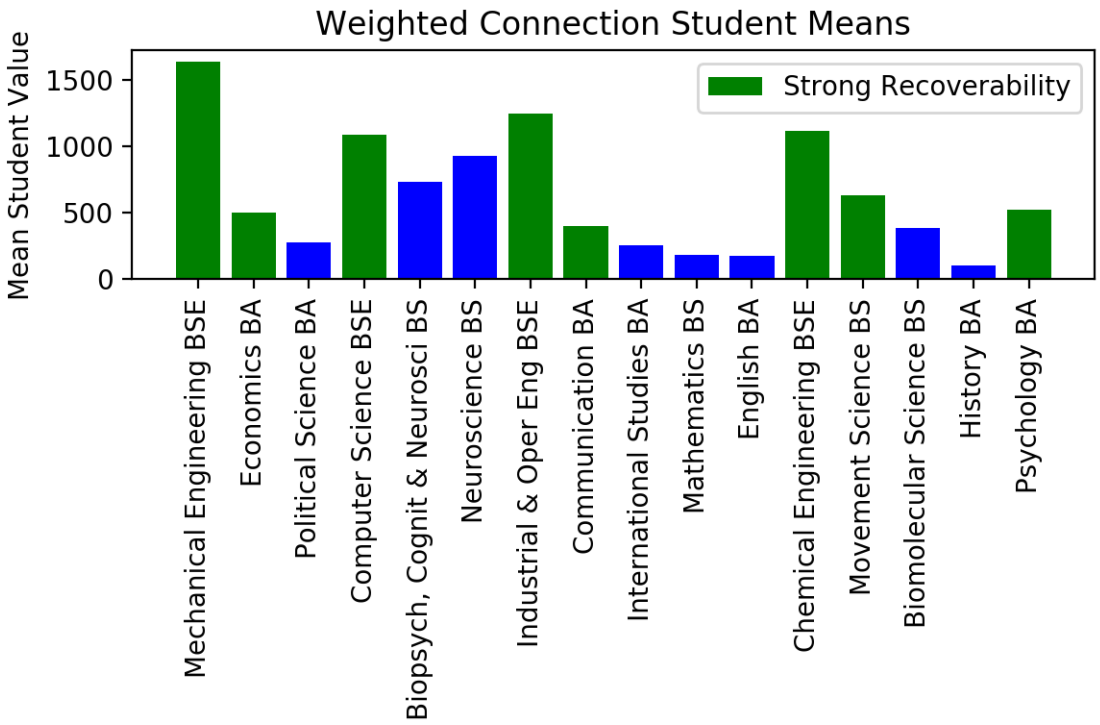


Figure 5.2: Average number of weighted connections by major with the strongly recoverable majors highlighted

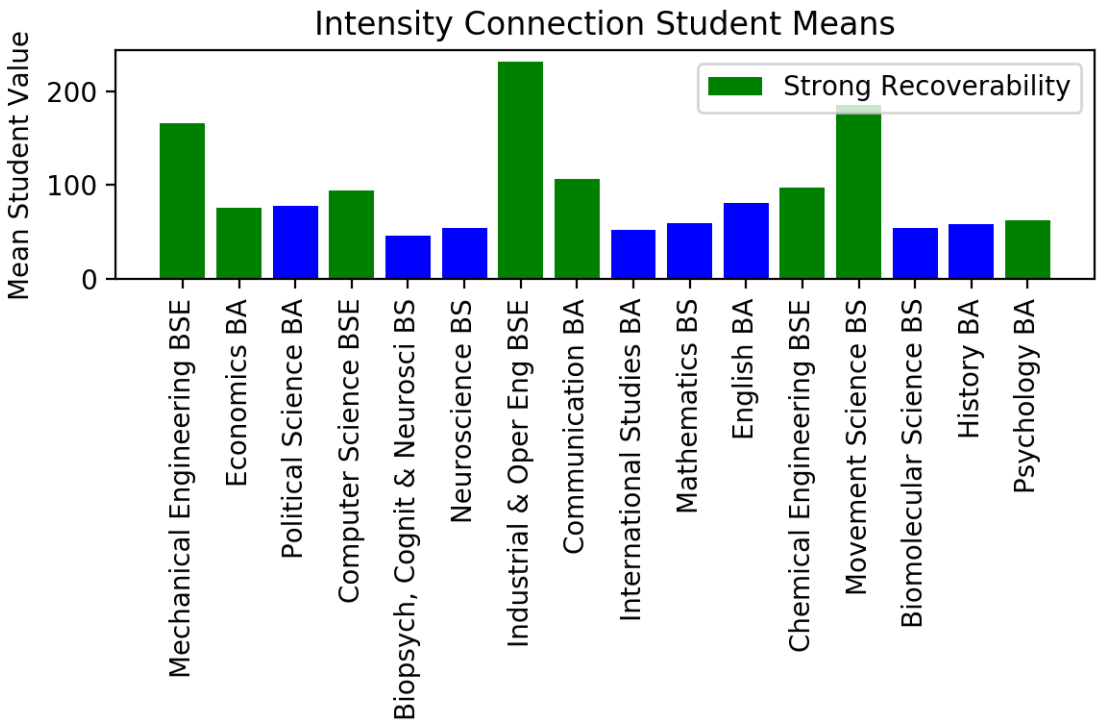


Figure 5.3: Average number of intensity connections by major with the strongly recoverable majors highlighted

student experience. In the dataset explored in this thesis co-location was known by students being enrolled in the same course at the same time, this will allow us to understand co-location outside of the classroom.

Future work could also apply our measures of coherence, recoverability, and robustness introduced in Chapter IV and apply it to other datasets. Because labeling or grouping items together is an integral part of what humans do and the inertia that some labelings have, it is of upmost importance we discover how useful these labels are.

APPENDICES

APPENDIX A

Chapter 3

A.1 Student Distributions for unique connections, weighted connections, and intensity connections by Major

```
array([[1.        , 0.8140199 , 0.21092172],
       [0.8140199 , 1.        , 0.20658923],
       [0.21092172, 0.20658923, 1.        ]])
```

Figure A.1: BMS unique, weighted, and intensity distributions

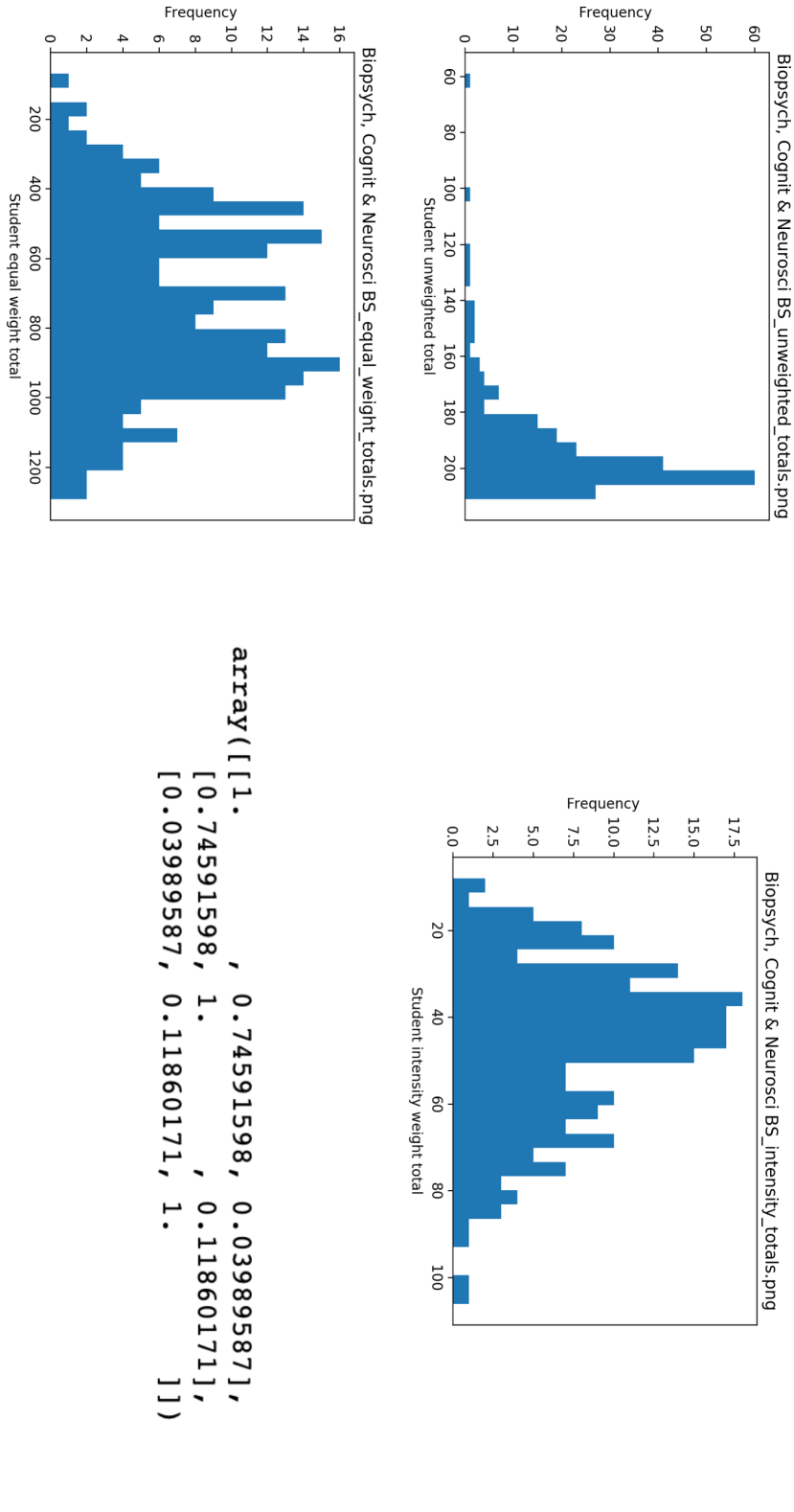
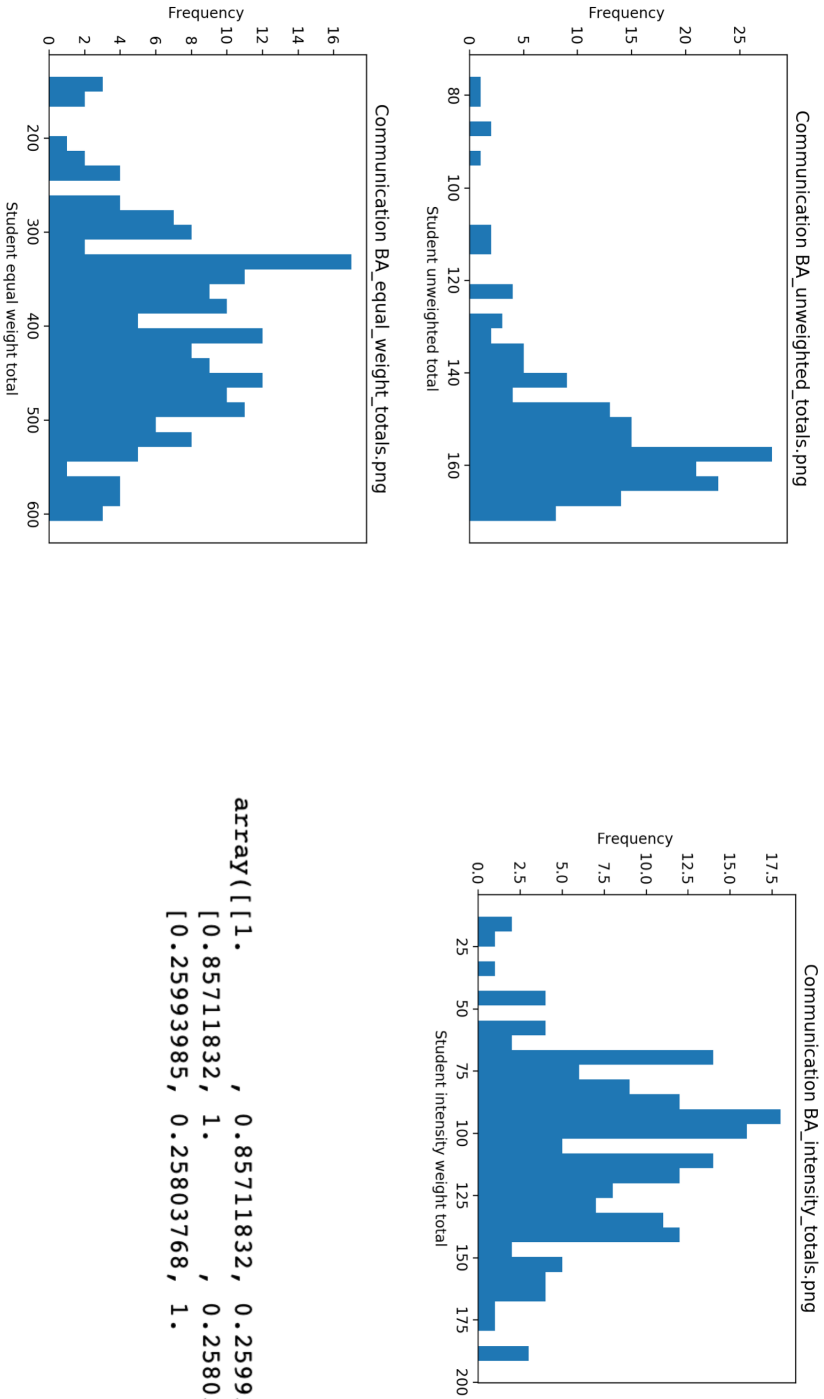


Figure A.2: BCN unique, weighted, and intensity distributions

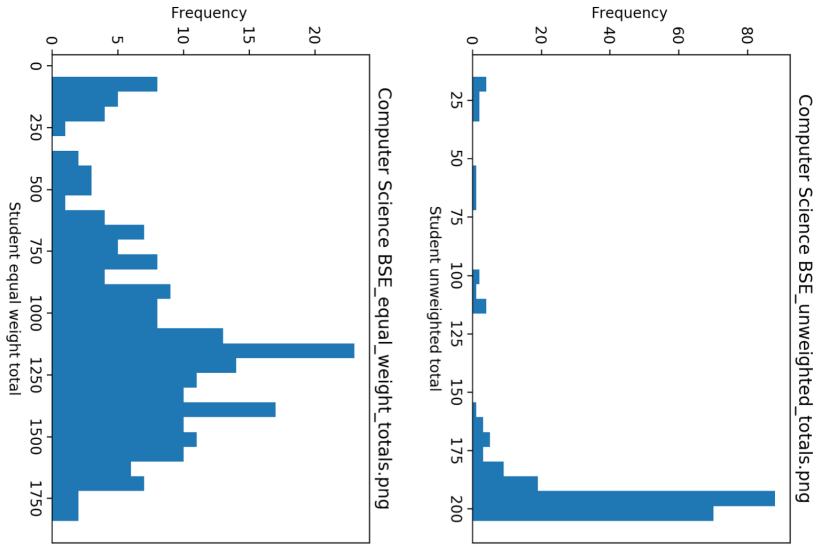
```
array([[ 1.        ,  0.72872344, -0.66306711],
       [ 0.72872344,  1.        , -0.28235007],
       [-0.66306711, -0.28235007,  1.        ]])
```

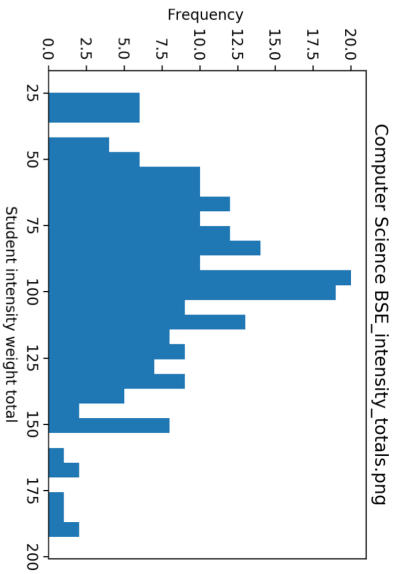
Figure A.3: CE unique, weighted, and intensity distributions




```
array([[1.        , 0.85711832, 0.25993985],
       [0.85711832, 1.        , 0.25803768],
       [0.25993985, 0.25803768, 1.        ]])
```

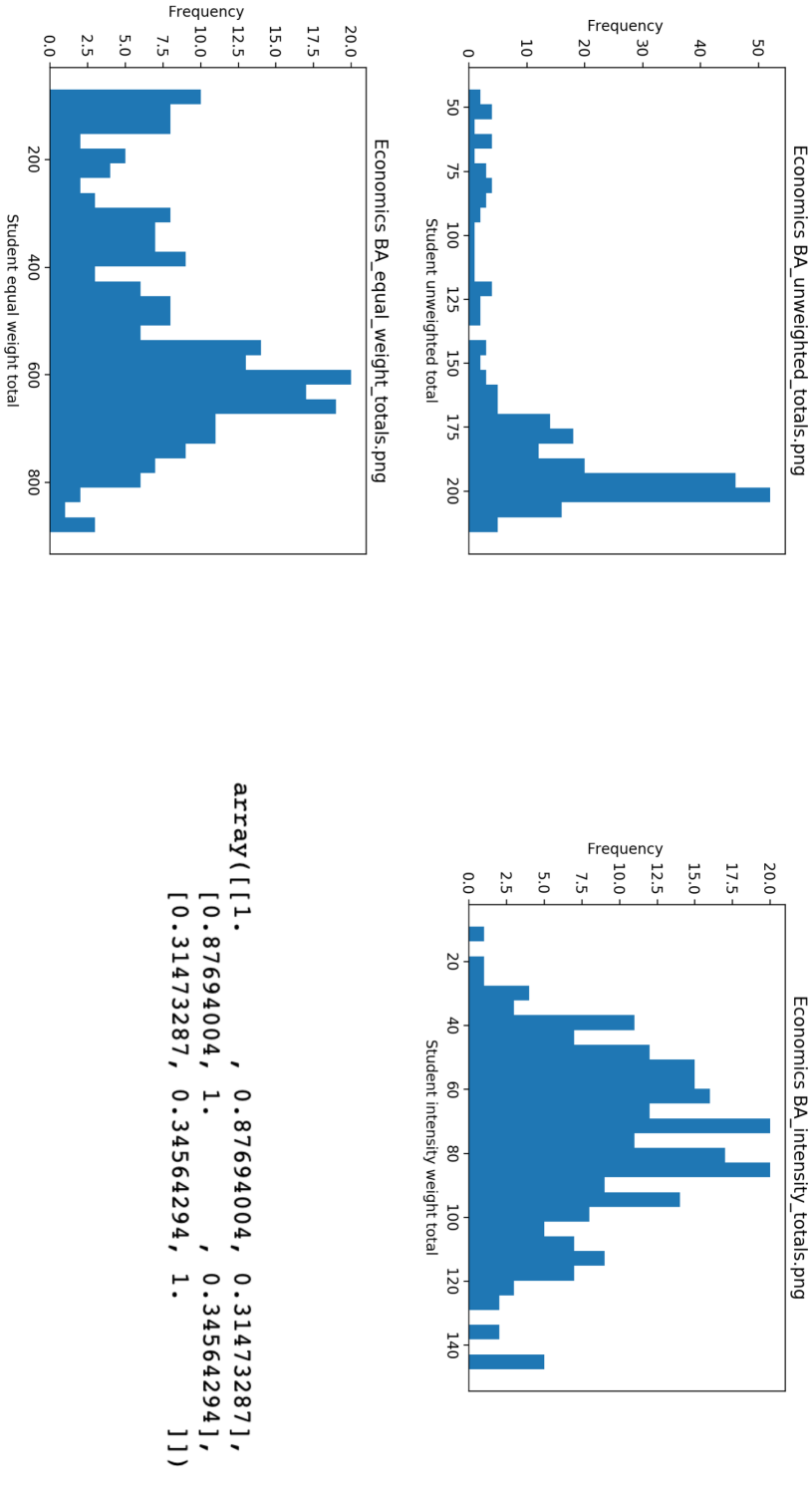
Figure A.4: COM unique, weighted, and intensity distributions





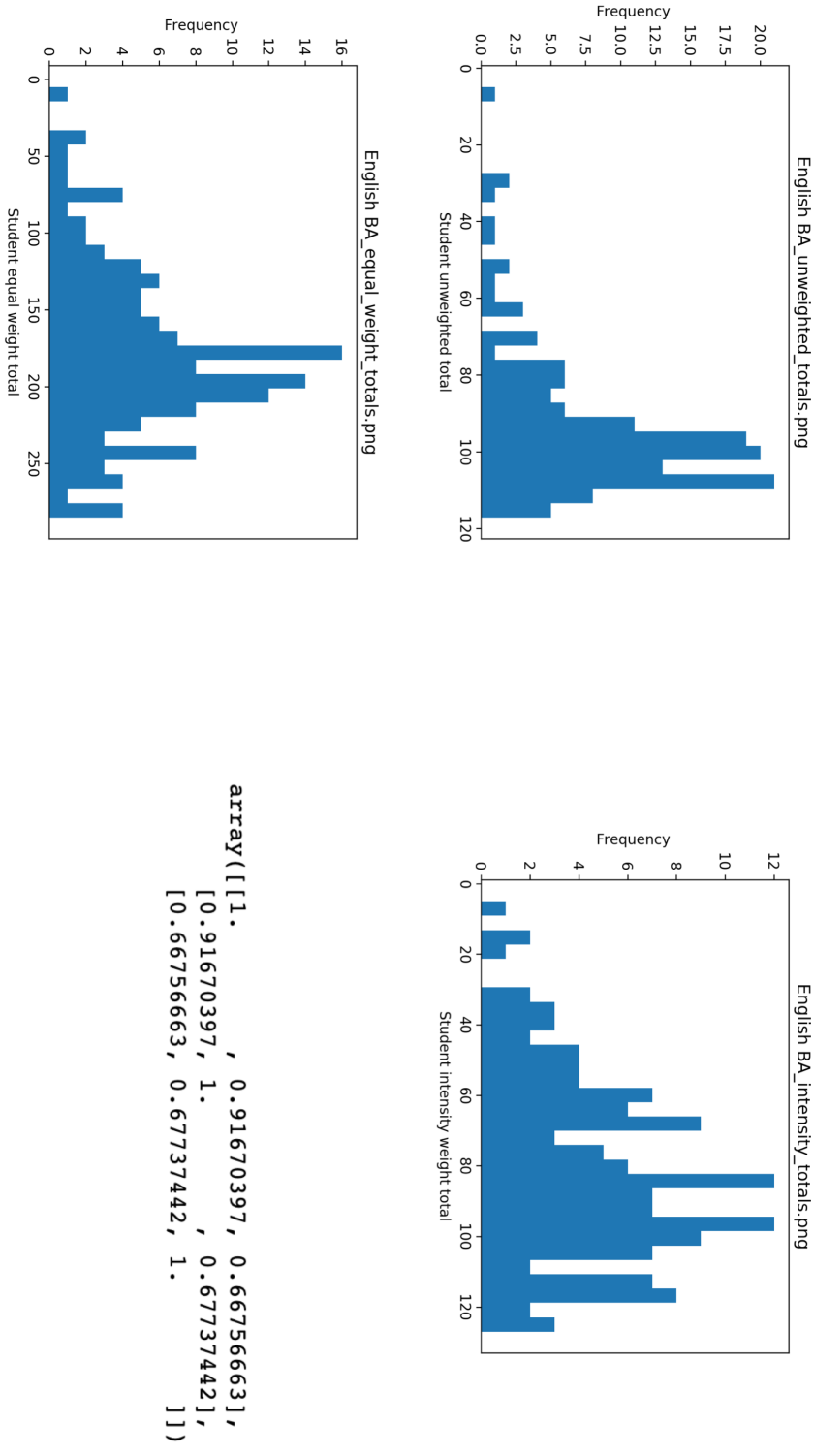

```
array([[ 1.        ,  0.75111875, -0.13747481],
       [ 0.75111875,  1.        ,  0.05994673],
       [-0.13747481,  0.05994673,  1.        ]])
```

Figure A.5: CS unique, weighted, and intensity distributions



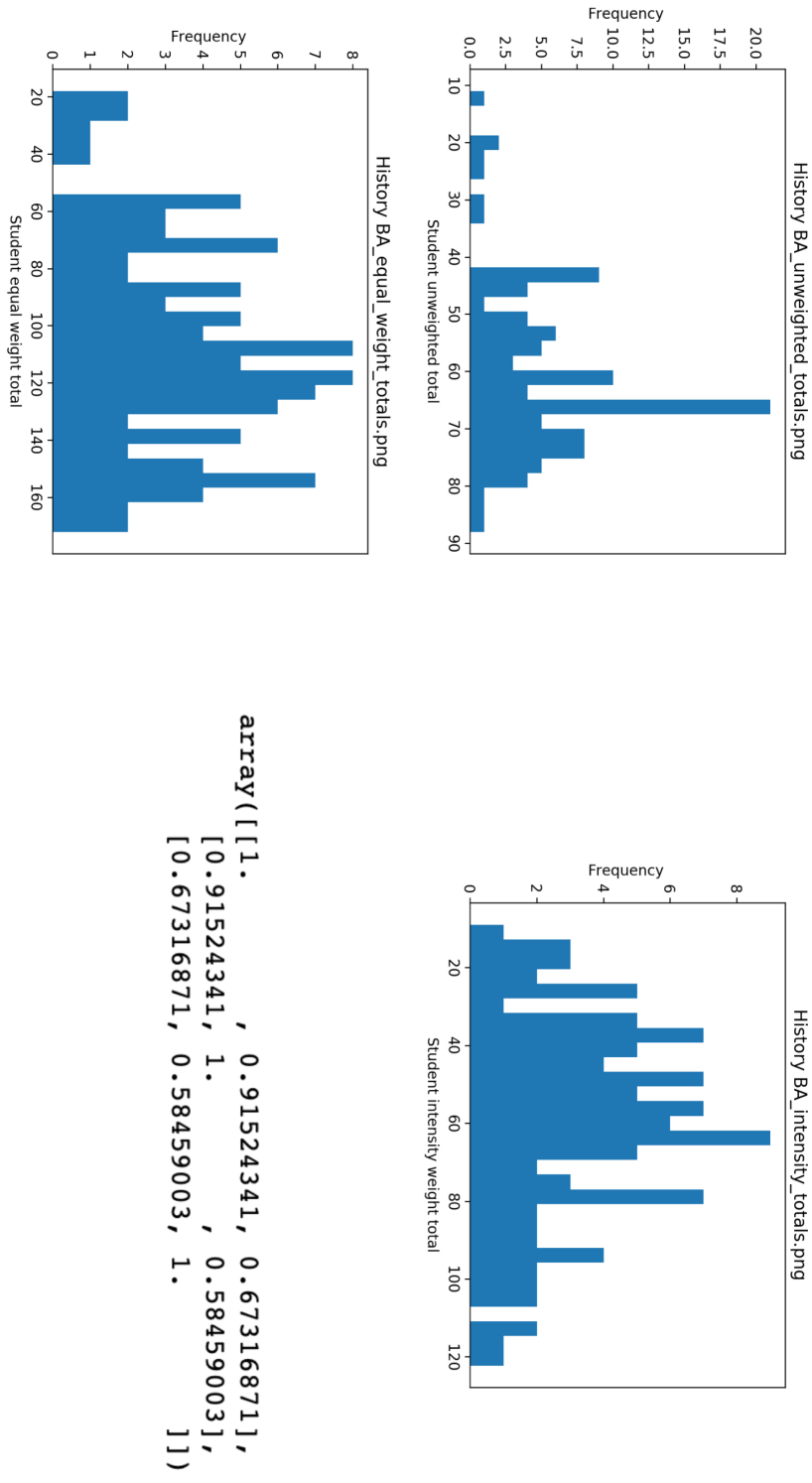

```
array([[1.        , 0.87694004, 0.31473287],
       [0.87694004, 1.        , 0.34564294],
       [0.31473287, 0.34564294, 1.        ]])
```

Figure A.6: ECN unique, weighted, and intensity distributions



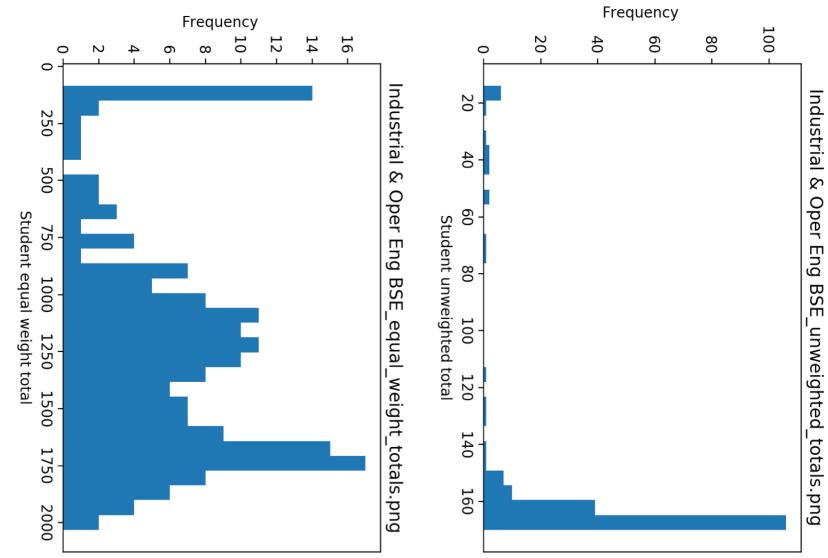

```
array([[1.        , 0.91670397, 0.66756663],
       [0.91670397, 1.        , 0.67737442],
       [0.66756663, 0.67737442, 1.        ]])
```

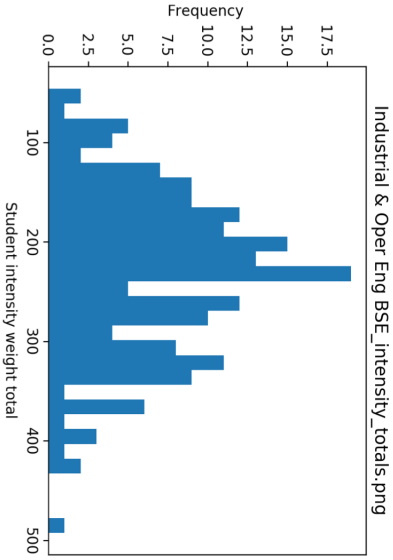
Figure A.7: ENG unique, weighted, and intensity distributions




```
array([[1.        , 0.91524341, 0.67316871],
       [0.91524341, 1.        , 0.58459003],
       [0.67316871, 0.58459003, 1.        ]])
```

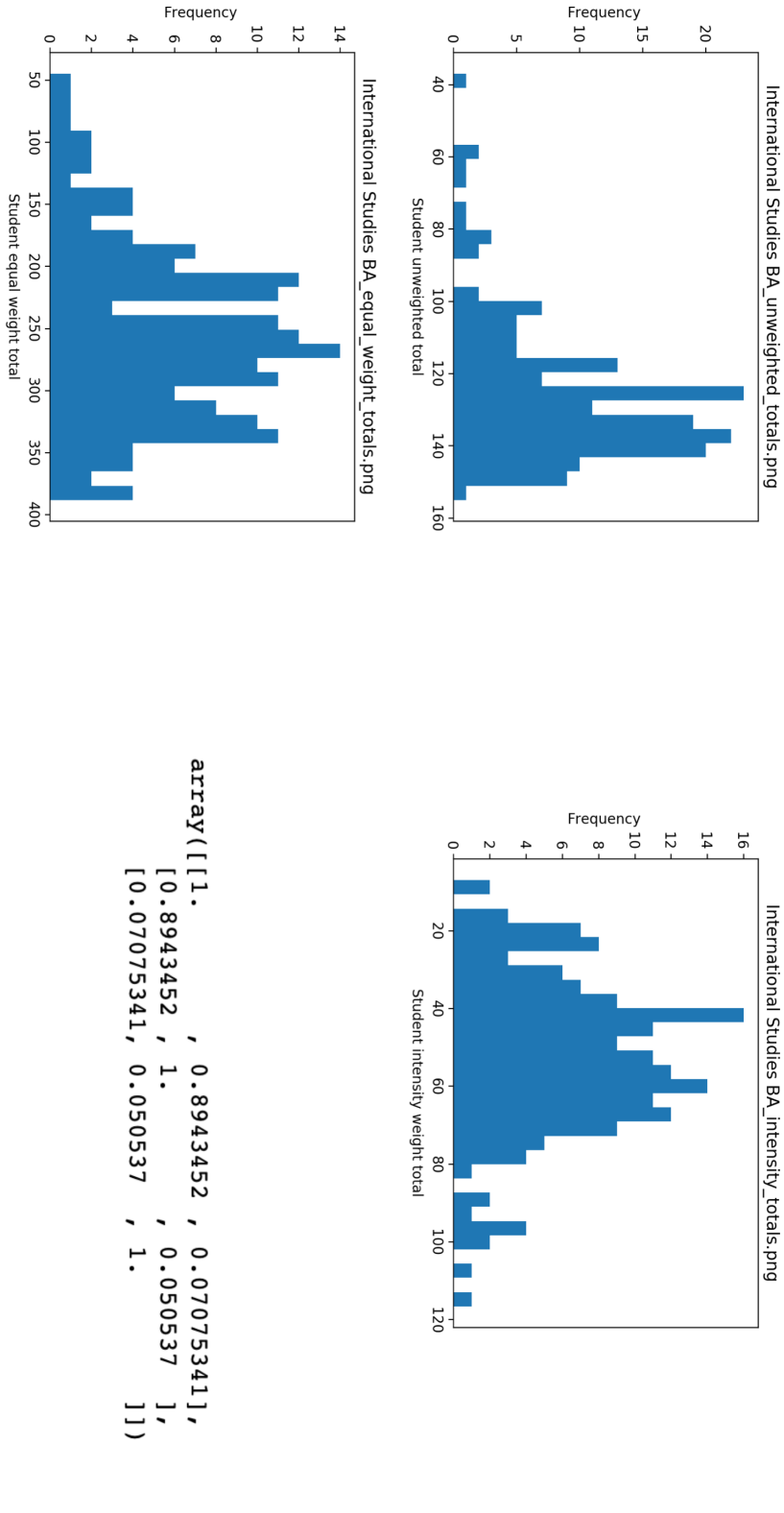
Figure A.8: HIS unique, weighted, and intensity distributions






```
array([[1.        , 0.76172713, 0.2853698 ],
       [0.76172713, 1.        , 0.44515339],
       [0.2853698 , 0.44515339, 1.        ]])
```

Figure A.9: IOE unique, weighted, and intensity distributions




```
array([[1.        , 0.8943452 , 0.07075341],
       [0.8943452 , 1.        , 0.050537  ],
       [0.07075341, 0.050537  , 1.        ]])
```

Figure A.10: IS unique, weighted, and intensity distributions

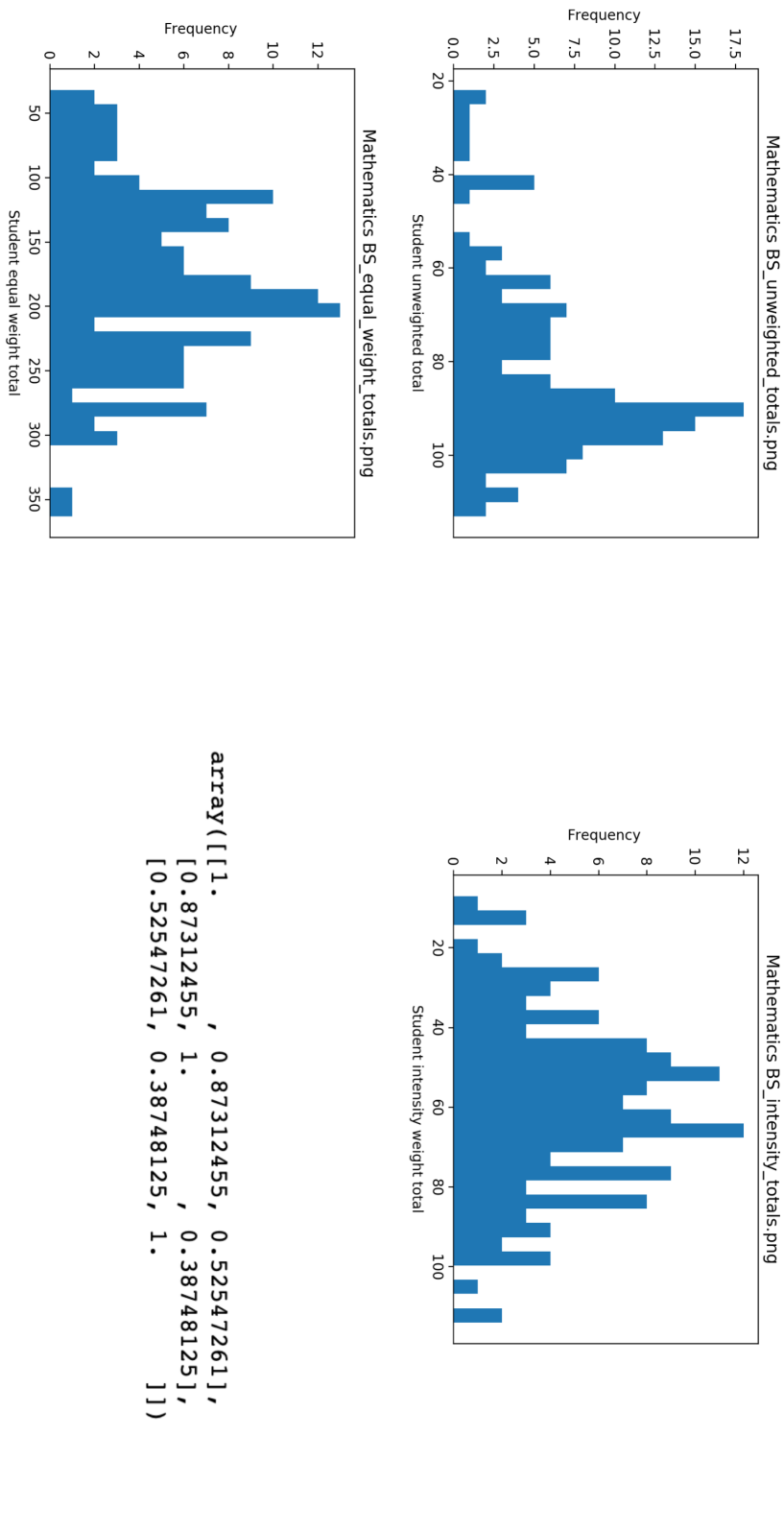
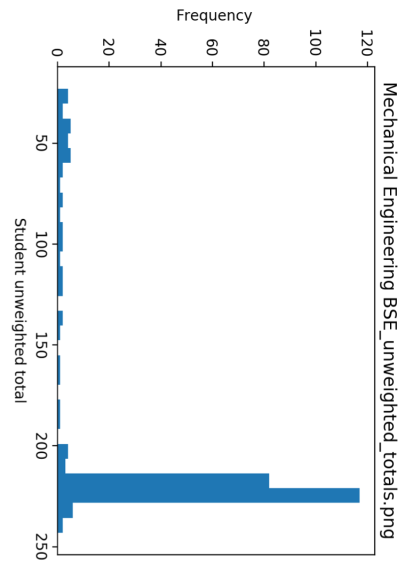
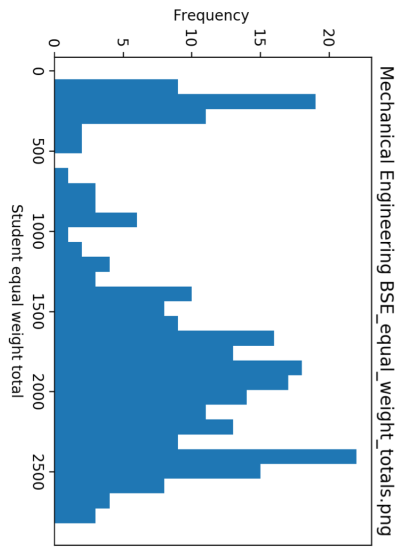
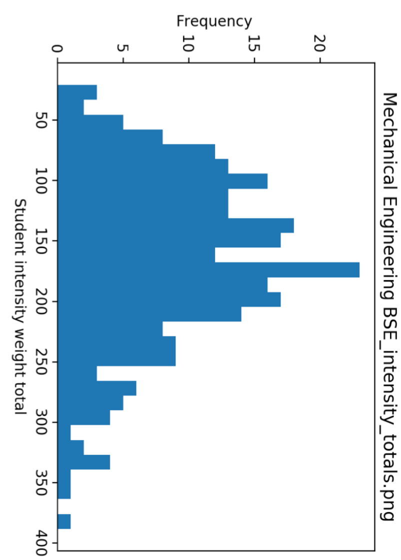


Figure A.11: MTH unique, weighted, and intensity distributions








```
array([[ 1.        ,  0.78981533, -0.20227767],
       [ 0.78981533,  1.        ,  0.06931772],
       [-0.20227767,  0.06931772,  1.        ]])
```

Figure A.12: ME unique, weighted, and intensity distributions

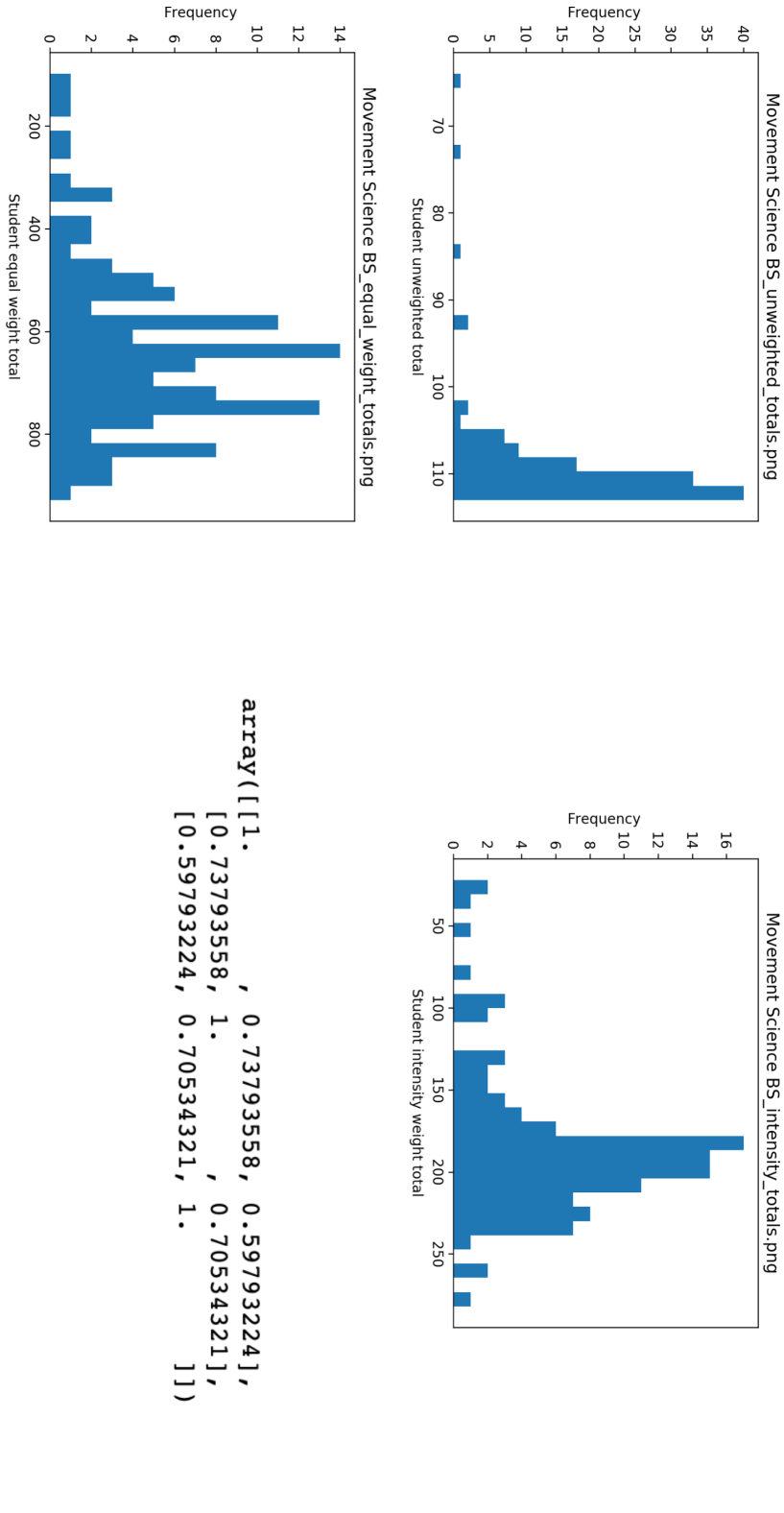


Figure A.13: MS unique, weighted, and intensity distributions

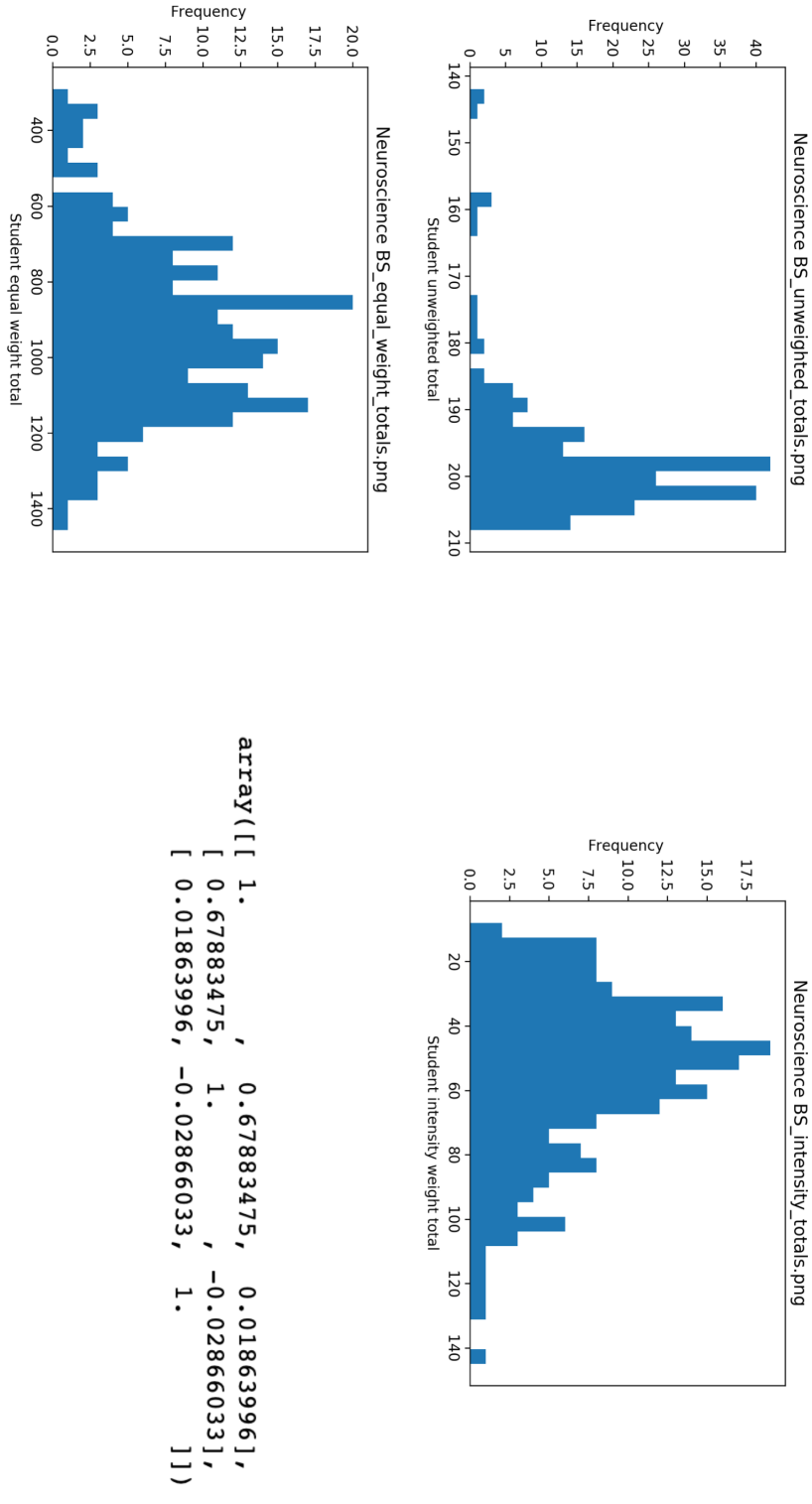
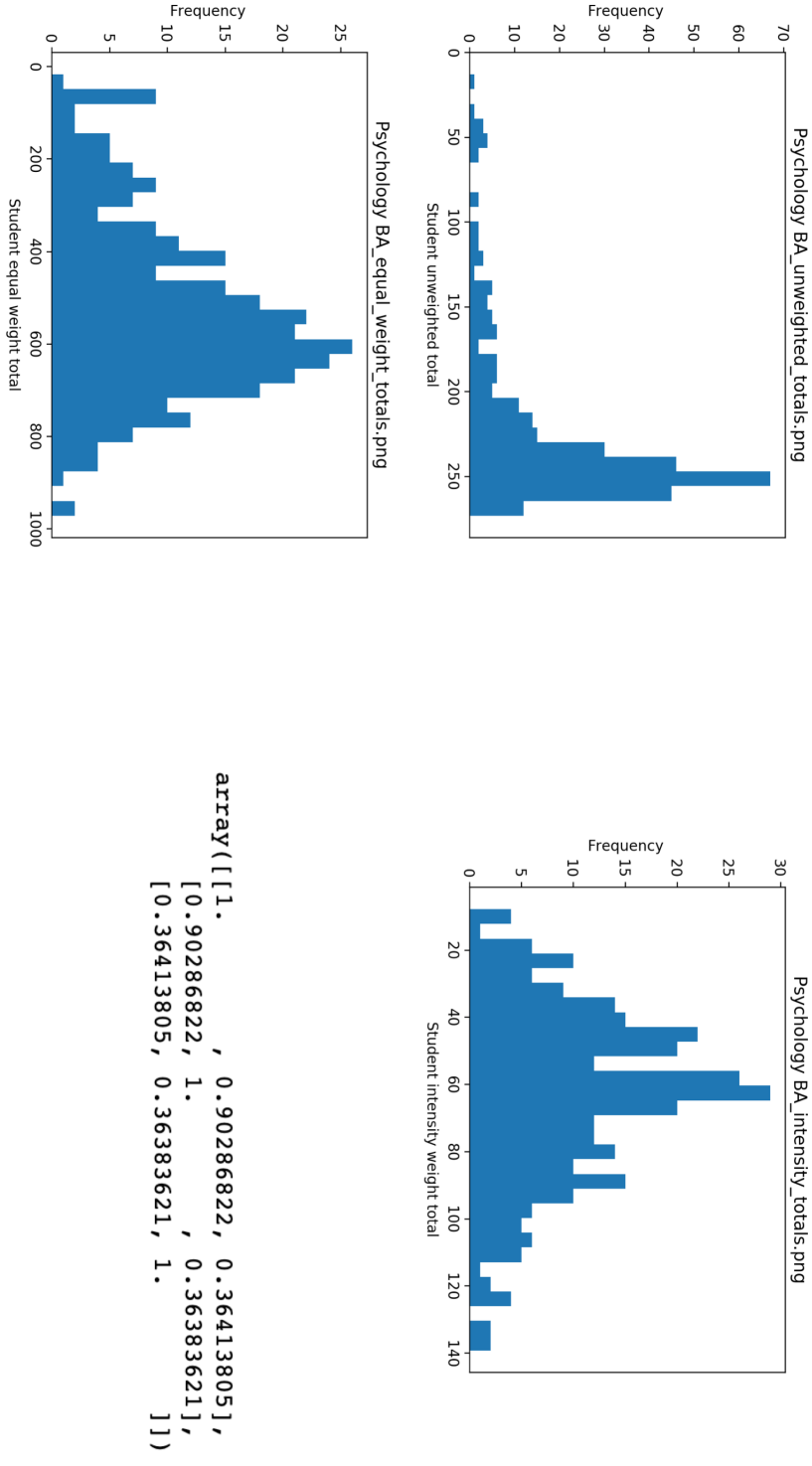


Figure A.14: NEU unique, weighted, and intensity distributions

Figure A.15: PS unique, weighted, and intensity distributions




```
array([[1.        , 0.90286822, 0.36413805],
       [0.90286822, 1.        , 0.36383621],
       [0.36413805, 0.36383621, 1.        ]])
```

Figure A.16: PSY unique, weighted, and intensity distributions

APPENDIX B

Chapter 4

B.1 Contingency Tables

	H	S	B	N	sum
cluster 1	201	1013	3	97	1314
cluster 2	4	2	0	444	450
cluster 3	40	93	643	121	897
cluster 4	0	0	0	258	258
sum	245	1108	646	920	

Table B.1: Contingency Table for HSBN and Clusters

	PSY	ME	ECN	PS	CS	BCN	NEU	IOE	COM	IS	MTH	ENG	CE	MS	BMS	HIS	sum
cluster 1	0	104	2	0	0	0	0	0	0	0	0	0	0	0	0	0	106
cluster 2	0	0	0	0	0	0	0	0	2	0	0	26	0	0	0	0	28
cluster 3	0	0	0	0	214	1	0	0	1	0	7	0	0	0	0	0	223
cluster 4	0	0	0	101	0	0	0	0	0	1	0	0	0	0	0	52	154
cluster 5	19	0	0	18	1	194	209	0	5	0	23	15	0	0	108	21	632
cluster 6	271	0	1	5	0	20	0	0	13	2	1	3	0	0	0	5	321
cluster 7	4	0	0	0	0	0	0	0	148	3	0	0	0	0	0	1	156
cluster 8	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	17
cluster 9	0	256	0	0	1	0	0	1	0	0	0	0	0	0	0	0	258
cluster 10	0	0	0	12	0	0	0	0	1	122	0	0	0	0	0	2	137
cluster 11	0	0	0	0	0	0	0	0	0	0	0	0	0	114	0	0	114
cluster 12	0	0	0	0	0	0	0	164	0	0	0	0	0	0	0	0	164
cluster 13	0	0	27	0	0	0	0	0	0	0	2	0	0	0	0	1	30
cluster 14	0	0	0	0	0	0	0	1	0	0	0	0	125	0	0	0	126
cluster 15	2	0	0	3	0	0	0	0	4	0	0	92	0	0	0	1	102
cluster 16	4	0	207	83	0	0	0	0	4	24	3	2	0	0	0	24	351
sum	300	256	237	222	216	215	209	183	178	171	140	138	125	114	108	107	

Table B.2: Contingency for Majors and Clusters

B.2 Coherence$_{Maj}$ and $P(M_{\mathbf{Maj}}|C^k_{\mathbf{Maj}})$ Results

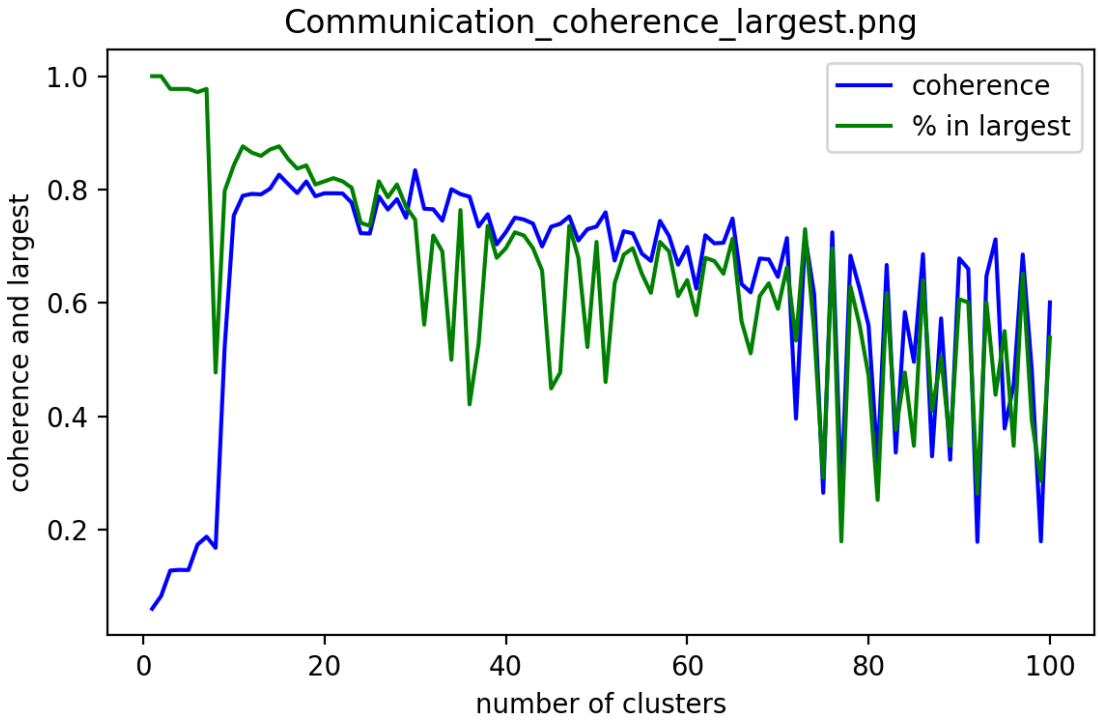


Figure B.1: Communication - Coherence$_{Maj}$ and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ change with k

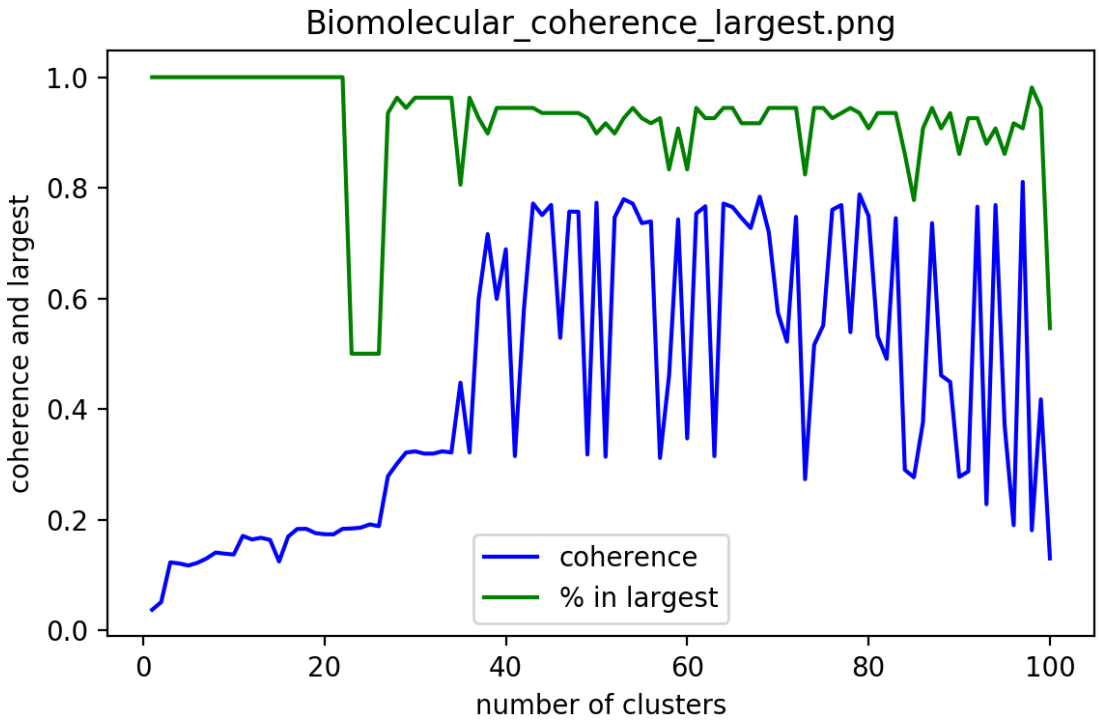


Figure B.2: Biomolecular Science - Coherence$_{Maj}$ and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ change with k

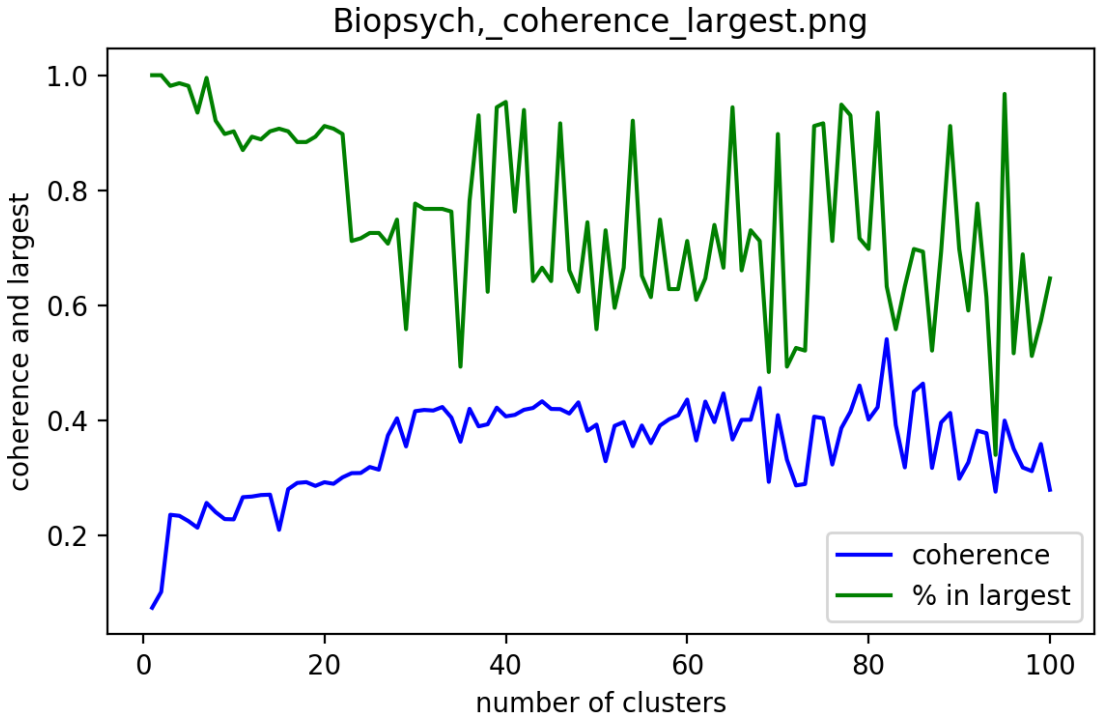


Figure B.3: Biopsych, Cognit & Neurosci - Coherence$_{Maj}$ and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ change with k

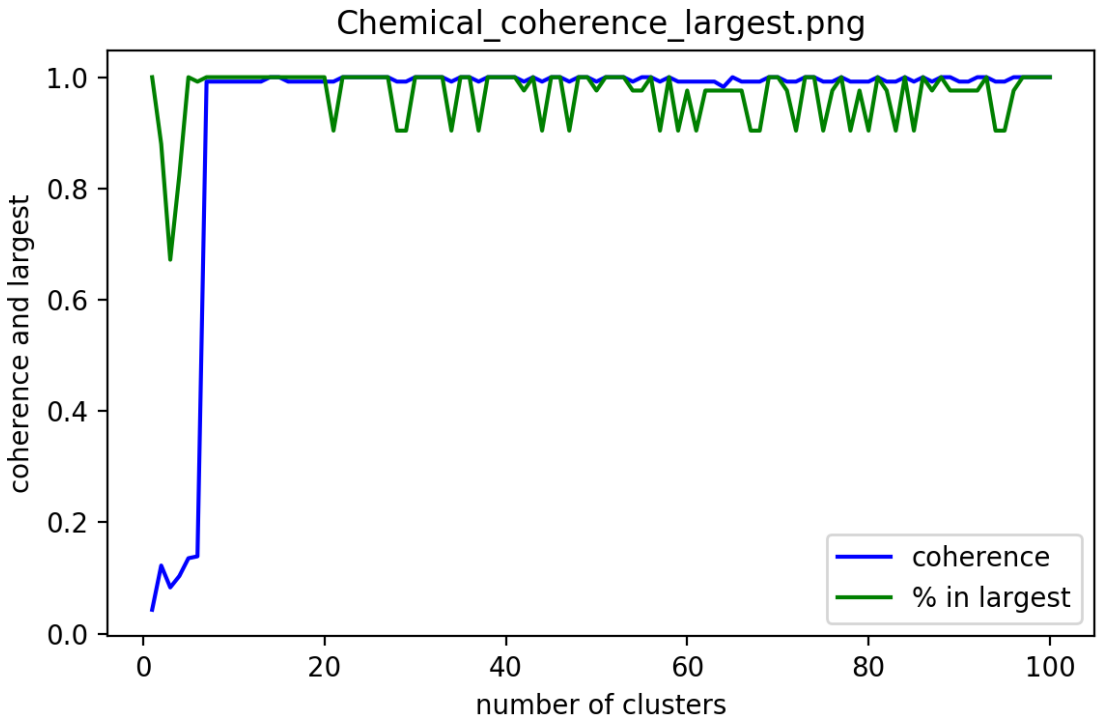


Figure B.4: Chemical Engineering - Coherence$_{Maj}$ and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ change with k

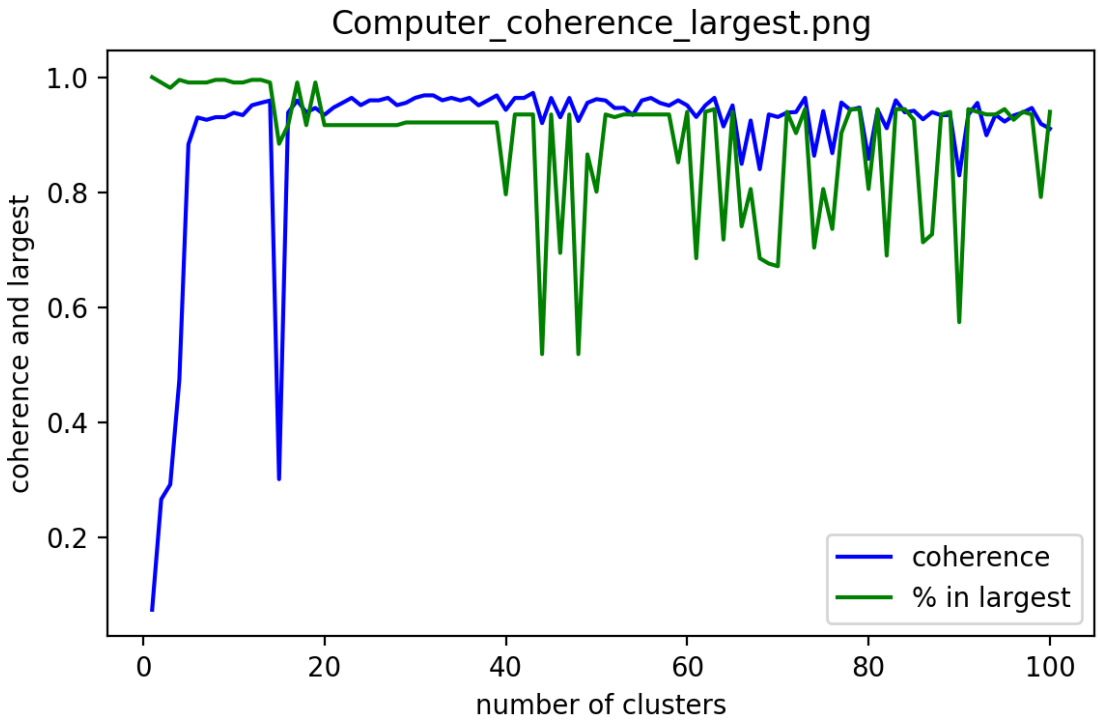


Figure B.5: Computer Science - Coherence$_{Maj}$ and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ change with k

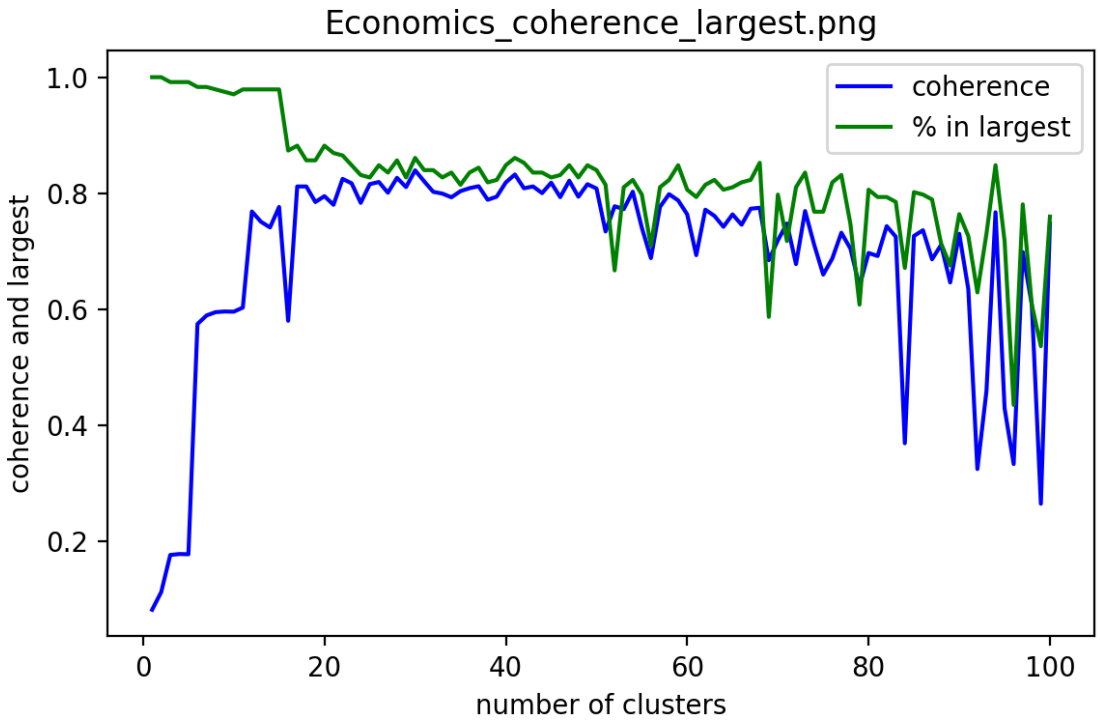


Figure B.6: Economics - Coherence$_{Maj}$ and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ change with k

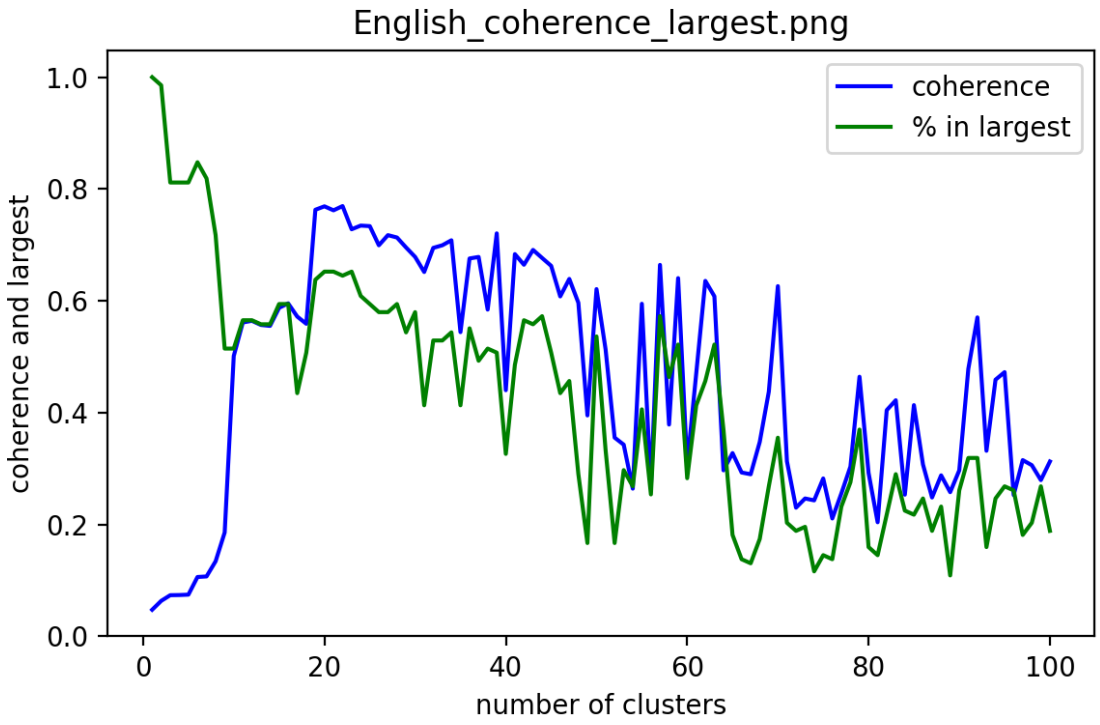


Figure B.7: English - Coherence$_{Maj}$ and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ change with k

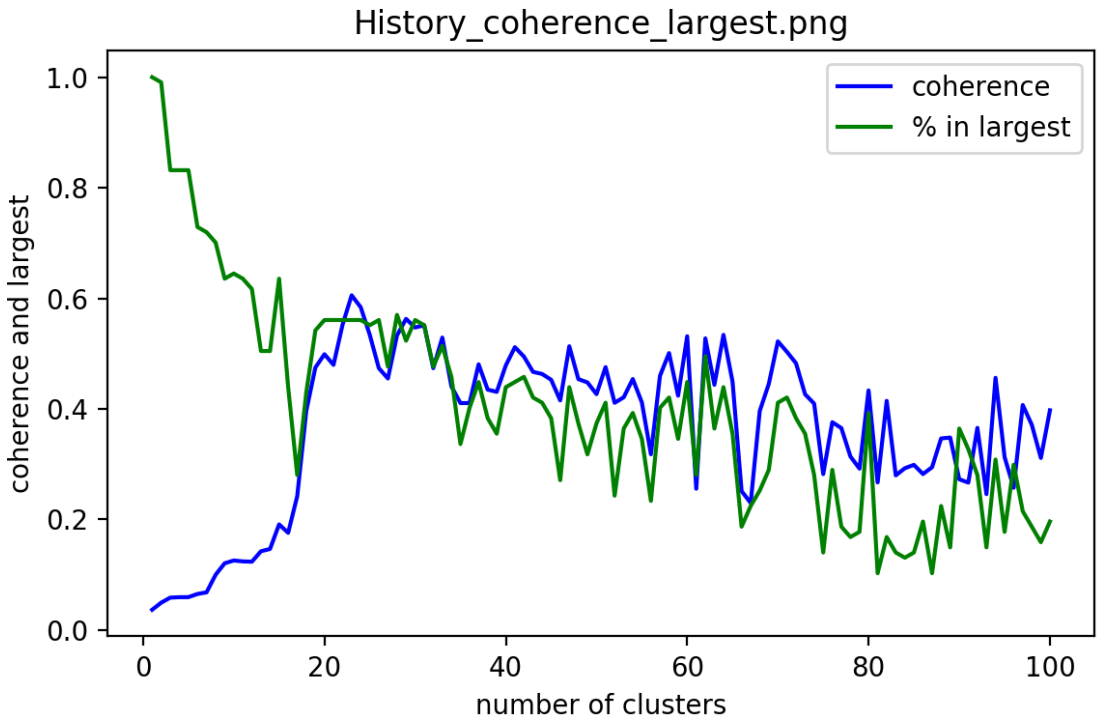


Figure B.8: History - Coherence$_{Maj}$ and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ change with k

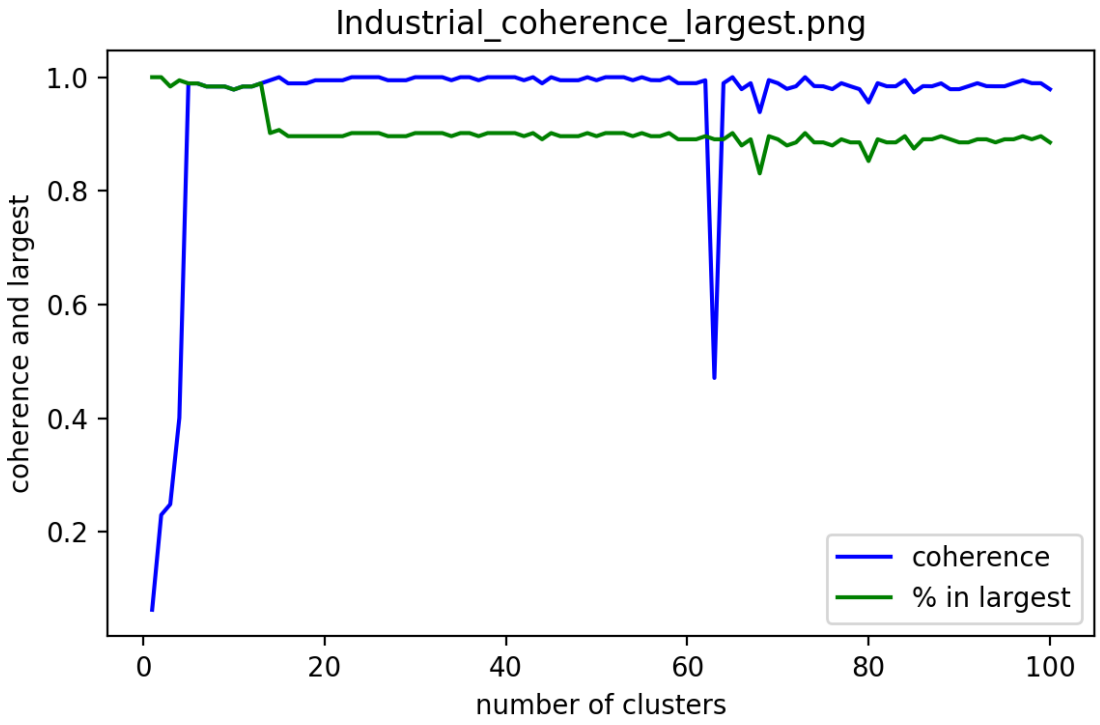


Figure B.9: Industrial & Oper Eng - Coherence$_{Maj}$ and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ change with k

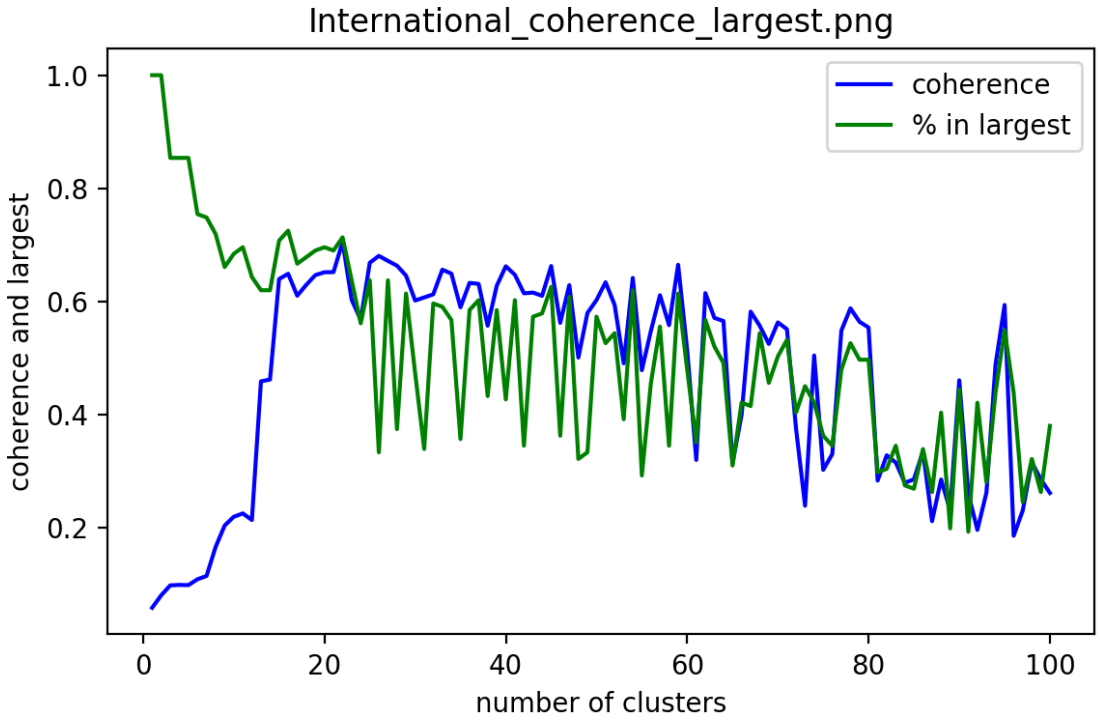


Figure B.10: International Studies - Coherence$_{Maj}$ and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ change with k

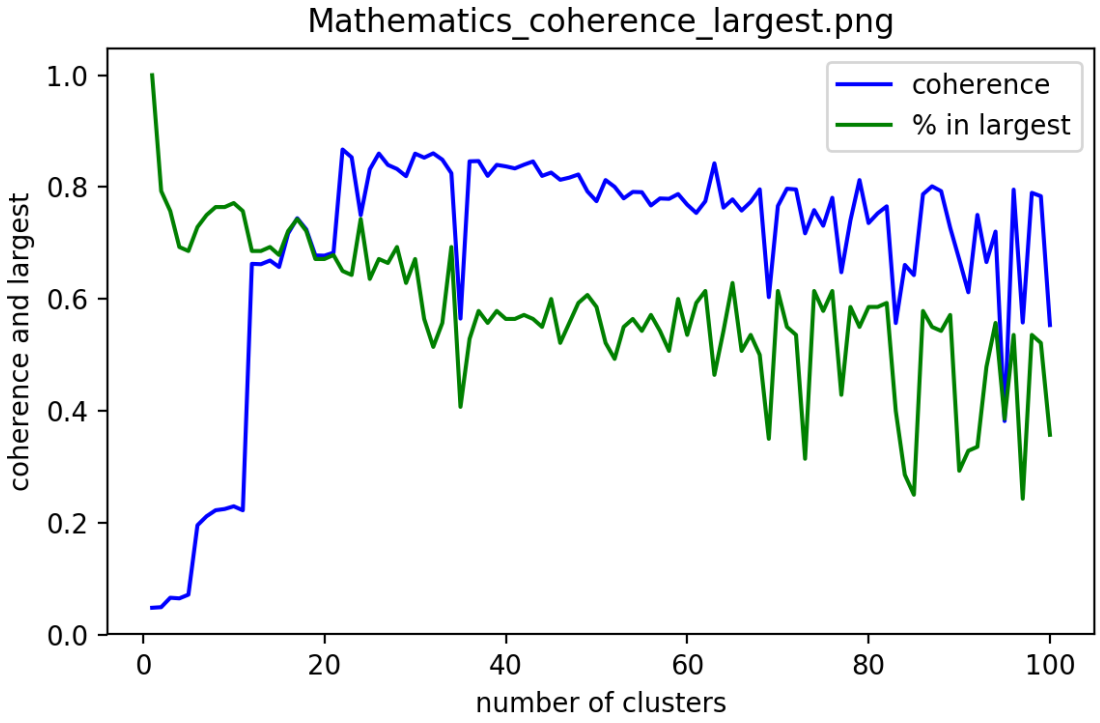


Figure B.11: Mathematics - Coherence$_{Maj}$ and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ change with k

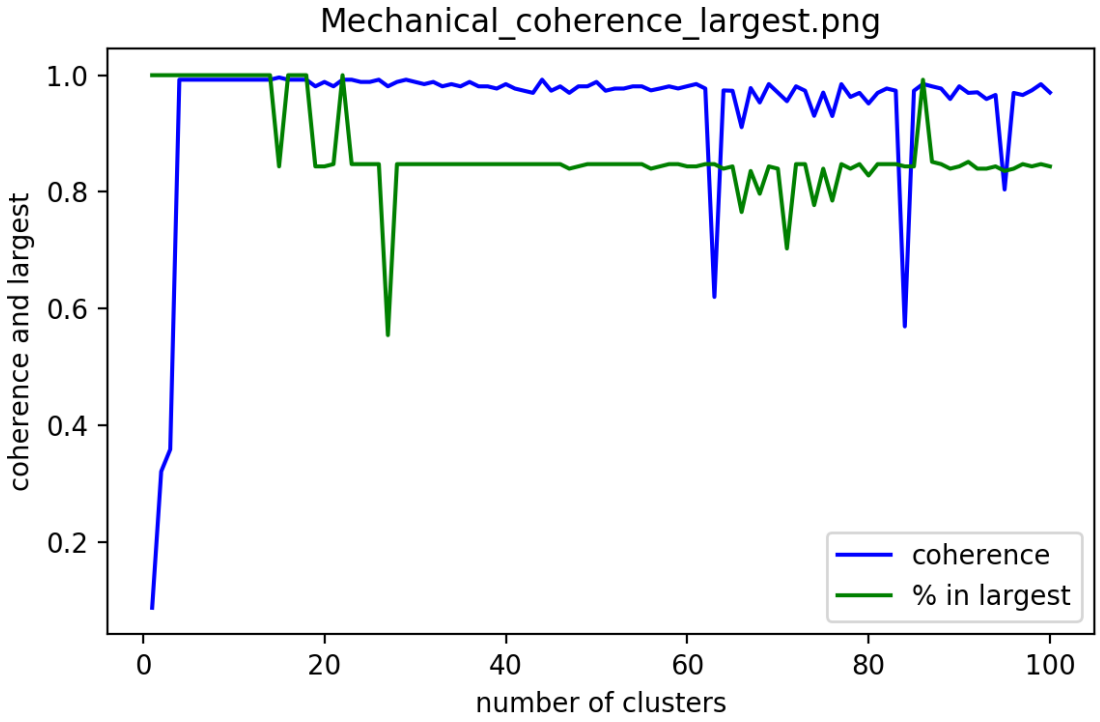


Figure B.12: Mechanical Engineering - Coherence$_{Maj}$ and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ change with k

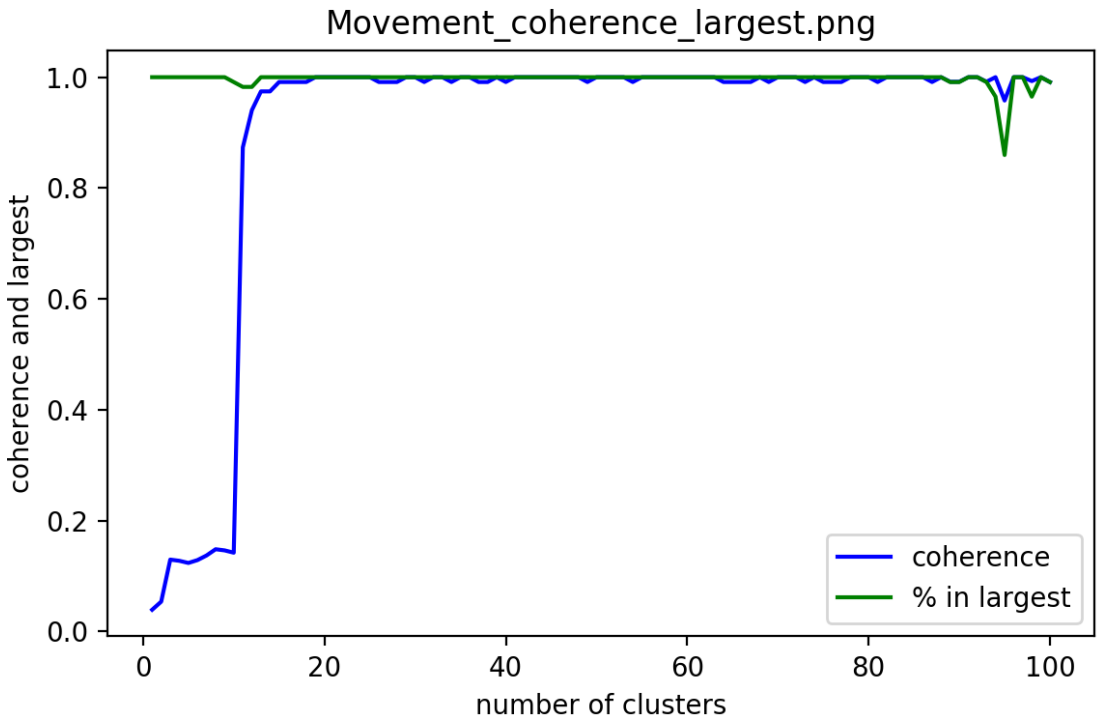


Figure B.13: Movement Science - Coherence$_{Maj}$ and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ change with k

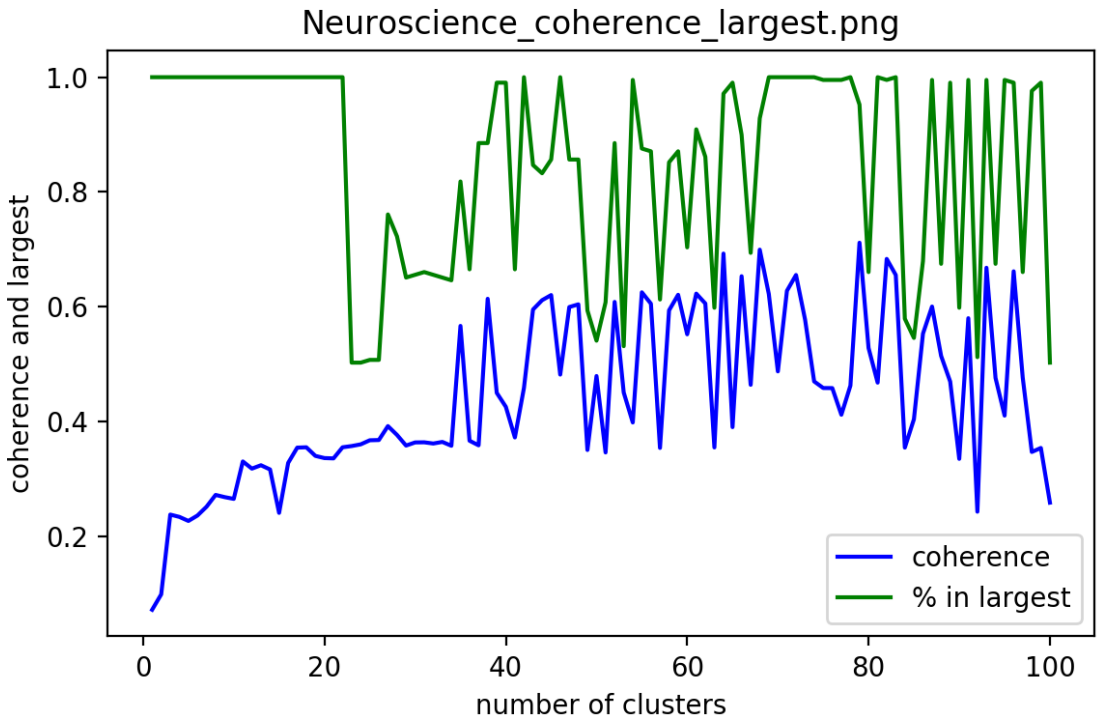


Figure B.14: Neuroscience - Coherence$_{Maj}$ and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ change with k

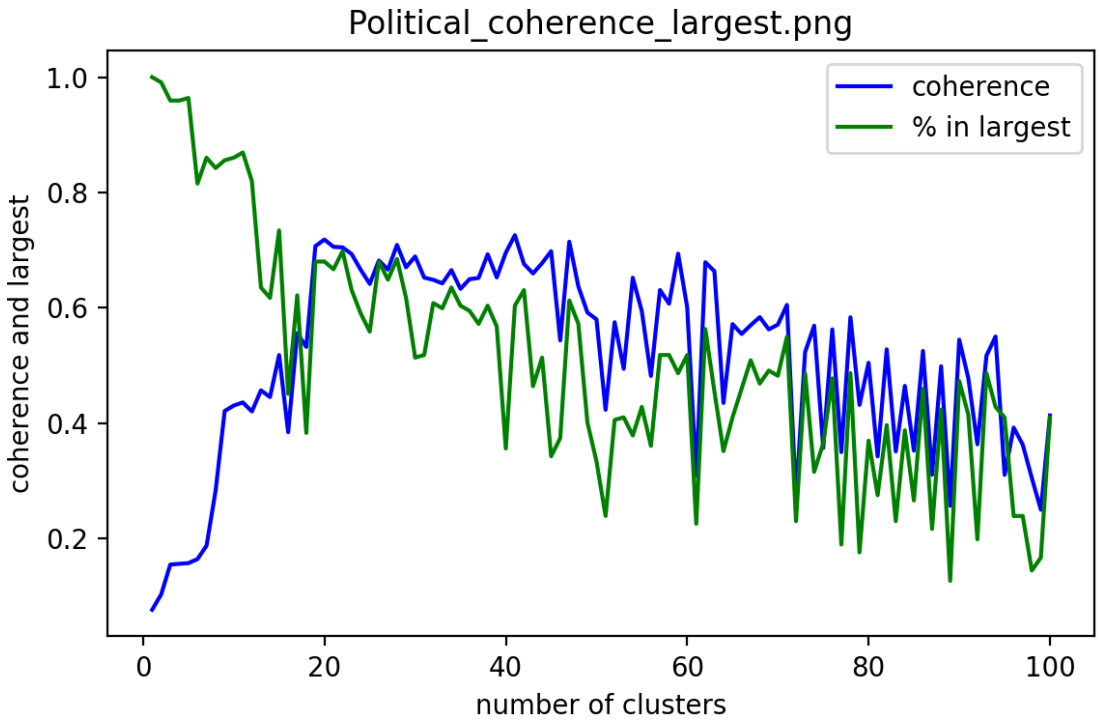


Figure B.15: Political Science - Coherence$_{Maj}$ and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ change with k

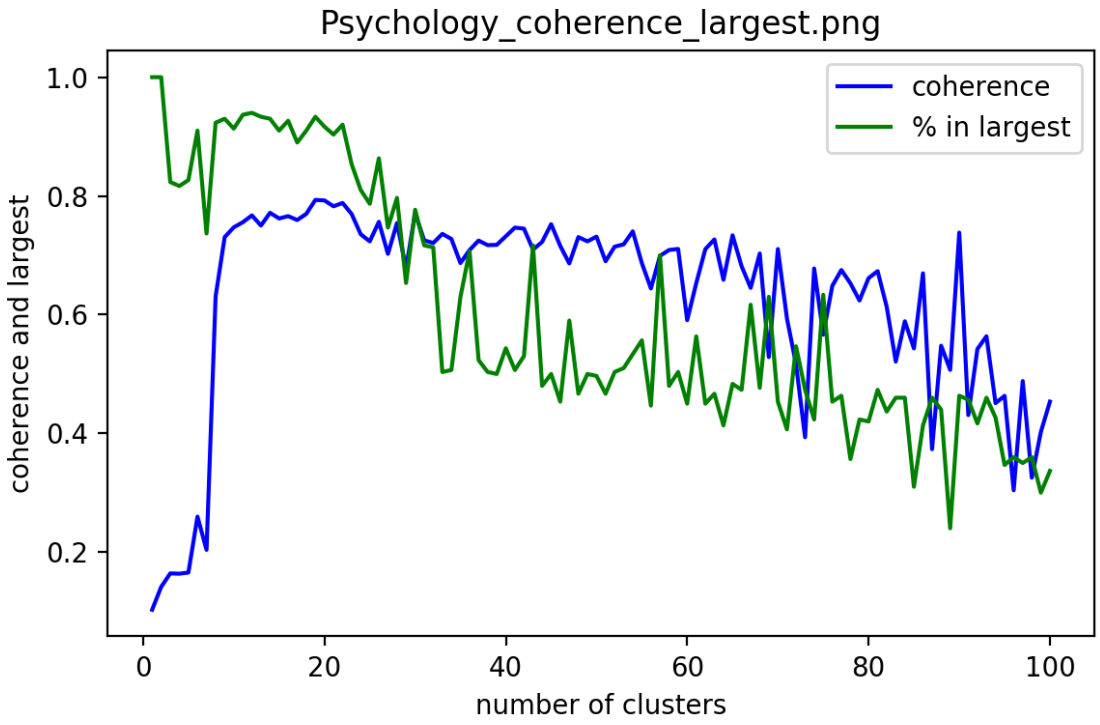


Figure B.16: Psychology - Coherence$_{Maj}$ and $P(M_{\text{Maj}}|C^k_{\text{Maj}})$ change with k

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Learning analytics data architecture (larc).

[2] JS Antrobus, R Dobbelaer, and S Salzinger. Social networks and college success, or grade point average and the friendly connection. *Social networks of children, adolescents, and college students*, 227:260, 1988.

[3] Susan Biancani and Daniel A McFarland. Social networks research in higher education. In *Higher education: Handbook of theory and research*, pages 151–215. Springer, 2013.

[4] Antoni Calvó-Armengol, Eleonora Patacchini, and Yves Zenou. Peer effects and social networks in education. *The Review of Economic Studies*, 76(4):1239–1267, 2009.

[5] Brian V Carolan. *Social network analysis and education: Theory, methods & applications*. Sage Publications, 2013.

[6] Karina L Cela, Miguel Ángel Sicilia, and Salvador Sánchez. Social network analysis in e-learning environments: A preliminary systematic review. *Educational Psychology Review*, 27(1):219–246, 2015.

[7] Anthony R D'Augelli and Scott L Hershberger. African american undergraduates on a predominantly white campus: Academic factors, social networks, and campus climate. *The Journal of Negro Education*, 62(1):67–81, 1993.

[8] Shane Dawson. A study of the relationship between student social networks and sense of community. *Journal of educational technology & society*, 11(3):224–238, 2008.

[9] David DiRamio, Ryan Theroux, and Anthony J Guarino. Faculty hiring at top-ranked higher education administration programs: An examination using social network analysis. *Innovative Higher Education*, 34(3):149–159, 2009.

[10] Ian Dobson, Benjamin A Carreras, Vickie E Lynch, and David E Newman. Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 17(2):026103, 2007.

[11] Greg J Duncan, Johanne Boisjoly, Michael Kremer, Dan M Levy, and Jacque Eccles. Peer effects in drug use and sex among college students. *Journal of abnormal child psychology*, 33(3):375–385, 2005.

[12] Nicole Ellison and Charles Steinfield. C., and c. lampe, 2007."the benefits of facebook 'friends': Social capital and college students use of online social network sites,". *Journal of Computer–Mediated Communication*, 12(4):1–143.

[13] Leon Festinger, Stanley Schachter, and Kurt Back. Social pressures in informal groups; a study of human factors in housing. 1950.

[14] Ed Fincham, Dragan Gaševic, and Abelardo Pardo. From social ties to network processes: Do tie definitions matter?. *Journal of Learning Analytics*, 5(2):9–28, 2018.

[15] Daniel Z Grunspan, Benjamin L Wiggins, and Steven M Goodreau. Understanding classrooms through social network analysis: A primer for social network analysis in education research. *CBE—Life Sciences Education*, 13(2):167–178, 2014.

[16] Patricia Gurin, Eric Dey, Sylvia Hurtado, and Gerald Gurin. Diversity and higher education: Theory and impact on educational outcomes. *Harvard educational review*, 72(3):330–367, 2002.

[17] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[18] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[19] Maureen E Kenny and Sonia Stryker. Social network characteristics of white, african-american, asian and latino/a college students and college adjustment: A longitudinal study. 1994.

[20] Kevin Lewis, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, and Nicholas Christakis. Tastes, ties, and time: A new social network dataset using facebook. com. *Social networks*, 30(4):330–342, 2008.

[21] Adalbert Mayer and Steven L Puller. The old boy (and girl) network: Social network formation on university campuses. *Journal of public economics*, 92(1-2):329–347, 2008.

[22] Samuel Messick. Validity of psychological assessment: Validation of inferences from persons'responses and performances as scientific inquiry into score meaning. *ETS Research Report Series*, 1994(2):i–28, 1994.

[23] Theodore M Newcomb. The acquaintance process as a prototype of human interaction. 1961.

[24] Mark Newman. *Networks*. Oxford university press, 2018.

[25] Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2):404–409, 2001.

[26] Mark EJ Newman, Stephanie Forrest, and Justin Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, 66(3):035101, 2002.

[27] Gaia Data Release. Aga brown et al 2016. *Astronomy & Astrophysics*, 595:A2, 1.

[28] Leslie Lane Salzinger. The ties that bind: The effect of clustering on dyadic relationships. *Social Networks*, 4(2):117–145, 1982.

[29] Parongama Sen, Subinay Dasgupta, Arnab Chatterjee, PA Sreeram, G Mukherjee, and SS Manna. Small-world properties of the indian railway network. *Physical Review E*, 67(3):036106, 2003.

[30] Scott L Thomas. Ties that bind: A social network approach to understanding student integration and persistence. *The journal of higher education*, 71(5):591–615, 2000.

[31] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.

[32] Demival Vasques Filho and Dion R J O'Neale. Degree distributions of bipartite networks and their projections. *Phys Rev E*, 98(2-1):022307, August 2018.

[33] George Veletsianos, Amy Collier, and Emily Schneider. Digging deeper into learners' experiences in mooc s: Participation in social networks outside of mooc s, notetaking and contexts surrounding content consumption. *British Journal of Educational Technology*, 46(3):570–587, 2015.

[34] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

[35] Stanley Wasserman, Katherine Faust, et al. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[36] Megumi Watanabe and Christina Falci. Workplace faculty friendships and work-family culture. *Innovative Higher Education*, 42(2):113–125, 2017.

[37] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440, 1998.

[38] Donald G York, J Adelman, John E Anderson Jr, Scott F Anderson, James Annis, Neta A Bahcall, JA Bakken, Robert Barkhouser, Steven Bastian, Eileen Berman, et al. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579, 2000.

[39] Tao Zhou, Jie Ren, Matús Medo, and Yi-Cheng Zhang. Bipartite network projection and personal recommendation. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 76(4 Pt 2):046115, October 2007.

[40] David J Zimmerman. Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and statistics*, 85(1):9–23, 2003.

[41] Justyna P Zwolak, Remy Dou, Eric A Williams, and Eric Brewe. Students' network integration as a predictor of persistence in introductory physics courses. *Physical Review Physics Education Research*, 13(1):010113, 2017.